Estimation of the yield coefficients in mass-balance based models for bioprocesses: Application to the anaerobic digestion process

B. Chachuat O. Bernard

Abstract – In this report, we investigate the estimation of the yield coefficients matrix in mass-balance based models with application to the anaerobic digestion process. The method is based on the former work of Bernard and Bastin [3]. As a starting point, it is assumed that the reaction scheme is known. The estimation is then twofold: (i) the reaction scheme is first validated from a mass-balance point of vue based for a given data set; (ii) then, the yield coefficients (or combinations) being structurally identifiable for the available measurements are estimated by deriving linear relationships. The procedure is applied to real data sets provided by the EC-TELEMAC project partners. They correspond to either lab scale, pilot scale or industrial scale digesters. Additionaly, two models of increasing complexity are considered, namely the AMH1 and AM2 models.

Contents

1 Introduction **1**

2 Mathematical background **2**
2.1 Determination of the number n_r of reactions 2
2.2 Estimation of the unknown yield coefficients 3
2.3 Extension to more general measurement sets 4

3 Application of the method to the anaerobic digestion process: AMH1 model **5**
3.1 Application to LBE-INRA experimental data: UASB pilot plant 7
3.2 Application to SAUZA experimental data: industrial digester 9
3.3 Application to USC experimental data: UASB and CSTR lab-plants 11

4 Application of the method to the anaerobic digestion process: AM2 model **13**
4.1 Application to LBE-INRA experimental data: UASB pilot plant 15

5 Conclusions **21**

References **22**

1 Introduction

The anaerobic wastewater treatment process presents very atractive features with respect to classical aerobic treatment plants (such as activated sludge plants). This is due to:

- its high capacity to degrade concentrated and difficult substrates;
- its very low sludge production;
- its limited energy requirements;
- and, in some cases, its ability to recover energy from methane combustion.

But in spite of these advantages, anaerobic treatment plants are still very rarely used at the industrial scale, probably beacause they are known to become easily unstable, *e.g.*, when the process operating condition are modified. The use of detailed dynamical models, along with efficient closed loop controllers, is therefore of paramount importance to improve the functionning of the process.

One of the main difficulties consists in designing models whose complexity is not too high (otherwise parameters could not be accuratly estimated and models would be hardly validated), while containing the main mechanisms that are responsible for the observed dynamics. In addition, the lack of phenomenological knowledge, the complexity of the process, its nonlinearities and the lack of sensors makes it imperative to derive models that would be as insensitive as possible to the aforementionned uncertainties by minimizing the number of asumptions in the model-building exercise. In particular, this can be acheived by applying mass-balance considerations which are known to be effective [1]. The dynamic behavior of a stirred tank bioreactor is then described by general models of the following form:

$$\frac{\mathrm{d}\boldsymbol{\xi}}{\mathrm{d}t} = \mathcal{K}\boldsymbol{r}\left(\boldsymbol{\xi}\right) + D\left(\boldsymbol{\xi^{in}} - \boldsymbol{\xi}\right) - \boldsymbol{q}\left(\boldsymbol{\xi}\right) \tag{1}$$

with $\boldsymbol{\xi} \in \mathbb{R}^{n_c}$ begin formed by the concentrations of the various species inside the liquid medium, $\boldsymbol{r} \in \mathbb{R}^{n_r}$ being a vector of the reaction rates, $\mathcal{K} \in \mathbb{R}^{n_c \times n_r}$ being a constant, stoichiometric-like, yield coefficient matrix, $\boldsymbol{q} \in \mathbb{R}^{n_c}$ being the gaseous flow rate representing the loss of mass by the gas phase, and D being the dilution rate.

Matrix $\mathcal{K}$ results from the knowledge of a reaction scheme and plays a key role in the mass-balance modelling. Each column of this matrix corresponds to a chemical or biological reaction, *e.g.*, the coefficients k_{ij} are associated to the j^{th} reaction. By assumption, a positive (resp. negative) coefficient means that ξ_i is a product (resp. substrate) of the j^{th} reaction; a coefficient equal to zero means that ξ_i does not intervene in the j^{th} reaction.

From these consideration, it is clear that designing matrix $\mathcal{K}$ is the first modelling step, and must be addressed prior to the determination of the reaction rates $r(\cdot)$ as functions of the state variables. Recently, Bernard and Bastin [3] proposed a method to guide the identification of matrix $\mathcal{K}$. In particular, it is shown how to determine the size and the best possible structure of this matrix from a set of available measurements, *i.e.*, the number and the configuration of the reactions that must be taken into account).

As a starting point to this work, it is assumed that a given reaction scheme has been chosen among several plausible choices. The objective is then to address the problem of the estimation and validation of the yield coefficients from a set of available data. The report is organised as follow. The mathematical background is recalled in section 2, and the results obtained by Bernard and Bastin [3] are extended to deal with more general measurement sets. The approach is illustrated on two anaerobic digestion models of increasing complexity in sections 3 and 4, namely the (AMH1) and (AM2) models (see *e.g.* [2]). Applications are considered for several experimental data sets which were produced in the scope of the EC-TELEMAC project at the lab-scale (USC), pilot-scale (LBE-INRA) and industrial-scale (Sauza).

2 Mathematical background

In this section, previous results obtained by Bernard and Bastin [3] are reviewed, and then readily extended to deal with more general input/output measurement sets.

Let's consider a bioprocess described by the mass-balance based model (1) with n_c components and n_r independent reactions, and assume that a subset $\boldsymbol{\xi_m}$ of n_m components of $\boldsymbol{\xi}$ is measured, where $n_m > n_r$. If these compounds have gaseous phases, it is also assumed that the associated gaseous flow rates $\boldsymbol{q_m}(\boldsymbol{\xi_m}, \boldsymbol{\xi_{nm}})$ are measured.

The equations describing the dynamics of $\boldsymbol{\xi_m}$ are thus:

$$\frac{\mathrm{d}\boldsymbol{\xi_m}}{\mathrm{d}t} = \mathcal{K}_m \boldsymbol{r}(\boldsymbol{\xi_m}, \boldsymbol{\xi_{nm}}) + D\left(\boldsymbol{\xi_m^{in}} - \boldsymbol{\xi_m}\right) - \boldsymbol{q_m}(\boldsymbol{\xi_m}, \boldsymbol{\xi_{nm}}) \tag{2}$$

where $\mathcal{K}_m \in \mathbb{R}^{n_m \times n_r}$ is the submatrix[1] extracted from $\mathcal{K}$ which is associated to $\boldsymbol{\xi_m}$. From submodel (2), it is clear that only the term $\mathcal{K}_m \boldsymbol{r}(\boldsymbol{\xi_m}, \boldsymbol{\xi_{nm}})$ needs to be mathematically expressed. It consists in (i) estimating matrix $\mathcal{K}_m$, and then (ii) determining the reaction rates as functions of the state variables. Only the former task is addressed in the sequel: the number n_r of reactions is calculated first; then, the $(n_c \times n_r)$ unknowm coefficients of matrix $\mathcal{K}$ (or combinations of these coefficients) are estimated. Both aspects are detailled hereafter.

2.1 Determination of the number n_r of reactions

Integrating equation (2) beween two time instants $t_1 < t_2$ gives:

$$\boldsymbol{u}(t_1, t_2) = \mathcal{K}_m \boldsymbol{v}(t_1, t_2) \tag{3}$$

$$\begin{aligned}
\text{where:}\quad \boldsymbol{u}(t_1, t_2) &= \boldsymbol{\xi_m}(t_2) - \boldsymbol{\xi_m}(t_1) \\
&\quad - \int_{t_1}^{t_2} \left[D(t)\left(\boldsymbol{\xi_m^{in}}(t) - \boldsymbol{\xi_m}(t)\right) + \boldsymbol{q_m}(\boldsymbol{\xi_m}(t), \boldsymbol{\xi_{um}}(t))\right] \mathrm{d}t \\
\boldsymbol{v}(t_1, t_2) &= \int_{t_1}^{t_2} \boldsymbol{r}(\boldsymbol{\xi_m}(t), \boldsymbol{\xi_{um}}(t))\, \mathrm{d}t
\end{aligned}$$

Now, consider N records of $\boldsymbol{u}(\cdot,\cdot)$ and $\boldsymbol{v}(\cdot,\cdot)$:

$$\mathcal{U} = (\boldsymbol{u}(t_1, t_2), \ldots, \boldsymbol{u}(t_N, t_{N+1})) \tag{4}$$

$$\mathcal{V} = (\boldsymbol{v}(t_1, t_2), \ldots, \boldsymbol{v}(t_N, t_{N+1})) \tag{5}$$

where $N > n_r$. Since the reactions are independant and $n_r < n_m$, matrix $\mathcal{V}$ is full rank, *i.e.*, $\mathrm{Rank}(\mathcal{V}) = n_r$. The following result can then be proved using basic linear algebra.

[1] Here, $\mathcal{K}_m$ is assumed to be full rank, otherwise it would mean that the same dynamical behavior can be obtained with a lower dimension matrix $\tilde{\mathcal{K}}_m$ by considering other reaction rates.

