

A numerical identifiability test for state-space models – application to optimal experimental design

M.E. Hidalgo* and E. Ayesa*, **

* Section of Environmental Engineering, CEIT, Po. Manuel Lardizabal, 15, 20018, San Sebastián, Spain

** ESII, University of Navarra, Po. Manuel Lardizabal, 13, 20018, San Sebastián, Spain

Abstract This paper describes a mathematical tool for identifiability analysis, easily applicable to high order non-linear systems modelled in state-space and implementable in simulators with a time-discrete approach. This procedure also permits a rigorous analysis of the expected estimation errors (average and maximum) in calibration experiments. The methodology is based on the recursive numerical evaluation of the information matrix during the simulation of a calibration experiment and in the setting-up of a group of information parameters based on geometric interpretations of this matrix. As an example of the utility of the proposed test, the paper presents its application to an optimal experimental design of ASM Model No.1 calibration, in order to estimate the maximum specific growth rate μ_H and the concentration of heterotrophic biomass X_{BH} .

Keywords Activated sludge; calibration; identifiability; optimal experimental design; Fisher Information Matrix; parameter estimation

Introduction

Mathematical modelling of a biological process is a useful tool for its design, operation and control, provided that these models are sufficiently calibrated and validated. In the calibration of model parameters, identifiability analysis plays an important role (Beck, 1989). Identifiability and optimal experimental design dealing with noisy and sampled data have generally been approached by analysing the properties of the Fisher information matrix (Walter and Pronzato, 1990; Munack, 1991). The way this approach should be implemented in discrete time simulators for nonlinear continuous systems with discrete outputs is not a secondary matter.

Most studies about identifiability and optimal experimental design for biodegradation processes have been applied to Monod-type models (Holmberg, 1982; Munack, 1989; Baltés *et al.*, 1994; Vanrolleghem *et al.*, 1995; Merkel *et al.*, 1996). However, an identifiability analysis of a global model like the Activated Sludge Model (ASM) No.1 (Henze *et al.*, 1986) should be done taking into account not only the parameters that are directly related to the measurements, but also the lack of knowledge of other parameters that can affect the calibration results due to compensations. This analysis is crucial considering that the parameter values obtained by calibration determinate, at a later phase, the results of design and control. For this reason, several studies have recently been published focused on the identifiability of the largest possible amount of parameters of global models from measurements obtained from continuous plant (Ayesa *et al.*, 1995; Weijers and Vanrolleghem, 1997) or batch tests (Brouwer *et al.*, 1998).

Another important issue in calibration tasks is the accuracy of parameter estimates. Indicators of expected variance in the estimation, functions of the information matrix, are well known but it would be very interesting to have also an indicator of the expected maximum error.

In this paper an identifiability test for multivariate nonlinear systems with discrete outputs is presented. It is easily implementable in computers and is based on the numerical evaluation of the Fisher information matrix. An example of an application of the developed

test for optimal experimental design is also given. In this test the global model is introduced in the processing of the global information matrix, so that not only the influences of the parameters directly related to the output through the model are considered, but also the indirect influence of all the parameters in the model. Moreover, it provides indicators of expected variance and expected maximum errors.

Theoretical framework

In order to simplify the computer implementation of the identifiability algorithm, the proposed methodology groups the state variables and the model parameters in an “extended” state vector (Ayessa *et al.*, 1995). The model parameters are introduced in the state vector as invariant states of the process. Then, the structure (1) and (2) can express a nonlinear system modelled in state space with scalar and discrete output,

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}(t)) \quad (1)$$

$$y(t_k) = h(\mathbf{x}(t_k)) \quad (2)$$

where $\mathbf{x} \in \mathbf{R}^n$ with $n = q + p$ is the “extended” state vector including the q original states of the system and the p parameters of the model and y is a scalar output of the model processed at each sampling time t_k . The p parameters are modelled by state equations like $\dot{x}_i = 0$ for $i = q + 1, \dots, n$.

