MODEL REDUCTION AND IDENTIFICATION OF WASTEWATERTREATMENT PLANTS – A SUBSPACE APPROACH

O. A. Z. SOTOMAYOR † , S. W. PARK † and C. GARCIA ‡

† *Department of Chemical Engineering, Polytechnic School of the University of São Paulo, SP, Brazil*
oscar@lscp.pqi.ep.usp.br
‡ *Dept. of Telecommunications and Control, Polytechnic School of the University of São Paulo, SP, Brazil*
clgarcia@lac.usp.br.

Abstract **--In this paper, a low-order linear time-invariant (LTI) state-space model that describes the nitrate concentrations in both anoxic and aerobic zones of an activated sludge wastewater treatment plant (WWTP), for biological treatment of municipal sewage, is identified around a given operating point (a model with lumped parameters). Several subspace identification methods, such as CCA, N4SID, MOESP and DSR are applied and their performance are compared, based on performance quality criteria, in order to select the best-reduced model. The selected model is validated with a data set not used in the identification procedure and it describes well the complex dynamics of the process. This model is asymptotically stable and it can be used for control, optimization, prediction and monitoring purposes. In this work the ASWWTP-USP benchmark is used as a data generator. This benchmark simulates the biological, chemical and physical interactions that occur in a complex activated sludge plant.**

Keywords **--subspace identification methods, reduced order models, state-space models, activated sludge process, wastewater treatment.**

I. INTRODUCTION

Advanced engineering applications require suitable mathematical models. System identification deals with the problem of obtaining "approximate" models of dynamic systems from measured input-output data. This issue is of interest in a variety of applications, ranging from chemical process simulation and control to identification of vibrational modes in flexible structures. The most traditional system identification techniques are the prediction error method (PEM) and the instrumental variable method (IVM). These methods are primarily used with the so-called black-box model structures (Viberg, 1995). However, several important problems remain to be solved. The PEM has excellent statistical properties provided the "true" PEM estimate can be found. Nevertheless, computing the PEM model can sometimes be overwhelmingly difficult. In general, a multi-dimensional non-linear optimization problem must be solved. On the other hand, the IVM attempts to deliver parameter estimates by only solving linear systems of equations. However, the use of these models is quite cumbersome in the general multivariable case, and the numerical reliability may be unacceptably high for complex cases involving large system orders and many outputs (Viberg, 2002). The preferred model structure for complex problems is therefore a state-space model.

Subspace methods have their origin in state-space realization. Subspace identification method is a technique that has been developed since the late 80's. It has attracted much attention, owing to its computational simplicity and effectiveness in identifying dynamic state-space linear multivariable systems. These algorithms are numerically robust and do not involve non-linear optimization techniques, i.e., they are fast (non-iterative) and accurate (since no problems with local minima occur). The computational complexity is modest compared to PEM, particularly when the number of inputs and outputs is large. Because applications of large dimensions are commonly found in the process industry, subspace identification methods are very promising in this field. As a result, a large number of successful applications of subspace identification methods for simulated and real processes have been reported in the literature. A general overview of the state-of-the-art in subspace identification methods is presented in De Moor *et a*l. (1999) and Favoreel *et a*l. (2000).

In this paper, a low-order LTI state-space multivariable model that describes the nitrate concentrations in the anoxic and aerobic zones of an activated sludge process is estimated around an operating point. Several subspace identification methods are applied and their performances are compared in order to select the best-obtained model. It can be used to control the process, e.g., as in Lindberg (1997), where a multivariable control algorithm based on a subspace model is used to regulate an activated sludge process. Previous performance comparisons of several subspace methods, applied to other processes, can be found in Abdelghani *et a*l. (1998), Katayama *et a*l. (1998) and Favoreel *et a*l. (1999).

In this work, the ASWWTP-USP (Activated Sludge Wastewater Treatment Plant – University of São Paulo) benchmark (Sotomayor *et a*l., 2001a) is used as a data generator. This benchmark simulates the biological, chemical and physical interactions that occur in a complex activated sludge plant.