Proposition 1 *The number n_r of reactions in the reaction scheme corresponds to the number of non-zero singular values of matrix $\mathcal{U}\mathcal{U}^t$. Moreover, an orthogonal basis of $\mathrm{Im}\,\mathcal{K}_m$ is given by the eigenvectors associated to the n_r non-zero singular values.*

Remarks – Although Proposition 1 gives conditions to determine n_r from a theoretical point of view, its practical application is often perturbed because of model mismatch, measurement noise and frequency, numerical burben, *etc...* Interpolating as well as filtering techniques can be useful to reduce these perturbations. Anyway, there are generally no zero eigenvalues for matrix $\mathcal{U}\mathcal{U}^T$ and the number of reactions is thus determined by selecting the n_r first principal axis which represent a total variance larger than a fixed threshold (*e.g.*, 95 %).

2.2 Estimation of the unknown yield coefficients

From Eqs. (3), (4) and (5), it is obvious that the following linear regression holds for any left kernel vector $\boldsymbol{\lambda} \in \mathbb{R}^{n_m}$ of matrix $\mathcal{K}_m$ (*i.e.*, $\boldsymbol{\lambda}^T\mathcal{K}_m = 0$) satisfying $\lambda_{i_0} = 1$ for an arbitrary $i_0 \in [1, n_m]$:

$$\mathcal{U}_{i_0} = \sum_{\substack{i=1 \\ i \neq i_0}}^{n_m} \lambda_i \mathcal{U}_i \qquad (6)$$

where $\mathcal{U}_j$ denotes the j^{th} rows of matrix $\mathcal{U}$. Note however that these rows are not independent *a priori* since they may be related by the relationships associated to other left kernel vectors of matrix $\mathcal{K}_m$. In particular, the rank condition $\mathrm{Rank}\,(\mathcal{U}) = n_r$ implies that a linear regression (6) cannot involve more than $n_r + 1$ independent terms. It is therefore of major importance to select the left kernel vectors which are associated to independent $\mathcal{U}_j$. In this feature, the following definitions have been introduced by Bernard and Bastin [3].

Definition 1 *A set of transformed data $\{\mathcal{U}_{i_a}, i_a \in [1, n_m]\}$, is said to be* associated to *a left kernel vector λ if $\lambda_{i_a} \neq 0, \forall i_a$.*

Definition 2 *Let $\widetilde{\mathcal{K}}_m \in \mathbb{R}^{n_a \times n_r}$ denote the submatrix extracted from $\mathcal{K}_m$ by removing the $n_m - n_a$ lines whose index j is such that $\lambda_j = 0$, and $\widetilde{\boldsymbol{\lambda}} \in \mathbb{R}^{n_a}$ be the vector obtained by removing all the zero entries of $\boldsymbol{\lambda}$. Then, $\boldsymbol{\lambda}$ is said to be* associated to *matrix $\widetilde{\mathcal{K}}_m$, and $d(\boldsymbol{\lambda}) = \dim\left(\widetilde{\boldsymbol{\lambda}}\right)$ is called the* regression dimension *with respect to $\boldsymbol{\lambda}$. In addition, a left kernel vector $\boldsymbol{\lambda}$ is said to be* sound *if any $(k \times n_r)$ submatrix (with $k < n_a$) of its associated matrix $\widetilde{\mathcal{K}}_m$ is full rank.*

Based on these definitions, the following results hold.

Proposition 2 *(i) A sound vector has* at most n_r+1 *non zero entries, i.e., $d(\boldsymbol{\lambda}) \leq n_r+1$. (ii) A set of transformed data associated to a sound vector cannot be associated to another left kernel vector.*

proof – (i) Consider a left kernel vector $\boldsymbol{\lambda} \in \mathbb{R}^{n_m}$ of matrix $\mathcal{K}_m$ with $n_a > n_r + 1$ non zero entries, and denote $\widetilde{\mathcal{K}}_m \in \mathbb{R}^{(n_a \times n_r)}$ its associated matrix. Since $\mathrm{Rank}\,\widetilde{\mathcal{K}}_m \leq \mathrm{Rank}\,\mathcal{K}_m = n_r$, the rank of any $(n_r + 1 \times n_r)$ submatrix of $\widetilde{\mathcal{K}}_m$ is at most n_r, *i.e.*, the dimension of its left kernel is at least 1 (from the *Rank Theorem*). In other words, the associated matrix $\widetilde{\mathcal{K}}_m \in \mathbb{R}^{(n_a \times n_r)}$ contains $(k \times n_r)$ submatrices $(k < n_a)$ whose rank is not full, which implies that the left kernel vector $\boldsymbol{\lambda}$ is not sound. The proof follows by contraposition.

(ii) Suppose $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \mathbb{R}^{n_m}$ are two left kernel vector associated to a given set of transformed data $\{\mathcal{U}_{i_a}, i_a \in \mathcal{E} \subset [1, n_m]\}$ where $|\mathcal{E}| = n_a \leq n_r + 1$, and denote $\widetilde{\mathcal{K}}_m \in \mathbb{R}^{n_a \times n_r}$ the matrix associated to the non zero entries of these vectors. If $\dim \langle \boldsymbol{\lambda}, \boldsymbol{\lambda}' \rangle = 2$ (*i.e.*, $\dim \operatorname{Ker} \widetilde{\mathcal{K}}_m \geq 2$), then any extracted submatrix of $\widetilde{\mathcal{K}}_m$ has a rank of at most $n_a - 2 < n_r$ (from the *Rank Theorem*) and, consequently, is not full rank. It follows that the left kernel vector $\boldsymbol{\lambda}$ is not sound. The contraposition completes the proof. □

Proposition 3 *The sound vectors span* $\operatorname{Ker} \mathcal{K}_m^t$.

proof – [to be completed]

Corollary 1 *A set of $n_c - n_r$ linearly independent sound vectors form a basis of* $\operatorname{Ker} \mathcal{K}_m^t$.

proof – From the rank theorem, $\dim \operatorname{Ker} \mathcal{K}_m^t = n_m - \operatorname{Rank} \mathcal{K} = n_m - n_r$. Since the sound vectors span $\operatorname{Ker} \mathcal{K}_m^t$, one can thus select $n_m - n_r$ linearly independant sound vectors –note that there are at least $n_m - n_r$ sound vectors, otherwise the sound vector would not span $\operatorname{Ker} \mathcal{K}_m^t$–. The result is obtained by the property that a set of n linearly independant vectors $v_1, \ldots, v_n \in \mathbb{R}^n$ form a basis of $\mathbb{R}^n$. □

Definition 3 *A basis of $\operatorname{Ker} \mathcal{K}_m^t$ formed by $n_c - n_r$ independent sound vectors is called a* sound kernel basis.

The question to be discussed now is to determine which yield coefficients k_{ij} can be identified from all the linear regressions defined by the sound vectors. The identifiability of the yield coefficients when the reaction rates are unknown is usually referred to as the C-identifiability [5]. Bernard and Bastin [3] provide the following theorem.

Theorem 1 *Let $\mathcal{K}$ be a $n_m \times n_r$ full rank matrix with $n_m > n_r$ The unknown element k_{ij} of $\mathcal{K}$ is C-identifiable if there exists a $k \times n_r$ full rank submatrix $\tilde{\mathcal{K}}$, with $k \leq n_r$, such that: (i) $\tilde{\mathcal{K}}$ does not contain any unknown element on its i^{th} column; (ii) the $(k+1) \times p$ submatrix $\begin{pmatrix} \tilde{\mathcal{K}} \\ \mathcal{K}_j \end{pmatrix}$ is of rank k, where $\mathcal{K}_j$ denotes the j^{th} row of $\mathcal{K}$.*

Corollary 2 *The yield coefficient k_{ij} is C-identifiable if and only if it can be computed from any combination of the components of the sound kernel basis* $(\boldsymbol{\lambda}^\star{}_i)_{i=1,\ldots,n_m-n_r}$.

2.3 Extension to more general measurement sets

In the previous subsections, the linear regressions are established by considering a subset $\boldsymbol{\xi}_m$ of n_m measured state variables. For bioprocesses however, the number of measurable state variables is (very) restricted, and the measurements often correspond to linear combinations of state variables, *e.g.*, the soluble COD (COD_s) or the volatile suspended solids (VSS). In addition, some measurements can be directly related to the reaction rates, *e.g.*, the outflow of gases with low solubility such as methane. It is therefore more convenient to base the analysis on the output vector $\boldsymbol{y} \in \mathbb{R}^{n_m}$ defined as:

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y'} \\ \boldsymbol{y''} \end{pmatrix} = \begin{pmatrix} \mathcal{C}'\boldsymbol{\xi} \\ \mathcal{C}''\boldsymbol{r}(\boldsymbol{\xi}) \end{pmatrix}$$

where $n_m = n'_m + n''_m$ is the total number of measurements, $\mathcal{C}' \in \mathbb{R}^{n'_m \times n_c}$ the matrix containing the coefficients of the linear relationships between the measurements $y_1, \dots, y_{n'_m}$ and the state variables $\xi_1, \dots, \xi_{n_c}$, and $\mathcal{C}'' \in \mathbb{R}^{n''_m \times n_r}$ the matrix containing the coefficients of the linear relationships between the measurements $y_{n'_m+1}, \dots, y_{n'_m+n''_m}$ and the reaction rates $r_1, \dots, r_{n_r}$. Note that both matrices $\mathcal{C}'$ and $\mathcal{C}''$ may contain additional (stoichiometric-like) unknown coefficients to be determined by the estimation procedure.