The information accumulated by the measurements y up to sampling time t_k is stored in the Fisher information matrix given by (Walter and Pronzato, 1990)

$$\mathbf{M}(t_k, t_0) = \sum_{j=1}^k \mathbf{s}_j^{0T} \cdot \mathbf{s}_j^0 \quad (3)$$

or in a recursive form,

$$\mathbf{M}(t_k, t_0) = \mathbf{M}(t_{k-1}, t_0) + \mathbf{s}_k^{0T} \cdot \mathbf{s}_k^0 \quad (4)$$

where k is the number of the discrete samples in the experiment up to t_k and \mathbf{s}_j^0 is a $(1 \times n)$ vector given by Eq. (5) and called the *sensitivity vector at the sampling time t_j with respect to the initial state $\mathbf{x}(t_0)$* :

$$\mathbf{s}_j^0 = \frac{\partial y(t_j)}{\partial \mathbf{x}(t_0)} = \left[\frac{\partial y(t_j)}{\partial x_1(t_0)} \quad \frac{\partial y(t_j)}{\partial x_2(t_0)} \quad \dots \quad \frac{\partial y(t_j)}{\partial x_n(t_0)} \right] \quad (5)$$

Hence, to evaluate the information matrix, $\mathbf{M}(t_k, t_0)$, we need to evaluate the sensitivity vector, \mathbf{s}_j^0 , at each sampling time. Using differential theory, \mathbf{s}_j^0 is obtained by Eq. (6):

$$\mathbf{s}_j^0 = \frac{\partial y(t_j)}{\partial \mathbf{x}(t_0)} = \frac{\partial y(t_j)}{\partial \mathbf{x}(t_j)} \cdot \left[\frac{\partial \mathbf{x}(t_j)}{\partial \mathbf{x}(t_j - T)} \cdot \frac{\partial \mathbf{x}(t_j - T)}{\partial \mathbf{x}(t_j - 2T)} \quad \dots \quad \frac{\partial \mathbf{x}(t_j - (m-1)T)}{\partial \mathbf{x}(t_0)} \right] \quad (6)$$

where t_j is the time corresponding to an integer number m of sampling periods, $t_j = t_0 + m \cdot T$. If a small enough integration step size, Δt , is chosen so that the jacobian matrix, \mathbf{J} , can be considered constant in the integration interval, $(t_j, t_j + \Delta T)$, and assuming a sampling period, T , multiple of the step size, $T = \alpha \cdot \Delta t$, the expression that updates the sensitivity vector, \mathbf{s}_j^0 , at each sampling time is (Hidalgo, 1999),

$$\mathbf{s}_j^0 = \nabla h(\mathbf{x}(t_j)) \cdot \left[\prod_{l=1}^{m\alpha} e^{\mathbf{J}(\mathbf{x}(t_j - l\Delta t) - \Delta t)} \right] \quad (7)$$

with $e^{\mathbf{J} \cdot \Delta t}$ the exponential matrix.

If r measurements were processed at each sampling time, the approach would be analogous. Then, instead of a sensitivity vector, a $(r \times n)$ instantaneous sensitivity matrix with

respect to the initial state, $\mathbf{S}_{j,r}^0$, would be obtained. The algorithm also permits different sampling periods of the measurements.

Therefore, the developed algorithm basically consists of calculating the state transition matrix at each integration step and the sensitivity vector and the accumulated information matrix at each sampling time. Hence, this approach allows the numerical evaluation of the information matrix during the simulation of a calibration experiment.

Geometric interpretation of the information matrix

The developed test also considers the partition of the global information matrix to obtain a reduced information matrix \mathbf{M}_l . This reduced matrix stores the information about the unknown components of the initial extended state vector but takes into account the correlation between all the components, not only unknown but also known, of the extended state vector. \mathbf{M}_l will be $l \times l$ dimensional, with l the number of unknown components of the initial extended state vector. In this way, because of reduction of the matrix dimension, the examination of the information is more manageable as well as the compensations between estimates are easier to analyse. When the reduced information matrix is non-singular, a geometric interpretation is considered in order to facilitate the understanding and the study of the information stored in an experiment. This geometric interpretation is based on the eigenvalue decomposition of the information matrix,

$$\mathbf{M}_l = \mathbf{V} \cdot \Lambda \cdot \mathbf{V}^T = \sum_{i=1}^l \lambda_i \cdot \mathbf{v}_i \cdot \mathbf{v}_i^T \quad (8)$$