II. SUBSPACE IDENTIFICATION METHODS

The subspace identification methods refer to a class of algorithms whose main characteristic is the approximation of subspaces generated by the row spaces of block-Hankel matrices of the input/output data, to calculate a reliable discrete-time state-space model of the following form:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + w_k \\ y_k &= Cx_k + Du_k + v_k \end{aligned} \tag{1}$$

with

$$\mathbf{E}\left[\begin{pmatrix} w_p \\ v_p \end{pmatrix}\begin{pmatrix} w_q^T & v_q^T \end{pmatrix}\right] = \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix}\delta_{pq} \geq 0 \tag{2}$$

where x represents the model state vector, u is the manipulated input vector and y is the process output vector. A is the system (state transition) matrix, B is the input matrix, C is the output matrix and D is the direct input to output matrix. w is called the process noise and v is called the measurement noise. They are assumed to be unmeasurable gaussian-distributed zero-mean white noise vector sequences and uncorrelated with the inputs u. The matrices Q, S and R are the covariance matrices of the noise sequences w and v. $\mathbf{E}$ denotes the expected value operator and δ_{pq} the Kronecker delta. The subscript index k denotes a time discrete (sampled) system. Related to Eq. (1), it is assumed that the system is asymptotically stable, the pair (A,C) is observable and the pair (A,B) is controllable.

It is common practice to distinguish among the three possible situations regarding the inputs acting on the system of the Eq. (1): (1) the purely deterministic case $(w = v = 0)$, (2) the purely stochastic case $(u = 0)$, and (3) the combined deterministic/stochastic case.

There are now many different versions of subspace algorithms, and they have reached a certain level of maturity. All subspace identification methods consist of three main step: estimating the predictable subspace for multiple future steps, then extracting state variables from this subspace and finally fitting the estimated states to a state-space model. See the main issues related to subspace identification methods and one particular technique (the "standard" N4SID) in Delgado *et a*l. (2001). Nevertheless, each subspace identification method looks quite different from other in concept, computation tools and interpretation. The major differences among these subspace identification methods lie in the regression or projection methods used in the fist step to remove the effect of the future inputs on the future outputs and thereby estimate the predictable subspace, and in the latent variable methods used in the second step to extract estimates of the states.

The major advantages of these algorithms are that they only need input-output data and very little prior knowledge about the system. In addition, these algorithms are based on system theory, geometry and numerically stable non-iterative linear algebra operations, such as QR (or LQ)–factorization, SVD (singular value decomposition) and its generalizations, for which good numerical tools are well-known (Golub and VanLoan, 1996). A drawback against subspace identification approach is that the physical insight of the process, in the obtained model, is lost, which is a characteristic of black-box models. Furthermore, a large amount of data is required to obtain accurate models. Actually, generating and collecting data of some processes can be too expensive.

The subspace identification algorithms considered in this study are:

a. Unconstrained and constrained version of the Canonical Correlation Analysis (CCA) algorithm, both in Peternell *et al*. (1996).
b) Past Output (PO) variant of the Multivariable Output-Error State-sPace model identification (MOESP) algorithm, in the SMI Toolbox by Haverkamp and Verhaegen (1997).
c) Numerical algorithm for Subspace State-Space System IDentification (N4SID): the "standard" version in the MATLAB System Identification Toolbox v.4.0.4 (Ljung, 1997), that implements the N4SID algorithm from Van Overschee and De Moor (1994), and the "robust" version from Van Overschee and De Moor (1996).
d) Deterministic and Stochastic subspace system identification and Realization (DSR) in the D-SR Toolbox by Di Ruscio (1997).

As previously mentioned, the purpose of the present paper is to compare the performance of these methods and not to analyze their implementational differences. As for the detailed algorithm, the difference between these subspace identification methods seems so large that it is hard to find the similarities between them.