By taking into account the measurement vector $\boldsymbol{y}$, the vector $\boldsymbol{u}(\cdot,\cdot)$ (4) and the linear relationship (3) are modified as follow:

$$\begin{aligned} \boldsymbol{u}(t_1, t_2) &= \begin{pmatrix} \boldsymbol{y'}(t_2) - \boldsymbol{y'}(t_1) - \displaystyle\int_{t_1}^{t_2} D(t)\left(\boldsymbol{y'^{in}}(t) - \boldsymbol{y'}(t)\right) - \mathcal{C}'\boldsymbol{q}(t)\,\mathrm{d}t \\ \displaystyle\int_{t_1}^{t_2} \boldsymbol{y''}(t)\,\mathrm{d}t \end{pmatrix} \\ \boldsymbol{u}(t_1, t_2) &= \begin{pmatrix} \mathcal{C}'\mathcal{K} \\ \mathcal{C}'' \end{pmatrix} \boldsymbol{v}(t_1, t_2) \end{aligned} \tag{7}$$

It is obvious that the results given in Propositions 1, 2, 3 and theorem 1 hold for the linear relationship (7) provided that matrix $\left(\mathcal{K}^t\,\mathcal{C}'^t\ ,\ \mathcal{C}''^t\right)^t$ is full rank.

Several applications of this methodology to the anaerobic digestion process are presented in the next sections.

3 Application of the method to the anaerobic digestion process: AMH1 model

Applications of the previous material are investigated, in this section and the following, for an anerobic digestion process in a stirred tank bioreactor. For the sake of simplicity, we consider first a simple model of the anaerobic digestion process by taking into account a single substrate (COD) and single bacterial population, denoted S_T and X_T respectively. The biological reaction can thus be represented as:

$$k\,S_T \xrightarrow{r(\boldsymbol{\xi})=\mu(\cdot)\,x} X_T$$

where k is a yield coefficient, and $\mu(\cdot)$ represents the bacterial growth rate. It is worth noting that several studies have shown that model (AMH1) has pretty good prediction capabilities, although its simple structure, provided that the process operates in normal conditions (*i.e.*, no destabilisation occurs).

By assuming perfect mixing in the digester (bioreactor), the following simple 2-dimensional dynamical system is obtained:

$$\begin{aligned} \dot{X}_T(t) &= -\alpha\,D(t)\;X_T(t) + \mu(\cdot)\;X_T(t) \\ \dot{S}_T(t) &= -D(t)\left(S_T(t) - S_T^{\mathrm{inf}}(t)\right) - k\,\mu(\cdot)\;X_T(t) \end{aligned} \tag{AMH1}$$

where $\alpha \in [0,1]$ is a coefficient representing the ratio of *unattached* biomass in the digester, $S_T^{\text{inf}}(t)$ is the concentration of substrate in the feeding stream (it is assumed that no biomass is brought by the influent), and $D(t)$ is the dilution rate (*i.e.*, the ratio between the influent flowrate and the volume of the digester). In addition, the methane flow rate q_{CH_4} is defined as follow:

$$q_{\text{CH}_4}(t) \quad = \quad k' \, \mu(\cdot) \; X_T(t)$$

with k' being a yield coefficient.

In addition to the gaseous methane flow rate q_{CH_4} which can be measured on-line, other usual measurements being available in anaerobic digestion processes through off-line analyses are:

- the concentration of dissolved COD (COD_s), which is related to the state variables as, $S_T = \text{COD}_s$;
- the concentration of volatile suspended solids (VSS), that can be expressed as a function of the total biomass concentration as, $\text{VSS} = \alpha X_T$.

From these considerations, the output vector $\boldsymbol{y}$ and the extended yield coefficient matrix $\mathscr{K}$ can be written as:

$$\boldsymbol{y} = \begin{pmatrix} \text{VSS} \\ \text{COD}_s \\ q_{\text{CH}_4} \end{pmatrix} \qquad \mathscr{K} = \begin{pmatrix} \mathcal{C}' \, \mathcal{K} \\ \mathcal{C}'' \end{pmatrix} = \begin{pmatrix} \alpha \\ k \\ k' \end{pmatrix}$$

The calculation of the left kernel vectors basis of this matrix is performed using `MAPLE` and yields:

$$\text{Ker}\mathscr{K}^t = \left\langle \begin{pmatrix} 0 \\ -\frac{k'}{k} \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -\frac{\alpha}{k} \\ 0 \end{pmatrix} \right\rangle \tag{8}$$

Any extracted matrix of the associated matrices of both kernel vectors is obviously full rank (since it is of dimension 1×1), which implies that (8) is a sound kernel basis for model (AMH1). From the application of Corollary 2, it is then straightforward that parameters k, k' and α are C-identifiable provided COD_s, VSS and q_{CH_4} are measured. The following linear relationships can be used to estimate their values:

$$\left[-u_{\text{COD}_s}\right] \frac{k'}{k} \quad = \quad \left[u_{q_{\text{CH}_4}}\right] \tag{9}$$

$$\left[u_{\text{COD}_s}\right] \frac{1}{k} + \left[u''_{\text{VSS}}\right] \frac{1}{\alpha} \quad = \quad \left[-u'_{\text{VSS}}\right] \tag{10}$$

$$\begin{aligned} \text{where:} \quad u_{\text{COD}_s} &= \text{COD}_s(t_2) - \text{COD}_s(t_1) - \int_{t_1}^{t_2} D(t) \left(\text{COD}_s^{\text{inf}}(t) - \text{COD}_s(t)\right) \text{d}t \\ u_{q_{\text{CH}_4}} &= \int_{t_1}^{t_2} q_{\text{CH}_4}(t) \; \text{d}t \\ u'_{\text{VSS}} &= \int_{t_1}^{t_2} D(t) \, \text{VSS}(t) \; \text{d}t \\ u''_{\text{VSS}} &= \text{VSS}(t_2) - \text{VSS}(t_1) \end{aligned}$$

In the case VSS measurements are unavailable, the left kernel of the extended yield matrix $\mathscr{K}$ is one-dimensional and defined as follow:

$$\mathrm{Ker}\mathscr{K}^t = \left\langle \begin{pmatrix} -\frac{k'}{k} \\ 1 \end{pmatrix} \right\rangle$$

Only the parameter combination $\frac{k'}{k}$ is thus C-identifiable, while parameter α is unidentifiable. The former can be estimated by using linear relationship (9).

3.1 Application to LBE-INRA experimental data: UASB pilot plant

The experimental data under consideration is this subsection have been obtained on a UASB pilot plant located at the LBE-INRA in Narbonne, France. The digester is $1\,\mathrm{m}^3$ and is fed with raw industrial wine distillery vinasses (diluted with water). Experimental data corresponding to the dilution rate D, the concentrations COD_s, $\mathrm{COD}_s^{\mathrm{inf}}$ and VSS, and the methane gaseous flow rate q_{CH_4} are available over more than 70 days. Further details about the plant and the experiments can be found in [4].

Before estimating the parameters k, k' and α, one should first study the ability of model (8) to accurately describe the behavior of the digester. This can be done by determining the theoretical number n_r of reactions corresponding to the experimental data (see subsection 2.1): if n_r is equal to 1 then the use of a reaction scheme involving a single reaction is validated, otherwise a more complex reaction scheme migth be necessary. For, we perform an eigenvalue decomposition of the matrix $\mathcal{U}\mathcal{U}^t$ (with scaled values), and then consider the total variance which is associated to each principal axis. We recall that matrix $\mathcal{U}$ is defined as:

$$\mathcal{U} \quad = \quad \begin{pmatrix} u_{q_{\mathrm{CH}_4}}(t_1, t_1 + h_I) & \dots & u_{q_{\mathrm{CH}_4}}(t_N, t_N + h_I) \\ u_{\mathrm{COD}_s}(t_1, t_1 + h_I) & \dots & u_{\mathrm{COD}_s}(t_N, t_N + h_I) \\ u_{\mathrm{VSS}}(t_1, t_1 + h_I) & \dots & u_{\mathrm{VSS}}(t_N, t_N + h_I) \end{pmatrix}$$

where the integral terms in $u_{q_{\mathrm{CH}_4}}$, u_{COD_s} and u_{VSS} are computed over a fixed period of $h_I = 10\,\mathrm{days}$, and the frame is moved by increments of $h_M = t_{i+1} - t_i = 1\,\mathrm{day}$ ($i = 1, \dots, N$). In addition, spline function are used in order to (i) extrapolate the experimental data at any time and, (ii) reduce the measurement noise.

Table 1: Total variance represented by the principal axis of matrix $\mathcal{U}\mathcal{U}^t$, based on the measurements of COD_s, VSS and q_{CH_4}, from the LBE-INRA experimental data.

principal axis	eigenvalue	cumulated variance
1	$0.274\ 10^2$	97.9%
2	$0.392\ 10^0$	99.3%
3	$0.198\ 10^0$	100.0%

The results of the eigenvalue decomposition are displayed in Tab. 1. It can be seen that most of the variance is concentrated in the first principal axis, hence indicating that a reaction scheme involving a single reaction is suitable to describe the behavior of the digester. Recall however that this analysis is based on the mass-balance part of the model only, and no hypothesis was made, at that point, on the mathematical expression of the reaction rate. Therefore, although the hypothesis of a single reaction is validated by the previous analysis, a suitable expression of the reaction rate as a function of the state variable must be chosen to allow good prediction capabilities of the model (*e.g.*, Monod or Haldane kinetic models). These considerations are out of the scope of this work.