where Λ is a diagonal matrix with the eigenvalues λ_i and \mathbf{V} is the orthogonal matrix of eigenvectors \mathbf{v}_i . Then a reduced information hyperellipsoid and a reduced uncertainty hyperellipsoid can be defined. The *information hyperellipsoid* is an indicator of the quantity of information about the system provided by the measurements. The axes of this hyperellipsoid are in the directions of the eigenvectors of \mathbf{M}_l (or principal directions) and have a half-length of the square root of the corresponding eigenvalues λ_i . The *principal directions* represent the linear combinations of the original state variables whose estimates are uncorrelated (Bard, 1974; Draper and Smith, 1981). The *uncertainty hyperellipsoid* is an index of the region containing the possible initial state vector estimates. Its axes are in the directions of the eigenvectors of \mathbf{M}_l and have a half-length of the square root of the inverse of the corresponding eigenvalues λ_i .

From these hyperellipsoids we can define some information parameters based on geometric interpretations and relate them to expected estimation errors. The information parameters and their corresponding geometric interpretation (for the case of 2 unknown components of $\mathbf{x}(t_0)$), are summarised in Table 1 and Figure 1, where λ_{\max} and λ_{\min} are the largest and smallest eigenvalues of \mathbf{M}_l respectively, $(m_l^{-1})_{ii}$ is the i th element of the main diagonal of $(\mathbf{M}_l)^{-1}$, UNI_i is the uncertainty indicator corresponding to the i th unknown component of $\mathbf{x}(t_0)$, λ_j is the j th eigenvalue of the reduced information matrix and v_{ji} is the i th component of the j th eigenvector of the reduced information matrix.

From Figure 1 it can be concluded that the uncertainty indicators depend on two factors. First, on the angles between the axes of the reduced uncertainty hyperellipsoid and the coordinate axes. These angles are given by the eigenvectors of \mathbf{M}_l and they are an index of compensations between the estimates of the initial state $\mathbf{x}(t_0)$. Secondly, on the lengths of the semiaxes given by the inverse of the PIC.

Information parameters and expected errors in the estimation

In principle, the unknown components of the initial extended state vector $\mathbf{x}(t_0)$ will be able to be estimated uniquely if the information matrix is non singular (Beck and Arnold, 1977).

Table 1 Geometric interpretation of the information parameters

Parameter	Formula	Related to
Information volume	$IV = \sqrt{ \mathbf{M}_I }$	Quantity of joint information
Information eccentricity	$ECC = \sqrt{\text{cond}(\mathbf{M}_I)} = \sqrt{\frac{\lambda_{\max}(\mathbf{M}_I)}{\lambda_{\min}(\mathbf{M}_I)}}$	Distribution of the joint information
Principal information components	$PIC = \{\sqrt{\lambda_1(\mathbf{M}_I)}, \sqrt{\lambda_2(\mathbf{M}_I)}, \dots, \sqrt{\lambda_l(\mathbf{M}_I)}\}$	Quantity of information in the principal directions
Co-ordinate information components	$CIC = \left\{ \frac{1}{\sqrt{(m_i^{-1})_{11}}}, \frac{1}{\sqrt{(m_i^{-1})_{22}}}, \dots, \frac{1}{\sqrt{(m_i^{-1})_{ll}}} \right\}$	Quantity of information about the unknown components of $\mathbf{x}(t_0)$
Uncertainty indicators	$UNI_i = \max \left\{ \frac{1}{\sqrt{\lambda_j}} \cdot v_{ji} \quad j = 1, \dots, l \right\}$	Maximum uncertainty in the unknown components of $\mathbf{x}(t_0)$

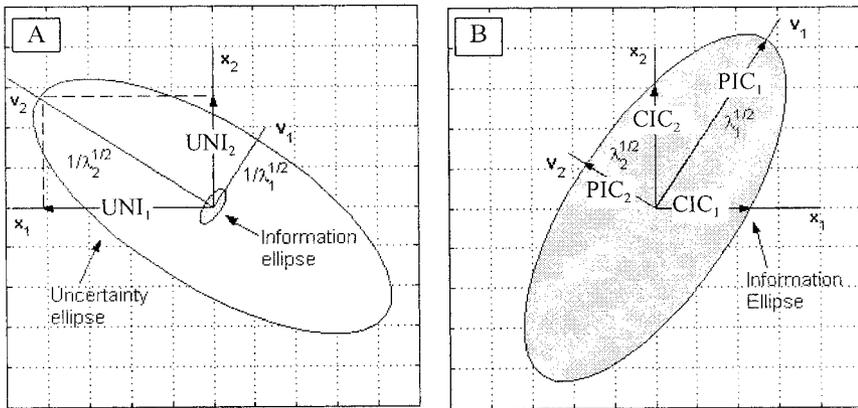


Figure 1 Geometric interpretation of information parameters PIC, CIC and UNI for $l = 2$