III. THE WASTEWATER TREATMENT PLANT SIMULATION

The ASWWTP-USP benchmark is a dynamic model, developed to simulate the processes that occur in a biological WWTP. The benchmark represents a continuous-flow predenitrifying activated sludge process, a frequently applied system for removal of organic matter and nitrogen from municipal sewage, predominantly domestic, operating at a constant temperature of 15°C and neutral pH. The layout of the process is formed by a bioreactor, composed of an anoxic zone and two aerobic zones, coupled with a secondary settler, as shown in Fig. 1.

For a reliable simulation of an activated sludge WWTP, the ASWWTP-USP benchmark is based on models widely accepted by the international community. Each bioreactor zone is modeled by the Activated Sludge Model ASM1 (Henze *et al*., 1987) and the secondary settler is modeled by the double-exponential

settling velocity model of Takács *et al*. (1991). The complete plant model includes 52 large, complex, coupled non-linear differential equations, which were implemented in MATLAB™/Simulink. The values of the process parameters are here omitted, but they can be found in Sotomayor *et al*. (2001a). For more realistic simulations, a white noise, with zero-mean and standard deviation 0.05, was added to the outputs produced by the benchmark.

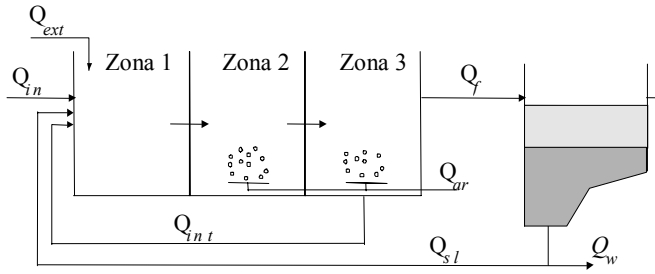

Figure 1. Layout of the ASWWTP-USP benchmark

IV. IDENTIFICATION OF A SUBSPACE MODEL OF THE PROCESS

A. Selection of Probing Signals, Generation and Pre-Treatment of Data Set

It is not very easy to select either the input or the output variables of the process. In this work, the nitrate concentrations in the anoxic zone Sno_1 (mg N/l) and in the last aerobic zone Sno_3 (mg N/l) are selected as outputs. The internal recirculation rate Q_{int} (m^3/h) and the external carbon dosage Q_{ext} (l/h) are considered as inputs. However, to improve the model influent flow rate Q_{in} (m^3/h), influent readily biodegradable substrate Ss_{in} (mg COD/l) and influent ammonium concentration Snh_{in} (mg N/l) are assumed as measurable disturbances, while influent nitrate concentration Sno_{in} (mg N/l) is assumed as an unmeasurable disturbance. The signals used in the identification procedure are summarized in Fig. 2.

Pseudo-random binary sequences (PRBS) are widely used in the identification of linear systems. However, since the PRBS consists of only two levels, the resulting data may not provide sufficient information to excite nonlinear dynamics. Additionally, a PRBS signal of a too large magnitude may bias the estimation of the linear Kernel. Multi-level (m-level) sequences, in contrast, allow the user to highlight nonlinear system behavior while manipulating the harmonic content of the signal, reducing the effect of nonlinearities in the resulting linear model (Godfrey, 1993). Moreover, the ill-conditioning (insufficient excitation) of probing inputs may lead to a substantial deterioration of performance of the subspace algorithms (Chiuso and Picci, 2000). In the present paper, the data signals correspond to m-level uniformly distributed random sequences. Their amplitudes and frequencies were chosen so as to adequately excite the system, without deviating too much from the normal operating point and, therefore, enabling the identification of a suitable linear model. All data signals are stored at a sampling rate of 0.16 hours to obtain 1400 samples.