As stated previously, the estimation of parameters k, k' and α from the linear relationships (9) and (10). The corresponding regressions are shown in Fig. 1, and the estimated parameters are given in Tab. 2 with their 95% confidence intervals.

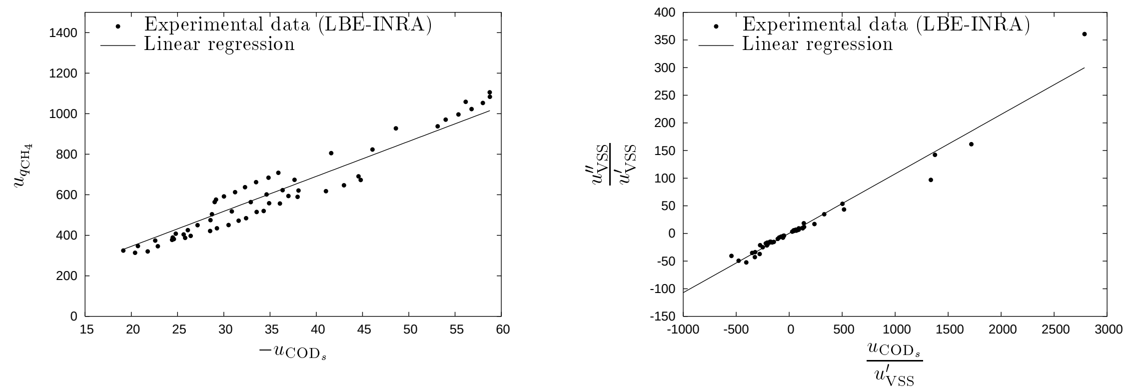

Figure 1: Estimation of the parameters k, k' and α, for model (AMH1), from the LBE-INRA experimental data (UASB pilot plant). *Left plot*: $y = \frac{k'}{k}x$. *Right plot*: $y = \frac{1}{k}x + \frac{1}{\alpha}$.

It can be seen that the estimates of $\frac{k'}{k}$ and $\frac{1}{k}$ are quite confident. Also note that the following estimates for k and k' can be easily deduced from these latter values: $k = 9.3\,\mathrm{g_{COD}}\cdot\mathrm{g_{cell}^{-1}}$, $k' = 160\,\mathrm{mmol_{CH_4}}\cdot\mathrm{g_{cell}^{-1}}$. On the contrary, a very loose confidence is placed on the estimate of $\frac{1}{\alpha}$ (for such confidence intervals, the corresponding estimate is indeed completely unreliable). In particular, we suspect that the use of the VSS measurements as an intermediate to represent the total biomass concentration might be responsible for this uncertainty. But although VSS appears to be a poor indicator of the total concentration of biomass, no alternative mesurements can be used here. In addition, the uncertainty on parameter α can be attributed to the fact that the coefficients u''_{VSS} are non-integral terms, and are therefore much lower than the other coefficients u_{COD_s} and u'_{VSS} whenever the integration period h_I is long (*e.g.*, several days).

Table 2: Estimation of the parameters k, k' and α, for model (AMH1), from the LBE-INRA experimental data (UASB pilot plant).

parameter		value	95% confidence interval
regression (9):			
$\frac{k'}{k}$	17.27	$\mathrm{mmol_{CH_4}}\cdot\mathrm{g_{COD}^{-1}}$	$\pm 2.5\%$
regression (10):			
$\frac{1}{k}$	0.107	$\mathrm{g_{cell}}\cdot\mathrm{g_{COD}^{-1}}$	$\pm 5.4\%$
$\frac{1}{\alpha}$	0.3	–	$\pm 185\%$

3.2 Application to SAUZA experimental data: industrial digester

The same approach is applied in this subsection to estimate the yield coefficients from a set of experimental data produced at the industrial-scale at Sauza (Mexico) and covering a 28-day period (from february 1st to 28th, 2002). The digester consits of a 20 m^3 reactor and is fed with vinasses from local tequilla distilleries. Daily measurements of the influent flow rate and COD concentration are available, along with the concentrations of COD$_s$ and VSS in the effluent and the methane gaseous flow rate q_{CH_4}. These measurements are displayed in Fig. 2.

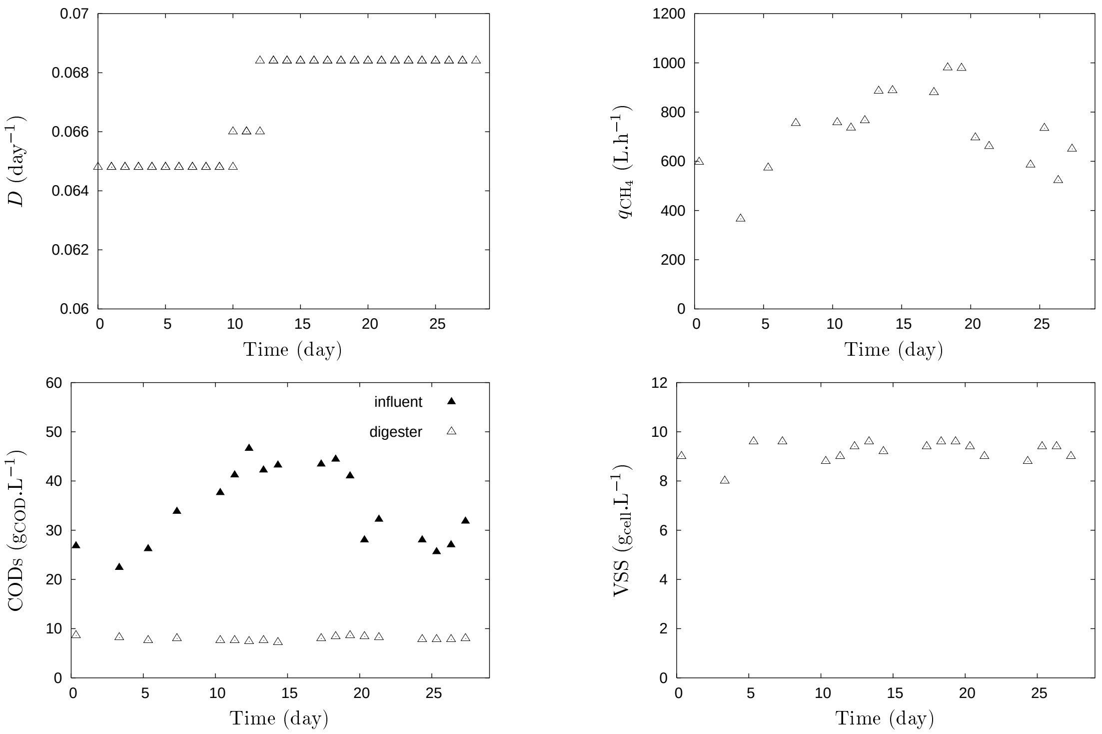

Figure 2: Influent flow rate, COD$_s$ and VSS concentrations, and methane gaseous flow rate for Sauza experimental data (industrial plant).

As previously, the first step consists in checking whether the use of a mass-balance based model involving a single biological reaction is convenient to describe the process. For, an eigenvalue decomposition of the matrix $\mathcal{U}\mathcal{U}^t$ is performed, and the variance associated to each principal axis is considered. These results are displayed in Tab. 3; they correspond to a moving frame of $h_I = 5$ days, being incremented by periods of $h_M = 1$ day.

On the contrary to the results obtained w.r.t. LBE-INRA pilot plant, large errors might be commited here by considering a single step reaction to describe the behavior of Sauza industrial digester. Indeed, it can be seen in Fig. 2 that the methane gaseous flow rate fluctuates between 400 and 1000 L.h^{-1}, while both COD$_s$ and VSS concentrations show little variations at the same time; this is clearly in contradiction with the assumption according to which the methane flow rate can be expressed as $q_{\mathrm{CH}_4} = k' r\,(S_T, X_T)$.

The fits obtained by regressing these experimental data based on the relationships (9) and (10) are pictured in Fig. 3; the corresponding estimates for the parameters (or parameter combinations) are given in Tab. 4.

Table 3: Total variance represented by the principal axis of matrix $\mathcal{U}\mathcal{U}^t$, based on the measurements of COD_s, VSS and q_{CH_4}, from the Sauza experimental data.

principal axis	eigenvalue	cumulated variance
1	$0.151\ 10^1$	71.7%
2	$0.503\ 10^0$	95.7%
3	$0.905\ 10^{-1}$	100.0%

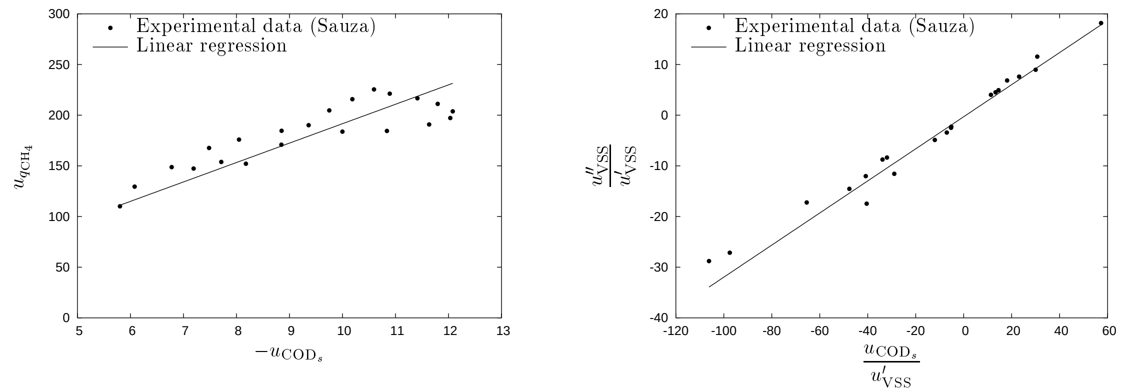

Figure 3: Estimation of the parameters k, k' and α, for model (AMH1), from Sauza experimental data (industrial plant). *Left plot*: $y = \frac{k'}{k}x$. *Right plot*: $y = \frac{1}{k}x + \frac{1}{\alpha}$.