But, generally, the measurements are noisy so that identifiability condition is insufficient and it is necessary to analyse the error propagation from the measurements towards the estimates. This analysis can be approached by means of the information parameters defined above. The relation between the geometric characteristics of the information matrix and the characteristic of the estimates are summarised in Table 2 (Hidalgo, 1999).

These parameters are closely related to the well known optimal design criteria (Munack, 1991). In fact IV is related with the D -criterion, the minimum PIC with the E -criterion and ECC with the modified E -criterion. These criteria provide the best experiment from a global point of view of the estimation. But if the experiment is aimed at the estimation of some parameters in particular, although more parameters are assumed to be unknown, design criteria based on CIC and UNI can be proposed.

Applications of the identifiability test

The described mathematical tool can be applied in a very simple manner to any global system (including the nonlinear ones) modelled in state-space. For example, we can use this methodology for the models of the different processes that can be found in the wastewater treatment area like phosphorus removal, anaerobic models, high-load processes, etc.

Table 2 Incidence of the information parameters in the estimation

Parameter	Geometric interpretation	Relation to estimation
<i>CIC</i> (on the directions $\mathbf{x}_1, \mathbf{x}_2$ in Figure 1b)	Lengths of radii of the information hyperellipsoid in the co-ordinate directions	Standard deviation of the estimates (inversely)
<i>PIC</i> (on the directions $\mathbf{v}_1, \mathbf{v}_2$ in Figure 1b)	Inverses of the lengths of the semiaxes of the uncertainty hyperellipsoid	Size of the confidence region of the estimation (inversely)
<i>UNI</i> (on the directions $\mathbf{x}_1, \mathbf{x}_2$ in Figure 1a)	Maximum projections of the semiaxes of the uncertainty hyperellipsoid on the co-ordinate directions	Maximum expected estimation error
<i>IV</i>	Scale factor of the volume of the information hyperellipsoid	Volume of the confidence region (inversely) Geometric mean of the estimation errors (inversely)
<i>ECC</i>	Eccentricity of the information hyper-ellipsoid (ratio between the largest and the smallest semiaxes)	Shape of the confidence region Numerical conditioning of the estimation problem

The identifiability test allows the calculation of the information parameters at each sampling time and to examine them in time. Hence it is recommended for the analysis of the evolution of the concepts underlying these information parameters, like compensations between parameter estimates, expected variance and maximum errors of the estimates, etc. The test also provides a better understanding of the calibration experiments because it gives very useful knowledge about the information flows in the different times of the experiment. Therefore, the test becomes a valuable tool for optimal experimental design.

As an example of application, in this paper we use the proposed methodology to compare the information stored in two typical batch test of calibration of ASM No. 1 and to select which of them is more adequate for the estimation of the maximum specific heterotrophic growth rate μ_H and the initial concentration of heterotrophic biomass X_{BH0} . These experiments are the OUR (Oxygen Uptake Rate) batch-tests at high and medium F/M with F/M the food to micro-organism ratio (Benfield and Randall, 1986).

Results and discussion

This case study has been chosen to illustrate the problem of uncoupling μ_H and X_{BH0} estimates on the basis of OUR measurements. The oxygen uptake rate for high oxygen concentrations is given by the equation,

$$\text{OUR} = \frac{1 - Y_H}{Y_H} \mu_H \frac{S_S}{K_S + S_S} X_{BH} \quad (9)$$

where Y_H is the yield for heterotrophic biomass, K_S is the half-saturation coefficient and S_S is the concentration of readily biodegradable substrate (Henze *et al.*, 1986). From Eq. (9) we can see that the information provided by the OUR measurements concerns the product $\mu_H X_{BH}$ and how this information can be separated for μ_H and X_{BH} is still a matter for discussion. Kappeler and Gujer (1992) proposed the OUR batch-test at high F/M with an initial phase of unlimited heterotrophic growth. But an endogenous phase only dependent on the product $b_H X_{BH}$ could also be useful to estimate μ_H and X_{BH0} . Then it would be interesting to compare quantitatively the information provided by the initial phase of unlimited growth with the information supplied by the endogenous phase and analyse their consequences in the expected estimation errors.