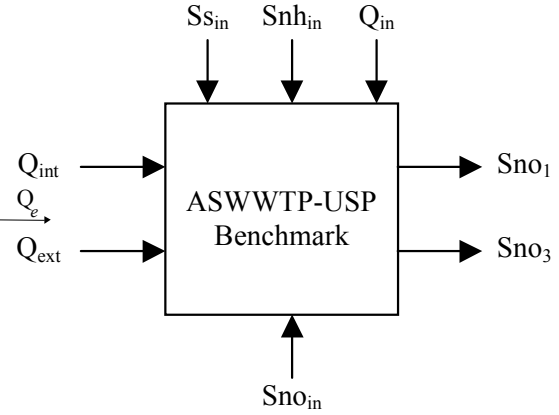

Figure 2. Signals for subspace identification

For a better identification result, the raw data set is pre-processed. As pointed out by Chui (1997), it is important to make sure that the scales of the input-output data are of comparable sizes. Therefore, all data signals are normalized aiming to be equally weighted. This operation is common in system identification. After, the data set is detrended in order to remove periodic components of known periodic length. This step is usual in signal processing (Bauer, 2000a). The pre-processed signals are shown in Fig. 3.

The identification process was carried out off-line in batch form by using the first 1000 points of the data set, whereas the remaining 400 points were applied for model validation. In the identification procedure is done in open loop and the purely deterministic case is considered.

B. Order Estimation

There is an extensive literature for order estimation algorithms for linear, dynamical, state-space systems. Nevertheless, there exist only few references dealing with the estimation of the order in the context of subspace identification methods (Bauer, 2001). In many cases, the determination of the system order *n* is very subtle. Normally, this information is obtained by detecting a gap in the spectrum of the singular values of the orthogonal (or oblique) projections of the row spaces of data block-Hankel matrices. In the present case, the gap is not easy to determine, as it is seen in figure 4, and hence the application of this strategy becomes subjective and the decision regarding the order of the model is an arbitrary one.

According to Bastogne *et al*. (1998), a more practical procedure is to choose the value n that minimizes the estimation errors, as shown in figure 5, which was generated using the "robust" N4SID algorithm. Comparing the relative estimation error indexes, it can be noticed that the 3rd, 6th and 7th-order model have practically the same mean error index. Nevertheless, the

choice of 6th or 7th-order does not bring enough improvement in comparison with a reduced order given by the 3rd-order model, which is the selected order estimation. For $n = 3$ the relative square error was 34.60% for the case of Sno_1 and 35.07% for the case of Sno_3, with a mean error of 34.84%.