One should first note that the dispersion of the points around the regression line in Fig. 3 (left plot) is markedly larger than in Fig. 1. This is obviouly in accordance to the results obtained from the principal component analysis of matrix $\mathcal{UU}^t$ in both cases. However, the confidence placed on the estimate of $\frac{k'}{k}$ remains reasonable, and so is the confidence on $\frac{1}{k}$. On the contrary, a confident estimate for the parameter α cannot be determined from these relationships and, again, the resulting value is completely unreliable (negative value!). We beleive that the reasons of such an uncertainty are similar to those given in the case of LBE-INRA pilot plant (see subsection 3.1).

Table 4: Estimation of the parameters k, k' and α, for model (AMH1), from Sauza experimental data (industrial plant).

parameter	value		95% confidence interval
regression (9):			
$\frac{k'}{k}$	19.16	$\text{mmol}_{CH_4}\cdot \text{g}_{\text{COD}}^{-1}$	$\pm 4.5\%$
regression (10):			
$\frac{1}{k}$	0.317	$\text{g}_{\text{cell}}\cdot \text{g}_{\text{COD}}^{-1}$	$\pm 8.3\%$
$\frac{1}{\alpha}$	-0.3	–	$\pm 148\%$

3.3 Application to USC experimental data: UASB and CSTR lab-plants

At last, we consider several data sets produced at the lab-scale at USC (University of *Santiago de Compostela*) in Spain. These data correspond to two digesters: a 4.8 L UASB unit and a 2 L CSTR unit. On the one hand, two measurement chronicles are available for the former: from 24th April to 4th June 2002 and from 10th Oct. to 18 dec. 2002; on the other hand, a single chronicle is available for the CSTR unit, from 20th April to 18th Aug. 2002. In any case, these data consist in the influent flow rate, the COD_s concentration and the methane gaseous outflow rate. Note in particular that since the VSS concentration was not measured, the number of parameters that can be estimated is considerably reduced. From the analysis performed in section 3, only the combination $\frac{k'}{k}$ is indeed C-identifiable in this case. Anyway, these data sets are of particular interest since they enable comparisons between (i) different lab-scale plants operated at the same time and fed with similar influents and, (ii) the same unit operated at different periods.

The results of the principal component analysis of matrix $\mathcal{UU}^t$ are given in Tab. 5 for each data set. These results are obtained by considering integration durations of $h_I = 10$ days for the UASB digester and $h_I = 20$ days for the CSTR unit; additionally, the moving frame is incremented by a fixed period of 1 day. Generally speaking, it can seen that the variance associated to the first principal axis is greater than 90%. These figures thus indicate that a single-step reaction scheme is valid for both UASB and CSTR lab plants from a mass-balance point of view. However, one should always keep in mind that this analysis relies on the measurements of q_{CH_4} and COD_s only. The results are indeed much less meaningful than for the LBE-INRA or Sauza data sets since the VSS measurements were unavailable here.

The left kernel of the extended yield matrix $\mathscr{K}$ is here 1-dimensional, and the only

Table 5: Total variance represented by the principal axis of matrix $\mathcal{U}\mathcal{U}^t$, based on the measurements of COD_s and q_{CH_4}, for the USC experimental data (UASB and CSTR lab-plants).

	UASB 24/4/02 → 4/6/02		UASB 10/10/02 → 18/12/02		CSTR 20/4/02 → 18/8/02	
princ. axis	eigenval.	cumul. variance	eigenval.	cumul. variance	eigenval.	cumul. variance
1	$0.611\ 10^1$	91.6%	$0.795\ 10^1$	93.6%	$0.371\ 10^1$	92.5%
2	$0.559\ 10^0$	100.0%	$0.542\ 10^0$	100.0%	$0.317\ 10^0$	100.0%

parameter being identifiable is thus the combination $\frac{k'}{k}$ (see section 3). Unfortunately, no further information can be derived from mass-balance considerations only, and in particular the yield coefficients k and k' cannot be estimated separately in this case. The values of $\frac{k'}{k}$ which are obtained by considering the linear relationship (9) can be found in Tab. 6 for each data set.

Table 6: Estimation of the parameter combination $\frac{k'}{k}$, for model (AMH1), from USC experimental data (UASB and CSTR lab-plants).

parameter	value	95% confidence interval
regression (9): UASB digester, 24/4/02 → 4/6/02		
$\frac{k'}{k}$	11.6 $\text{mmol}_{CH_4}\cdot\text{g}_{COD}^{-1}$	±6.0%
regression (9): UASB digester, 10/10/02 → 18/12/02		
$\frac{k'}{k}$	13.7 $\text{mmol}_{CH_4}\cdot\text{g}_{COD}^{-1}$	±5.1%
regression (9): CSTR digester, 20/4/02 → 18/8/02		
$\frac{k'}{k}$	7.6 $\text{mmol}_{CH_4}\cdot\text{g}_{COD}^{-1}$	±3.4%

Reliable estimates of $\frac{k'}{k}$ are produced in any case since the uncertainties remain limited, around 5% (at a 95% confidence level). It is also noteworthy that the estimated values are close for both UASB data sets, whereas the difference is mardedly larger beween both UASB and CSTR plants. From these results, it is therefore of great interest to consider both UASB data sets simultaneously in order to derive a single estimate for parameter $\frac{k'}{k}$. The following value is obtained (with its associated uncertainty at a 95% confidence level):

$$\frac{k'}{k} = 13.5\ \text{mmol}_{CH_4}\cdot\text{g}_{COD}^{-1}\ \pm 4.0\%$$

The corresponding regression is shown in Fig. 4 (left plot). The regression with respect to the CSTR data set is displayed on the rigth plot.

Parameter $\frac{k'}{k}$ can be interpreted as the methane to COD conversion factor. It is found from the experimental data that its value remains stable in the UASB digester at different times, and is roughly twice larger in the UASB unit than in the CSTR unit. One might

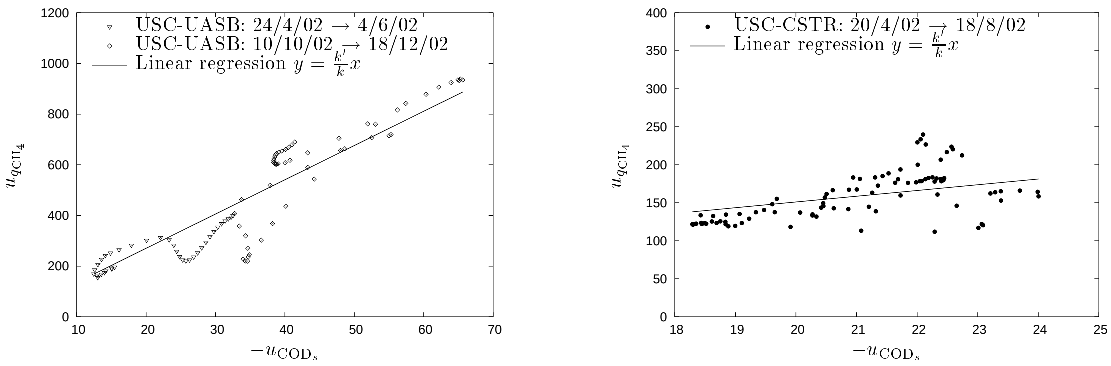

Figure 4: [Estimation of the parameter combination $\frac{k'}{k}$, for model (AMH1), from USC experimental data. *Left plot*: UASB lab plant. *Right plot*: CSTR lab plant.

therefore be tempted to generalise by concluding that the UASB plant is more efficient from the biogas production point of vue. However, the following issues should be raised:

(i) the ratio $\frac{k'}{k}$ strongly depends on the influent characteristics;

(ii) the UASB unit is run in normal conditions, whereas the CSTR unit is operated at a much lower pH; inhibition effect might therefore be responsible for the lower methane production rate in the CSTR digester.

Complimentary results are therefore required to conclude on this point.

4 Application of the method to the anaerobic digestion process: AM2 model

We now address the anaerobic digestion of an organic substrate according to a two-step reaction scheme. In the first step, the organic substrate is converted to volatile fatty acids (VFAs) by means of acidogenic bacteria, and carbon dioxyde CO_2 is produced; VFAs are then degradated to methane CH_4 and CO_2 by methanogenic bacteria in the second step.

$$\text{Acidogenesis:} \quad k_1 S_1 \xrightarrow{r_1(\boldsymbol{\xi})} X_1 + k_2 S_2 + k_4 \mathrm{CO}_2 \tag{11}$$

$$\text{Methanization:} \quad k_3 S_2 \xrightarrow{r_2(\boldsymbol{\xi})} X_2 + k_5 \mathrm{CO}_2 + k_6 \mathrm{CH}_4 \tag{12}$$

where S_1 is the primary substrate (organic substrate expressed as COD_s), S_2 the VFAs, X_1 the acidogenic bacteria, X_2 the methanogenic bacteria and C the total inorganic carbon (TIC) –*i.e.*, the sum of dissolved CO_2 and carbonate HCO_3^- concentrations–.