The analysis of these experiments is focused on the estimation of μ_H and X_{BH0} but other initial states and parameters and of ASM No. 1 are assumed unknown: S_{S0} , X_{S0} (initial concentration of slowly biodegradable substrate), K_S , k_h (maximum specific hydrolysis rate) and K_X (half-saturation coefficient for hydrolysis of X_S). So the influence of that lack of knowledge is taken into account in the estimation of μ_H and X_{BH0} . Two cases are considered in both experiments. First, the decay coefficient b_H is assumed to be known and secondly the information is processed assuming b_H to be unknown. Only the CIC values and the error indicators for μ_H and X_{BH0} are presented although the identifiability test calculates the information parameters for all the unknown components. The evolution of CIC and OUR values during the simulation of both experiments are shown in Figure 2.

First of all, examining the OUR evolution from Figure 2 it can be observed that during the first period of the high F/M batch-test (phase A), the OUR increases due to the unlimited heterotrophic growth. Suddenly OUR decreases (phase B), because of limiting concentrations of readily biodegradable substrate, to a second level dominated by the hydrolysis. A new drop to a third level occurs (phase C) when the slowly biodegradable substrate is limiting. The OUR in this third level (phase D) is dominated by the decay process. In an OUR batch-test at medium F/M a first constant OUR is measured (phase A), corresponding to a maximum growth rate. Suddenly OUR decreases as in high F/M batch-test (phase B), and after that the OUR is dominated first by the hydrolysis (phase C), and afterwards by the decay process (phase D).

Analysing the evolution of the co-ordinate information components CIC it is observed that when b_H is known the endogenous phase is shown to be very important in the experiment at medium F/M because in this phase independent information about X_{BH0} is attained through the decay process. So it is possible to uncouple the information about μ_H and X_{BH0} . In high F/M experiments this phase is not important because the uncoupling between μ_H and X_{BH0} has taken place before (in phases A and B). However, when b_H is unknown, the endogenous phase in medium F/M batch-test cannot uncouple the information about μ_H and X_{BH0} because the information provided in this phase concerns the product $b_H X_{BH}$ and there is no independent information about X_{BH0} . In high F/M batch-tests the uncoupling of μ_H and X_{BH0} is possible when information about the hydrolysis and the decay process are supplied in the transition between phases C (hydrolysis) and D (decay).

Comparing the CIC values reached in both experiments it can be concluded that when b_H is known, much more information about μ_H and X_{BH0} is obtained in an OUR batch at medium F/M with a long length of endogenous phase than in an experiment at high F/M with an exponential growth phase. However, when b_H is unknown, more information is