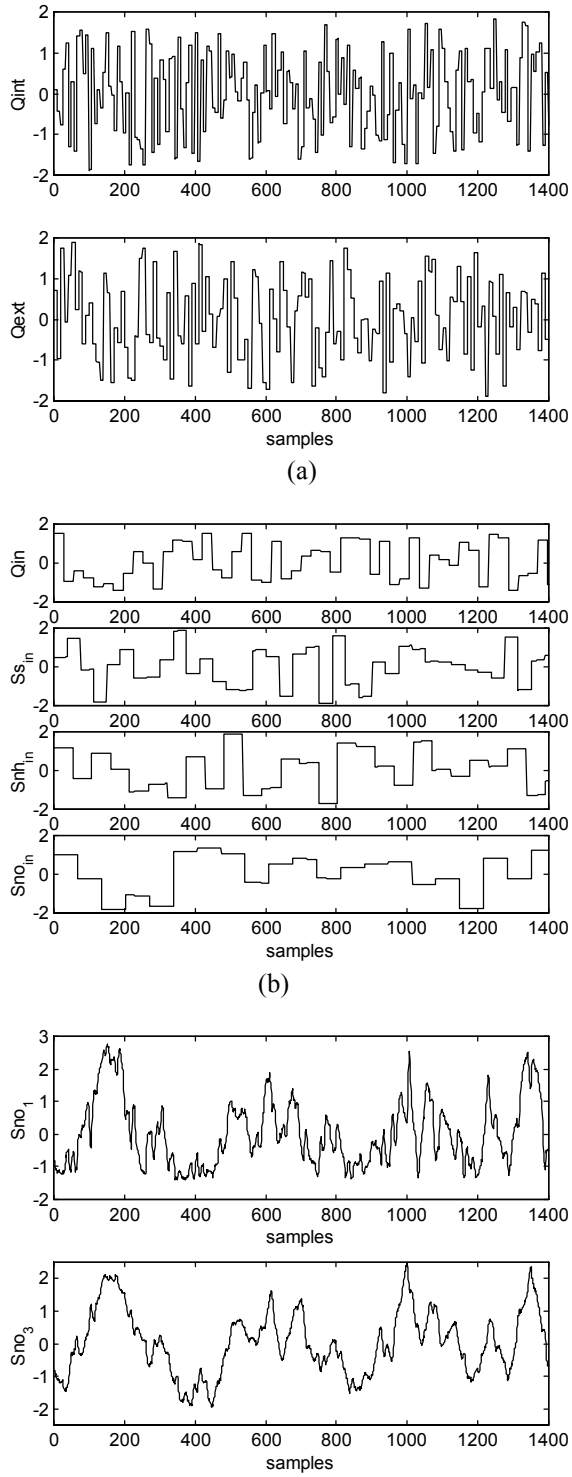

Figure 3. Data sequences of the process: (a) inputs; (b) disturbances; (c) outputs.

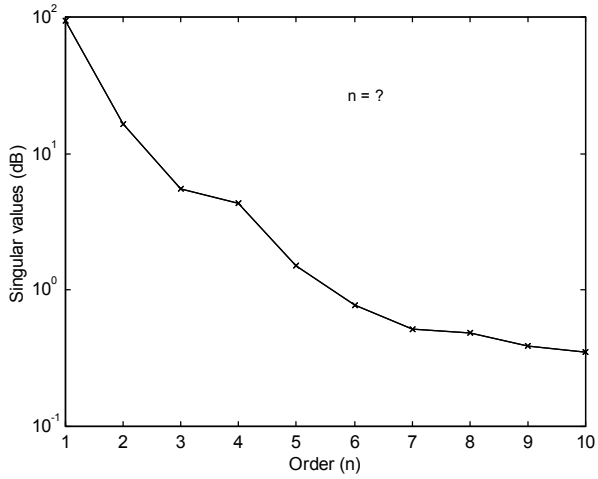

Figure 4. Singular value spectrum

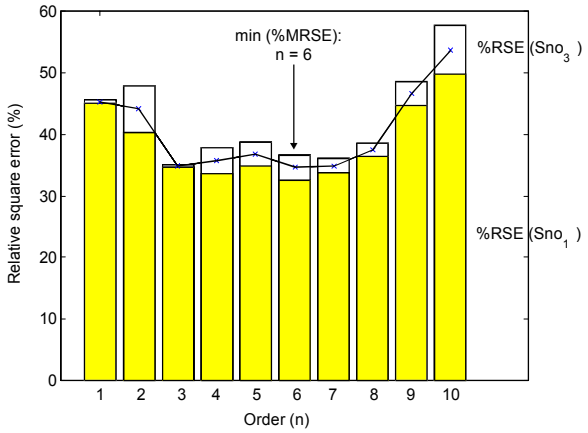

Figure 5. Estimation error spectrum

C. Performance Quality Criteria

In the present paper, two performance indicators are proposed to measure identification/validation error, in order to obtain the best 3rd-order state-space model. The performance indicators are:

Mean relative square error (MRSE):

$$\%MRSE = \frac{1}{l} \cdot \sum_{i=1}^{l} \sqrt{\frac{\sum_{j=1}^{N} (y_i(j) - \hat{y}_i(j))^2}{\sum_{j=1}^{N} (y_i(j))^2}} \times 100 \qquad (3)$$

Mean variance-accounted-for (MVAF):

$$\%MVAF = \frac{1}{l} \cdot \sum_{i=1}^{l} \left(1 - \frac{\text{variance}(y_i - \hat{y}_i)}{\text{variance}(y_i)} \right) \times 100 \qquad (4)$$

being N the number of identification data points, l the number of outputs, y_i the *i-th* real output and $\hat{y}_i$ the *i-th* simulated output produced by the model. The MRSE index is widely used in the literature, while the MVAF index is specifically used by the SMI Toolbox.