By considering perfect mixing in the digester, the following 5-dimensional dynamical system is obtained:

$$\begin{aligned}
\dot{X}_1(t) &= -\alpha D(t)\, X_1(t) + \mu_1(\cdot)\, X_1(t) \\
\dot{X}_2(t) &= -\alpha D(t)\, X_2(t) + \mu_2(\cdot)\, X_2(t) \\
\dot{S}_1(t) &= -D(t)\left(S_1(t) - S_1^{\mathrm{inf}}(t)\right) - k_1\,\mu_1(\cdot)\, X_1(t) \\
\dot{S}_2(t) &= -D(t)\left(S_2(t) - S_2^{\mathrm{inf}}(t)\right) + k_2\,\mu_1(\cdot)\, X_1(t) - k_3\,\mu_2(\cdot)\, X_2(t) \\
\dot{C}(t) &= -D(t)\left(C(t) - C^{\mathrm{inf}}(t)\right) + k_4\,\mu_1(\cdot)\, X_1(t) + k_5\,\mu_2(\cdot)\, X_2(t) - q_{\mathrm{CO}_2}
\end{aligned} \tag{AM2}$$

Note that this models falls into the scope of the general dynamic model (1) with the state, reaction rate and gaseous flowrate vectors and the yield coefficient matrix being defined as:

$$\boldsymbol{\xi} = \begin{pmatrix} X_1 \\ X_2 \\ S_1 \\ S_2 \\ C \end{pmatrix} \qquad \boldsymbol{r} = \begin{pmatrix} r_1(\boldsymbol{\xi}) \\ r_2(\boldsymbol{\xi}) \end{pmatrix} \qquad \boldsymbol{q} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ q_{\mathrm{CO}_2} \end{pmatrix} \qquad \mathcal{K} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -k_1 & 0 \\ k_2 & -k_3 \\ k_4 & k_5 \end{pmatrix}$$

In addition to on-line measurements of the CO_2 gaseous flowrate, the other measurements being usually available through off-line analyses are the concentrations of COD_s, VFA, TIC and, in some cases, VSS.

- VFA and TIC measurements can be directly associated to the state variables S_2 and C respectively;
- COD_s is related to the state variables S_1 and S_2 by the linear relationship $\mathrm{COD}_s = S_1 + k_{\mathrm{VFA}}\, S_2$ (with $k_{\mathrm{VFA}} \approx 0.072\,\mathrm{g_{COD}}.\mathrm{mmol}_{\mathrm{VFA}}^{-1}$ in the case VFAs are assimilated to pure acetate);
- VSS measurements can be related to the biomass concentrations as, VSS$=\alpha\,(X_1 + X_2)$, with $0 \leq \alpha \leq 1$ being the fraction of unattached biomass in the digester.

Similarly to model (AMH1), the methane flow rate q_{CH_4} can also be expressed as a function of the reaction rates since it has a low solubility [4]:

$$q_{\mathrm{CH}_4} = k_6\, r_2(\boldsymbol{\xi})$$

where k_6 is an additional yield coefficients whose value is to be estimated.

The measurement vector and the extended yield coefficient matrix are then defined as:

$$\boldsymbol{y} = \begin{pmatrix} \mathrm{VSS} \\ \mathrm{COD}_s \\ \mathrm{VFA} \\ \mathrm{TIC} \\ q_{\mathrm{CH}_4} \end{pmatrix} \qquad \mathscr{K} = \begin{pmatrix} \mathcal{C}'\,\mathcal{K} \\ \mathcal{C}'' \end{pmatrix} = \begin{pmatrix} \alpha & \alpha \\ -k_1 + k_{\mathrm{VFA}} k_2 & -k_{\mathrm{VFA}} k_3 \\ k_2 & -k_3 \\ k_4 & k_5 \\ 0 & k_6 \end{pmatrix}$$

The calculation of the left kernel vectors basis of this matrix, using `MAPLE`, yields:

$$\mathrm{Ker}\mathscr{K}^t = \left\langle \begin{pmatrix} 0 \\ 0 \\ 1 \\ -\frac{k_2}{k_4} \\ -\frac{k_2}{k_4}\frac{k_5}{k_6} - \frac{k_3}{k_6} \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ \frac{k_1 - k_{\mathrm{VFA}} k_2}{k_4} \\ -\frac{k_1 - k_{\mathrm{VFA}} k_2}{k_4}\frac{k_5}{k_6} - k_{\mathrm{VFA}}\frac{k_3}{k_6} \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ -\frac{\alpha}{k_4} \\ -\frac{k_5}{k_6}\frac{\alpha}{k_4} + \frac{\alpha}{k_6} \end{pmatrix} \right\rangle \tag{13}$$

In the sequel, these vectors are denoted $(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3)$ (from left to right). In addition, their associated matrices are denoted $\widetilde{\mathscr{K}_1}$, $\widetilde{\mathscr{K}_2}$ and $\widetilde{\mathscr{K}_3}$:

$$\widetilde{\mathscr{K}_1} = \begin{pmatrix} k_2 & -k_3 \\ k_4 & k_5 \\ 0 & k_6 \end{pmatrix} \qquad \widetilde{\mathscr{K}_2} = \begin{pmatrix} -k_1 + k_{\mathrm{VFA}} k_2 & -k_{\mathrm{VFA}} k_3 \\ k_4 & k_5 \\ 0 & k_6 \end{pmatrix} \qquad \widetilde{\mathscr{K}_3} = \begin{pmatrix} \alpha & \alpha \\ k_4 & k_5 \\ 0 & k_6 \end{pmatrix}$$

Any (2×2) submatrix of $\widetilde{\mathscr{K}_1}$, $\widetilde{\mathscr{K}_2}$ or $\widetilde{\mathscr{K}_3}$ is obviously full rank, which implies that the left kernel vectors $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$ and $\boldsymbol{\lambda}_3$ are sound, and (13) is a sound basis of $\mathrm{Ker}\,\mathscr{K}^t$ for model (AM2). From the application of Corollary 2, it immediatly follows that all the yield coefficients $(k_1, \ldots, k_6)$, as well as parameter α are C-identifiable if COD_s, VFA, TIC, VSS and q_{CH_4} are measured. The subsequent linear relationships, obtained by combining the vectors of the sound kernel basis, can be used to estimate the values of these parameters:

$$\begin{aligned}
\left[u_{\mathrm{COD}_s} - k_{\mathrm{VFA}} u_{\mathrm{VFA}}\right] \frac{k_2}{k_1} + \left[u_{q_{\mathrm{CH}_4}}\right] \frac{k_3}{k_6} &= \left[-u_{\mathrm{VFA}}\right] && (14)\\
\left[-u_{\mathrm{COD}_s} + k_{\mathrm{VFA}} u_{\mathrm{VFA}}\right] \frac{k_4}{k_1} + \left[u_{q_{\mathrm{CH}_4}}\right] \frac{k_5}{k_6} &= \left[u_{\mathrm{TIC}}\right] && (15)\\
\left[u_{\mathrm{COD}_s} - k_{\mathrm{VFA}} u_{\mathrm{VFA}}\right] \frac{1}{k_1} + \left[-u_{q_{\mathrm{CH}_4}}\right] \frac{1}{k_6} + \left[u''_{\mathrm{VSS}}\right] \frac{1}{\alpha} &= \left[-u'_{\mathrm{VSS}}\right] && (16)\\
\left[u_{\mathrm{VFA}}\right] \frac{1}{k_2} + \left[u_{q_{\mathrm{CH}_4}}\right] \left(1 + \frac{k_3}{k_2}\right) \frac{1}{k_6} + \left[-u''_{\mathrm{VSS}}\right] \frac{1}{\alpha} &= \left[u'_{\mathrm{VSS}}\right] && (17)\\
\left[u_{\mathrm{TIC}}\right] \frac{1}{k_4} + \left[u_{q_{\mathrm{CH}_4}}\right] \left(1 - \frac{k_5}{k_4}\right) \frac{1}{k_6} + \left[-u''_{\mathrm{VSS}}\right] \frac{1}{\alpha} &= \left[u'_{\mathrm{VSS}}\right] && (18)
\end{aligned}$$

$$\begin{aligned}
\text{where:}\quad u_{\mathrm{COD}_s} &= \mathrm{COD}_s(t_2) - \mathrm{COD}_s(t_1) - \int_{t_1}^{t_2} D(t) \left(\mathrm{COD}_s^{\mathrm{inf}}(t) - \mathrm{COD}_s(t)\right) \mathrm{d}t\\
u_{\mathrm{VFA}} &= \mathrm{VFA}(t_2) - \mathrm{VFA}(t_1) - \int_{t_1}^{t_2} D(t) \left(\mathrm{VFA}^{\mathrm{inf}}(t) - \mathrm{VFA}(t)\right) \mathrm{d}t\\
u_{\mathrm{TIC}} &= \mathrm{TIC}(t_2) - \mathrm{TIC}(t_1) - \int_{t_1}^{t_2} D(t) \left(\mathrm{TIC}^{\mathrm{inf}}(t) - \mathrm{TIC}(t)\right) - q_{\mathrm{CO}_2}(t)\, \mathrm{d}t\\
u_{q_{\mathrm{CH}_4}} &= \int_{t_1}^{t_2} q_{\mathrm{CH}_4}(t)\, \mathrm{d}t\\
u'_{\mathrm{VSS}} &= \int_{t_1}^{t_2} D(t)\, \mathrm{VSS}(t)\, \mathrm{d}t\\
u''_{\mathrm{VSS}} &= \mathrm{VSS}(t_2) - \mathrm{VSS}(t_1)
\end{aligned}$$

It is worth noting that relationships (14) and (15) are 2-dimensional relationships, whereas (16), (17) and (18) are 3-dimensional since they involve the additional parameter α. We therefore anticipate on the fact that these latter relationships might lead to poor estimates. These considerations will be illustrated in subsection 4.1.