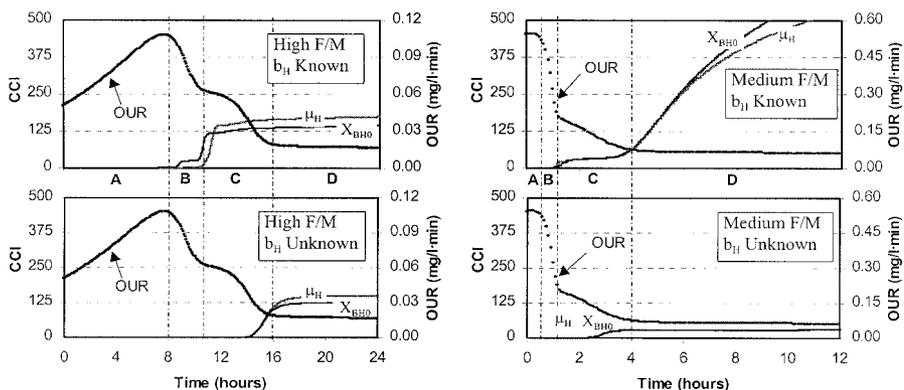


Figure 2 Evolution of CIC values in high and medium OUR batch-tests

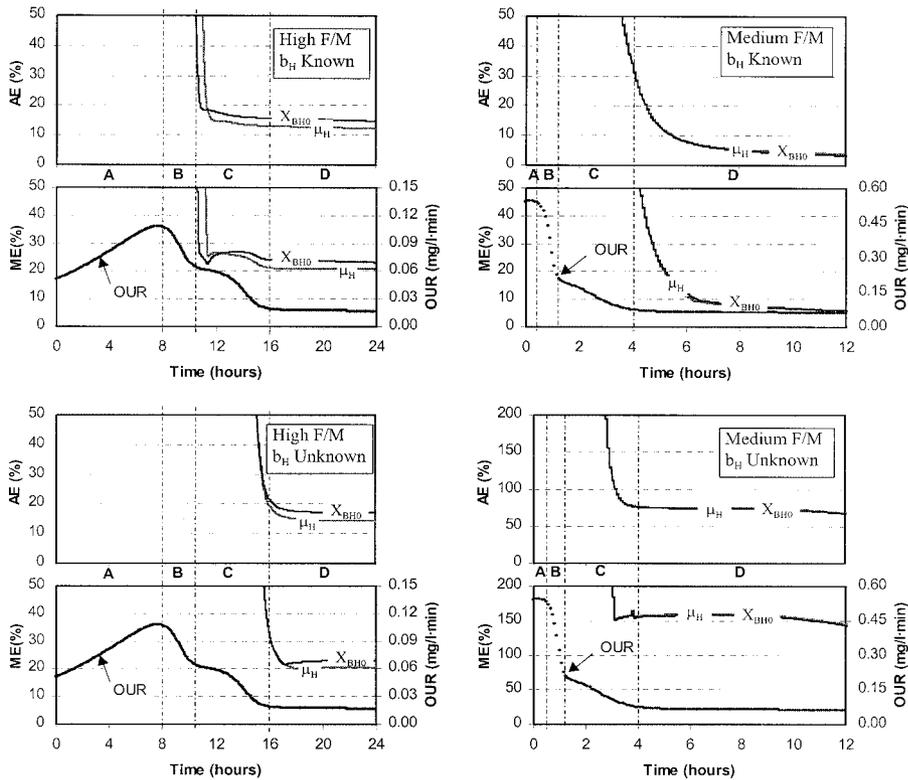


Figure 3 Evolution of standard deviation and maximum error indicators in high and medium OUR batch-tests

received about both parameters in the experiment at high F/M with an exponential growth phase and prolonged up to the endogenous phase.

The developed identifiability test not only allows a qualitative analysis but also predicts medium and maximum errors providing indicators of standard deviation (AE) and maximum error (ME) of the estimates. Next, the evolution of AE and ME indicators during the simulation of the high F/M and medium F/M OUR batch-tests are presented in Figure 3.

Examining the errors shown above it is confirmed what was concluded in the previous analysis of CIC values. When b_H is known the average errors at medium F/M are of the order of 4% and maximum errors around 6%, whereas in high F/M batch-test the AE indicators are about 15% and ME around 22%. Thus, AE and ME indicators at high F/M are of the order of 3–4 times the expected errors in OUR batch-test at medium F/M. However, when b_H is unknown, the expected errors are very high in the experiment at medium F/M (AE around 70% and maximum error around 150%), whereas at high F/M they are approximately the same as when b_H is known. Hence, from this discussion it can be concluded that when b_H is known the best experiment to estimate μ_H and X_{BH0} is the OUR batch-test at medium F/M with a long length of endogenous phase. But if b_H is assumed to be unknown the medium F/M batch-test has to be rejected and the best experiment is the high F/M batch-test with an exponential growth phase and prolonged up to the endogenous phase.

Conclusions

This paper has introduced an identifiability test for state-space models based on the recursive calculation of the information matrix stored in calibration experiments, directly applicable to high non-linear systems and easily implementable in simulators, with a

time-discrete approach. A group of information parameters are also defined on the basis of geometric interpretations of the information matrix. Two of them are of special relevance in the analysis of expected errors in the estimation of unknown parameters of a model. They are the co-ordinate information components (CIC), indices of the expected variance of the estimates, and the uncertainty indicators (UNI), indices of expected maximum errors in the estimation.