Analyzing the values in Table 1, the PO-MOESP model seems to produce a better model in terms of identification, while the DSR model seems to produce a better model in terms of validation. Hence, in this work, the 3rd-order DSR model was chosen to describe the process.

Table 1. Performance of the subspace-based algorithms

Algorithm	%MRSE		%MVAF	
	Identific.	Validation	Identific.	Validation
uCCA	40.4417	69.9404	83.5750	73.5628
cCCA	40.1652	69.2404	83.7998	73.9129
PO-MOESP	**31.8091**	57.5806	**89.9037**	79.3096
"standard" N4SID	44.4914	72.9242	80.0546	74.2431
"robust" N4SID	34.8394	57.7508	87.8739	81.2475
DSR	34.2450	**50.9904**	88.2237	**84.4274**

D. Identification Results

The selected deterministic model (proper model) is described by the following matrices:

$$A=\begin{bmatrix} 0.9763 & 0.0194 & 0.3268 \\ 0.0061 & 0.8815 & 0.0893 \\ -0.0023 & 0.0071 & 0.9763 \end{bmatrix},$$

$$B=\begin{bmatrix} 0.0238 & -0.0459 & -0.1488 & -0.0403 & 0.0002 \\ -0.1295 & 0.0299 & 0.0230 & 0.0185 & -0.0052 \\ 0.0097 & -0.0082 & -0.0082 & 0.0004 & 0.0036 \end{bmatrix}$$

$$C=\begin{bmatrix} 0.2253 & -0.4032 & -0.1823 \\ 0.2668 & 0.2880 & -0.4626 \end{bmatrix},$$

$$D=\begin{bmatrix} 0.1292 & -0.0193 & -0.0651 & -0.0312 & 0.0053 \\ -0.0387 & 0.0086 & 0.0126 & 0.0105 & -0.0026 \end{bmatrix}$$

A strictly proper model (i.e, with $D=0$) is also identified, and it is described by:

$$A=\begin{bmatrix} 0.9763 & 0.0199 & 0.3263 \\ 0.0062 & 0.8818 & 0.0907 \\ -0.0024 & 0.0072 & 0.9758 \end{bmatrix},$$

$$B=\begin{bmatrix} 0.0368 & -0.0434 & -0.1537 & -0.0431 & -0.0045 \\ -0.1505 & 0.0234 & 0.0357 & 0.0283 & -0.0044 \\ 0.0167 & -0.0100 & -0.0091 & 0.0003 & 0.0039 \end{bmatrix}$$

$$C=\begin{bmatrix} 0.2259 & -0.4026 & -0.1810 \\ 0.2664 & 0.2876 & -0.4633 \end{bmatrix} \quad (5)$$

The poles (eigenvalues of A) of the proper and the strictly proper model are closer to the unit circle and they shown the slower dynamics of the process. In addition, the poles close to 1 show that the data set seems to contain a phenomenon known as "co-integration" in econometrics. Based on this observation, it is possible to obtain models which produce a one-step-ahead prediction error much smaller (Bauer, 2000b).

Figure 6 shows the outputs generated by the identified strictly proper model (dotted line). As it can be observed, the identified model for a given operating points correctly reproduces the main dynamic characteristics of the activated sludge process. In these graphics, either the identification or the validation data were introduced in the obtained model. In both cases the simulation started at zero initial conditions.