Now if the measurements of VSS are unavailable, then $\dim \mathrm{Ker}\,\mathscr{K}^t = 2$ and a sound kernel basis is given by:

$$\mathrm{Ker}\mathscr{K}^t = \left\langle \begin{pmatrix} 0 \\ 1 \\ -\frac{k_2}{k_4} \\ -\frac{k_2}{k_4}\frac{k_5}{k_6} - \frac{k_3}{k_6} \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ \frac{k_1 - k_{\mathrm{VFA}} k_2}{k_4} \\ -\frac{k_1 - k_{\mathrm{VFA}} k_2}{k_4}\frac{k_5}{k_6} - k_{\mathrm{VFA}}\frac{k_3}{k_6} \end{pmatrix} \right\rangle$$

The C-identifiability analysis shows that only the following yield coefficient combinations can be estimated: $\frac{k_1}{k_4}$, $\frac{k_2}{k_4}$, $\frac{k_3}{k_6}$ and $\frac{k_5}{k_6}$. Their values can be computed, *e.g.*, by considering linear relationships (14) and (15).

4.1 Application to LBE-INRA experimental data: UASB pilot plant

In order to illustrate the methodology in the case of model (AM2), we first consider experimental data produced at the LBE-INRA in Narbonne, France, on a UASB pilot-plant. These data are the same as those used in subsection 3.1 for estimating the yield coefficients of model (AMH1). The available measurements consist of D, $q_{\mathrm{CH_4}}$, $q_{\mathrm{CO_2}}$, COD_s, $\mathrm{COD}_s^{\mathrm{inf}}$, VFA, $\mathrm{VFA}^{\mathrm{inf}}$, $\mathrm{TIC}^{\mathrm{inf}}$, TIC and VSS.

The first step in the parameter estimation procedure consists in validating the use of the two-step reaction scheme (11)–(12) from these measurements. This is done by considering the singular value decomposition of the matrix $\mathcal{U}$ defined as:

$$\mathcal{U} = \begin{pmatrix} u_{q_{\mathrm{CH_4}}}(t_1, t_1 + h_I) & \dots & u_{q_{\mathrm{CH_4}}}(t_N, t_N + h_I) \\ u_{\mathrm{COD}_s}(t_1, t_1 + h_I) & \dots & u_{\mathrm{COD}_s}(t_N, t_N + h_I) \\ u_{\mathrm{VFA}}(t_1, t_1 + h_I) & \dots & u_{\mathrm{VFA}}(t_N, t_N + h_I) \\ u_{\mathrm{TIC}}(t_1, t_1 + h_I) & \dots & u_{\mathrm{TIC}}(t_N, t_N + h_I) \\ u'_{\mathrm{VSS}}(t_1, t_1 + h_I) & \dots & u'_{\mathrm{VSS}}(t_N, t_N + h_I) \end{pmatrix}$$

with the integration duration h_I and the increment h_M being chosen here as $h_I = 10\,\mathrm{days}$ and $h_M = 1\,\mathrm{day}$, and by using spline function in order to extrapolate the discrete measurements at any time. The results of the eigenvalue decomposition of matrix $\mathcal{U}\mathcal{U}^t$ are displayed in Tab. 7. It can be noted that most of the variance is concentrated in the first two principal axis, hence corresponding to a cumulated variance greater than 98%. The use of a two-step reaction scheme is therefore validated from the mass-balance point of view.

Table 7: Total variance represented by the principal axis of matrix $\mathcal{U}\mathcal{U}^t$, based on the measurements of $q_{\mathrm{CH_4}}$, COD_s, VFA, TIC and VSS, for LBE-INRA UASB pilot plant.

principal axis	eigenvalue	cumulated variance
1	$0.387\ 10^{2}$	92.3%
2	$0.266\ 10^{1}$	98.6%
3	$0.504\ 10^{0}$	99.8%
4	$0.537\ 10^{-1}$	100.0%
5	$0.124\ 10^{-1}$	100.0%

We now address the problem of parameter estimation from the experimental data. It should be outlined first that the parameters $k_1, \dots, k_6$ and α are C-identifiable; their values can be estimated, *e.g.*, from the linear relationships (14), (15), (16), (17) and (18). The same values of h_I and h_M are used here to compute these regressions. The resulting estimates are given in Tab. 8.

Regressions (14) and (15) provide rather confident estimates for both parameter combinations $\frac{k_5}{k_6}$ and $\frac{k_3}{k_6}$. These regressions are pictured in Fig. 5. On the contrary, poor estimates are obtained for the combinations $\frac{k_4}{k_1}$ and $\frac{k_2}{k_1}$ (these values correspond to the slopes of the least-square regressions shown in fig. 5). Note that these large uncertainties are partly due to the experimental protocol. Indeed, the methanization step is nearly independent on the former acidogenic step since a large amount of VFAs is brougth by the influent. Consequently, the mass-balance equations for VFA and TIC exhibit a low sensitivity towards the yield coefficients k_2 and k_4. These considerations were confirmed by the sensitivity analysis of the dynamical model. Also note that comparable uncertainties were obtained by Bernard *et al.* [4] based on steady-state measurements only for the same data set; in addition, these authors stress that the composition of substrate might change

Table 8: Estimation of the yield coefficients from the linear relationships (14), (15), (16), (17) and (18), for the LBE-INRA experimental data (UASB pilot plant).

Combination	Estimated value		Standard deviation	95% confidence interval
Regression (14):				
$\frac{k_2}{k_1}$	1.73	$\mathrm{mmol}_{\mathrm{VFA}}.\mathrm{g}_{\mathrm{COD}}^{-1}$	±33.3%	±67.8%
$\frac{k_3}{k_6}$	0.538	$\mathrm{mmol}_{\mathrm{VFA}}.\mathrm{mmol}_{\mathrm{CH_4}}^{-1}$	±2.4%	±4.7%
Regression (15):				
$\frac{k_4}{k_1}$	0.735	$\mathrm{mmol}_{\mathrm{CO_2}}.\mathrm{g}_{\mathrm{COD}}^{-1}$	±61.1%	±122.5%
$\frac{k_5}{k_6}$	0.816	$\mathrm{mmol}_{\mathrm{CO_2}}.\mathrm{mmol}_{\mathrm{CH_4}}^{-1}$	±1.2%	±2.4%
Regression (16):				
$\frac{1}{k_1}$	$-1.13\ 10^{-2}$	$\mathrm{g}_{\mathrm{cell}}.\mathrm{g}_{\mathrm{COD}}^{-1}$	±198.8%	±398.6%
$\frac{1}{k_6}$	$6.48\ 10^{-3}$	$\mathrm{g}_{\mathrm{cell}}.\mathrm{mmol}_{\mathrm{CH_4}}^{-1}$	±7.6%	±15.3%
$\frac{1}{\alpha}$	0.232	–	±91.6%	±183.6%
Regression (17):				
$\frac{1}{k_2}$	$2.11\ 10^{-2}$	$\mathrm{g}_{\mathrm{cell}}.\mathrm{mmol}_{\mathrm{VFA}}^{-1}$	±20.6%	±41.4%
$\left(1+\frac{k_3}{k_2}\right)\frac{1}{k_6}$	$1.68\ 10^{-2}$	$\mathrm{g}_{\mathrm{cell}}.\mathrm{mmol}_{\mathrm{CH_4}}^{-1}$	±13.0%	±26.1%
$\frac{1}{\alpha}$	−0.109	–	±175.5%	±351.8%
Regression (18):				
$\frac{1}{k_4}$	$1.79\ 10^{-2}$	$\mathrm{g}_{\mathrm{cell}}.\mathrm{mmol}_{\mathrm{VFA}}^{-1}$	±39.6%	±79.5%
$\left(1-\frac{k_5}{k_4}\right)\frac{1}{k_6}$	$-8.64\ 10^{-3}$	$\mathrm{g}_{\mathrm{cell}}.\mathrm{mmol}_{\mathrm{CH_4}}^{-1}$	±68.4%	±137.3%
$\frac{1}{\alpha}$	0.061	–	±344.7%	±691.9%

throughout the experiment, hence inducing (possibly large) fluctuations of the yield coefficient k_1. This large uncertainty on parameter k_1 is confirmed elsewhere by the results of regression (16). On the other hand, this latter regression provides a confident estimate for $\frac{1}{k_6}$ and, consequently, for parameter k_6. Likewise, reasonable estimates can be obtained for parameters k_2 and k_4 from regressions (17) and (18). At last, note that the estimates for parameter α are completely unreliable in any regression involving the VSS measurements; these values are indeed outside the physical range $[0,1]$ and their associated 95% confidence intervals exceed 100%.