As an example of the multiple uses of this kind of mathematical analysis, an application related to optimal experimental design has been presented. Two typical calibration experiments of ASM No. 1 are analysed from the point of view of the estimation of μ_H and X_{BH0} . These are the OUR batch-tests at high and medium F/M. The results obtained reveal that the lack of knowledge about other additional parameters of the Model No.1 (for example b_H) can be decisive in the selection of the best experiment. In this way, when b_H is known, the best experiment to estimate μ_H and X_{BH0} is the OUR batch-test at medium F/M with a long length of endogenous phase. If b_H is assumed unknown the best experiment is the high F/M batch-test with an exponential growth phase and prolonged up to the endogenous phase.

References

- Ayesa, E., Oyarbide, G., Larrea, L. and García-Heras, J.L. (1995). Observability of reduced order models – Application to a model for control of ALPHA process. *Wat. Sci. Tech.*, **31**(2), 161–170.
- Baltes, M., Schneider, R., Sturm, C. and Reuss, M. (1994). Optimal Experimental Design for Parameter Estimation in Unstructured Growth Models. *Biotechnol. Prog.*, **10**(5), 480–488.
- Bard, Y. (1974). *Non-linear Parameter Estimation*. Academic Press Inc., London.
- Beck, J.V. (1989). System Identification and Control. *Dynamic Modelling and Expert Systems in Wastewater Engineering*. Lewis Publishers, Inc., Chelsea, Michigan.
- Beck, J.V and Arnold, K.J. (1977). *Parameter Estimation in Engineering and Science*. J. Wiley and Sons, NY.
- Benefield, L.D. and Randall, C.W. (1980). *Biological Process Design for Wastewater Treatment*. Prentice Hall International, Englewood Cliffs, New Jersey.
- Brouwer, H., Klapwijk, A. and Keesman, K.J. (1998). Identification of Activated Sludge and Wastewater Characteristics using Respirometric Batch-Experiments. *Wat. Res.*, **32**(4), 1240–1254.
- Draper, N. and Smith, H. (1981). *Applied Regression Analysis*. Second Edition. J. Wiley & Sons, New York.
- Henze, M., Grady, Jr. C.P.L., Gujer, W., Marais, G.vR. and Matsuo, T. (1987). *Activated Sludge Model No. 1*. IAWPRC Scientific and Technical Report No. 1. London, UK.
- Hidalgo, M.E. (1999). Desarrollo y verificación de un test de identificabilidad para modelos en espacio de estado. *Aplicación al proceso de fangos activos*. PhD Thesis, ESII, Universidad de Navarra, San Sebastián.
- Holmberg, A. (1982). On the Practical Identifiability of Microbial Growth Models Incorporating Michaelis-Menten Type Nonlinearities. *Math. Biosci.*, **62**, 23–43.
- Kappeler, J. and Gujer, W. (1992). Estimation of kinetics parameters of heterotrophic biomass under aerobic conditions and characterisation of wastewater for activate sludge modelling. *Wat. Sci. Tech.*, **25**(6), 125–139.
- Merkel, W., Schwarz, A., Fritz, S., Reuss, M. and Krauth, K. (1996). New Strategies for Estimating Kinetic Parameters in Anaerobic Wastewater Treatment Plants. *Wat. Sci. Tech.*, **34**(5–6), 393–401.
- Munack, A. (1989). Optimal Feeding Strategy for Identification of Monod-type Models by Fed-batch Experiments. *Computer Applications in Fermentation Technology: Modeling and Control of Biotechnological Processes*. N. Fish, R. Fox and N. Thornhill (eds.). Elsevier, pp. 105–204.
- Munack, A. (1991). Optimization of Sampling. *Biotechnology, a Multi Volume Comprehensive Treatise*, **4**, 251–264.
- Vanrolleghem, P.A., Van Daele, M. and Dochain, D. (1995). Practical Identifiability of a Biokinetic Model of Activated Sludge Respiration. *Wat. Res.*, **29**(11), 2561–2570.
- Walter, E. and Pronzato, L. (1990). Qualitative and Quantitative Experiment Design for Phenomenological Models – A Survey. *Automatica*, **26**(2), 195–213.
- Weijers, S.R. and Vanrolleghem, P.A. (1997). A procedure for Selecting Best Identifiable Parameters in Calibrating Activated Sludge Model No. 1 to Full-Scale Plant Data. *Wat. Sci. Tech.*, **36**(5), 69–79.