Low-order state-space models sufficiently representative of the nominal system behavior are a prerequisite to the systematic design of control systems. So, the strictly proper model (5) has been successfully used to develop an infinite-horizon optimal control (Sotomayor *et al*., 2001b) and a model predictive control (Sotomayor *et al*., 2001c).

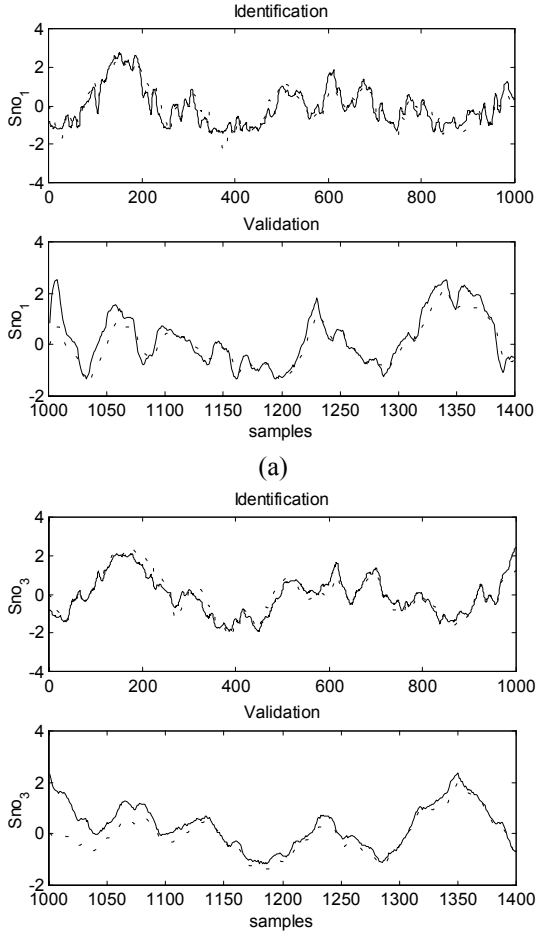

Figure 6. Output comparison (a) for Sno_1, (b) for Sno_3

V. CONCLUSIONS

The use of subspace identification methods has proved to be a valuable tool in the estimation of LTI state-space models for the activated sludge process. The performance of different identification algorithms (CCA, MOESP, N4SID and DSR) was compared. Although the used simulation benchmark consists of 52 differential equations, the 3rd-order DSR model, a very reduced order model, manages to describe sufficiently well the process dynamics. It is well suited for model-based control as well as for monitoring applications.

Acknowledgements

The authors thank the financial support from FAPESP (Brazil) under grant 98/12375-7. They are also grateful to Dietmar Bauer from the Technische Universität Wien (Austria), Bert Haverkamp from the Delft University of Technology (the Netherlands), Peter Van Overschee from the Katholieke Universiteit Leuven (Belgium) and David Di Ruscio from the Telemark Institute of Technology (Norway), for having supplied their subspace Matlab-codes.

REFERENCES

Abdelghani, M., Verhaegen, M., Van Overschee, P., De Moor, B., "Comparison study of subspace identification methods applied to flexible structures", *Mechanical Systems and Signal Processing* **12**(5), 679-692 (1998).

Bastogne, T., Noura, H., Sibille, P.,Richard, "A. Multivariable identification of a winding process by subspace methods for tension control", *Control Engineering Practice* **6**(9), 1077-1088 (1998).

Bauer, D., "On data preprocessing for subspace methods", *Proc. of the 39th IEEE Conference on Decision and Control*. Sydney, Australia (2000a).

Bauer, D., *Personal Communication* (2000b).

Bauer, D., "Order estimation for subspace methods", *Automatica* **37**(10), 1561-1573, (2001).

Chiuso, A.; Picci, G., "Probing inputs for subspace identification", *Proc. of the 39th IEEE Conference on Decision and Control*. Sydney, Australia (2000).

Chui, N. L., "Subspace methods and informative experiments for system identification", *PhD Thesis*. University of Cambridge, United Kingdom (1997).

Delgado, C. J. M., Lopes dos Santos, P., Martins de Carvalho, J. L., "On-line subspace identification", In: *Proc. of the European Control Conference ECC 2001*. Porto, Portugal (2001).

De Moor, B., Van Overschee, P., Favoreel, W., "Algorithms for subspace state space system identification – an overview", in: Biswa Datta, editor, *Applied and Computational Control, Signal and Circuits*, Vol. 1, Chapter 6, pp.247-311, Birkhäuser Boston (1999).

Di Ruscio, D., "A method for identification of combined deterministic and stochastic systems", In: M. Aoki and A. Heavenner, editors, *Applications of Computer Aided Time Series Modeling*, 181-235, Springer-Verlag (1997).

Favoreel, W., Van Huffel, S., De Moor, B., Sima, V., "Comparative study between three subspace identification algorithms", *Proc. of the European Control Conference ECC'99*. Karlsruhe, Germany (1999).

Favoreel, W., De Moor, B., Van Overschee, P., "Subspace state space system identification for industrial processes", *Journal of Process Control* **10** (2-3), 149-155 (2000).

Godfrey, K. *Perturbation Signals for System Identification*. Prentice-Hall Int., Hertfordshire, UK (1993).

Golub, G. H., VanLoan, C. F. *Matrix Computation*. Johns Hopkins University Press. Baltimore, MD, 3rd edition (1996).

Haverkamp, B., Verhaegen, M. *SMI Toolbox: state space model identification software for multivariable dynamical systems*, v.1.0. Delft University of Technology (1997).

Henze, M., Gujer, W., Marais, G., Matsuo, M. Activated sludge model N°1. Scientific and Technical Report No.1. IAWQ, London (1987).

Katayama, T., Omori, S., Picci, G., "A comparison of some subspace identification methods", *Proc. of the 37th. IEEE Conference on Decision and Control*. Tampa, Florida (1998).

Lindberg, C. F., *Control and estimation strategies applied to the activated sludge processes*, Ph.D. Thesis, Uppsala University, Sweden (1997).

Peternell, K., Scherrer, W., Deistler, M., "Statistical analysis of novel subspace identification methods", *Signal Processing* **52**(2), 161-177(1996).

Sotomayor, O.A.Z., Park, S. W., Garcia, C., "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 (2001a).

Sotomayor, O.A.Z., Park, S. W., Garcia, C., "Subspace-based optimal control of N-removal activated sludge plants", *Proc. of the 10th. IEEE International Conference on Control Applications CCA'2001*. Mexico City, Mexico (2001b).

Sotomayor, O.A.Z., Park, S. W., Garcia, C., "MPC control of a predenitrification plant using linear subspace models", in: *12th European Symposium on Computer Aided Process Engineering, ESCAPE-12*, The Hague, The Netherlands, 553-558 (2001c).

Takács, I., Patry, G., Nolasco, D., "A dynamic model of clarification-thickening process", *Water Research* **25**(10), 1263-1271 (1991).

Van Overschee, P., De Moor, B., "N4SID: subspace algorithms for the identification of combined deterministic-stochastic systems", *Automatica, Special Issue on Statistical Signal Processing and Control* **30**(1), 75-93 (1994).

Van Overschee, P., De Moor, B., *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer Academic Publishers. Dordrecht (1996).

Viberg, M., "Subspace-based methods for the identification of linear time-invariant systems", *Automatica, Special Issue on Trends in System Identification* **31**(12), 1835-1851 (1995).

Viberg, M., "Subspace-based state-space system identification", *Circuits, Systems and Signal Processing*, **21**,(1), 23-27 (2000).

Received: September 16, 2001.
Accepted for publication: August 28, 2002.
Recommended by Guest Editors: J. Cerdá, S. Díaz and A. Bandoni.