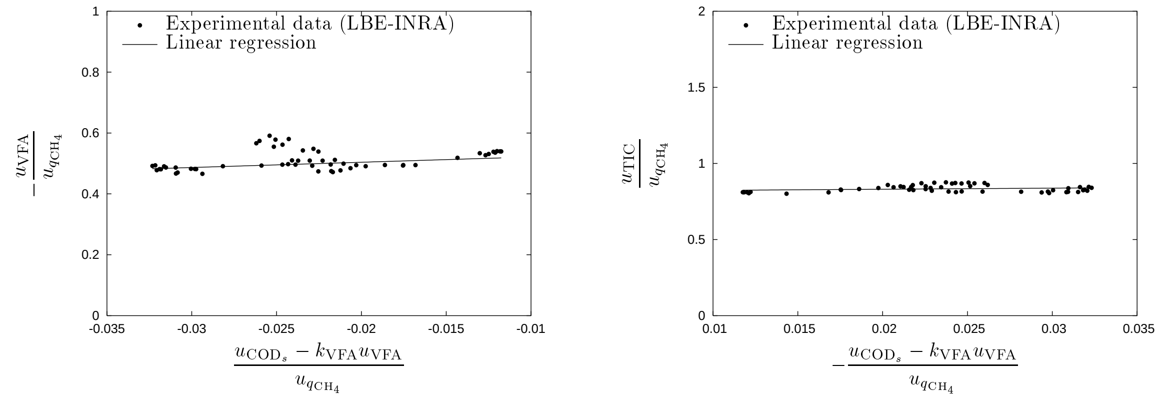

Figure 5: Estimation of the yield coefficients from the linear relationships (14) and (15), for the LBE-INRA experimental data (UASB pilot plant). *Left plot*: $y = \frac{k_2}{k_1}x + \frac{k_3}{k_6}$. *Right plot*: $y = -\frac{k_4}{k_1}x + \frac{k_5}{k_6}$.

By using the previous values of parameter combinations, estimates can be derived for the individual yield coefficients $k_1, \ldots, k_6$, along with their 95% confidence intervals; in the counterpart, reliable estimates could not be obtained for parameter α from mass-balance considerations. The estimates are displayed in Tab. 9 and can be compared to the values obtained by Bernard *et al.* [4] based on steady-state measurements only.

Table 9: Estimates of the yield coefficients $k_1, \ldots, k_6$ for the LBE-INRA experimental data (UASB pilot plant).

Yield coefficient	Estimated value		95% confidence interval	Value from Bernard *et al* [4]	
k_1	27.4	$\mathrm{g_{COD}{\cdot}g_{cell}^{-1}}$	$[19.2; 232.7]$	42.1	$\mathrm{g_{COD}{\cdot}g_{cell}^{-1}}$
k_2	47.3	$\mathrm{mmol_{VFA}{\cdot}g_{cell}^{-1}}$	$[33.5; 80.7]$	116.5	$\mathrm{mmol_{VFA}{\cdot}g_{cell}^{-1}}$
k_3	83.1	$\mathrm{mmol_{VFA}{\cdot}g_{cell}^{-1}}$	$[68.6; 102.7]$	268.0	$\mathrm{mmol_{VFA}{\cdot}g_{cell}^{-1}}$
k_4	55.7	$\mathrm{mmol_{CO_2}{\cdot}g_{cell}^{-1}}$	$[31.0; 272.0]$	50.6	$\mathrm{mmol_{CO_2}{\cdot}g_{cell}^{-1}}$
k_5	125.9	$\mathrm{mmol_{CO_2}{\cdot}g_{cell}^{-1}}$	$[106.5; 152.2]$	343.6	$\mathrm{mmol_{CO_2}{\cdot}g_{cell}^{-1}}$
k_6	154.2	$\mathrm{mmol_{CH_4}{\cdot}g_{cell}^{-1}}$	$[133.8; 182.1]$	453.0	$\mathrm{mmol_{CH_4}{\cdot}g_{cell}^{-1}}$

It can be seen that the confidence which is placed on the estimates of parameter k_2 and, almost, parameters k_1 and k_4 is rather low. In order to improve these estimates, an alternative approach is applied subsequently which consists in deriving lower dimensional regressions than the 3-dimensional regressions (16), (17) and (18) used previously. In this feature, it is assumed that the ratio ν of acidogenic and methanogenic bacteria, defined as $\nu = \frac{X_1}{X_1+X_2}$, remains constant during the operation. Note that this assumption was formerly used by Bernard *et al.* [4] to estimate the yield coefficients of a similar dynamical model based on steady-state measurements only. Taking part of the additional parameter ν, the extended yield coefficient matrix $\mathscr{K}$ is modified as:

$$\mathscr{K} = \begin{pmatrix} \nu\alpha & 0 \\ 0 & (1-\nu)\,\alpha \\ -k_1 + k_{\mathrm{VFA}} k_2 & -k_{\mathrm{VFA}} k_3 \\ k_2 & -k_3 \\ k_4 & k_5 \\ 0 & k_6 \end{pmatrix}$$

and the following linear relationships are derived from its associated sound kernel vectors:

$$\left[u_{\mathrm{COD}_s} - k_{\mathrm{VFA}} u_{\mathrm{VFA}}\right] \frac{1}{\nu k_1} + \left[u''_{\mathrm{VSS}}\right] \frac{1}{\alpha} = \left[u'_{\mathrm{VSS}}\right] \tag{19}$$

$$\left[u_{q\mathrm{CH}_4}\right] \frac{1}{(1-\nu)\,k_6} + \left[u''_{\mathrm{VSS}}\right] \frac{1}{\alpha} = \left[u'_{\mathrm{VSS}}\right] \tag{20}$$

where u'_{COD_s}, u''_{VFA}, $u'_{q\mathrm{CH}_4}$, u'_{VSS} and u''_{VSS} have been defined previously in section 4. Again, an integration horizon $h_I = 10$ days is considered, and moved by increments $h_M = 1$ day. The results are given in Tab. 10 and the regressions are pictured in Fig. 6.

Table 10: Estimation of the yield coefficients k_1 and k_6, along with α, for the LBE-INRA experimental data (UASB pilot plant).

Combination	Value		95% confidence interval
Regression (19):			
$\frac{1}{\nu k_1}$	$2.72\ 10^{-1}$	$\mathrm{g_{cell}} \cdot \mathrm{g_{COD}^{-1}}$	$\pm 8.7\%$
$\frac{1}{\alpha}$	$7.27\ 10^{-2}$	–	$\pm 1187\%$
Regression (20):			
$\frac{1}{(1-\nu)k_6}$	$6.24\ 10^{-3}$	$\mathrm{g_{cell}} \cdot \mathrm{mmol_{CH_4}^{-1}}$	$\pm 4.1\%$
$\frac{1}{\alpha}$	$2.27\ 10^{-2}$	–	$\pm 186\%$

It can be seen that the estimates of νk_1 and $(1-\nu)\,k_6$ are quite confident, which is confimed by the quality of the regression fit. In the counterpart, the estimated values of α are again completely unreliable (these values should be within the range $[0,1]$) due to (i) the slow variations of the biomass concentration in the reactor and, (ii) the fact that u''_{VSS} is a non-integral term whereas any other terms (u'_{COD_s}, u''_{VFA}, $u'_{q\mathrm{CH}_4}$, u'_{VSS}) contain an integral part.

At that point, note that no assumption was made upon the value of the additional parameter ν. A typical value of $\nu = 0.2$ can be found in the litterature [6]). The estimates for both parameters k_1 and k_6 are here computed based on this value, and the confidence intervals are derived for $\nu \in [0.1, 0.3]$. These values can be found in Tab. 9.

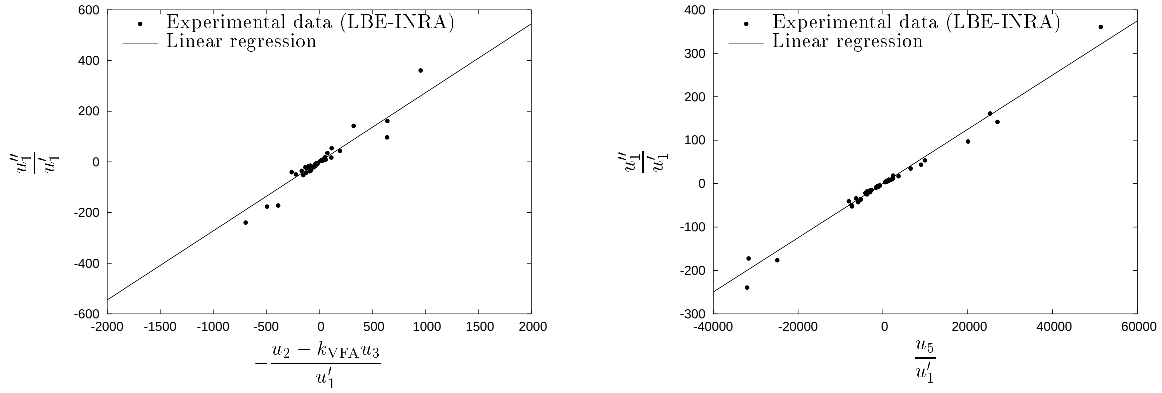

Figure 6: Estimation of the yield coefficients from linear regressions, for the LBE-INRA experimental data (UASB pilot plant). *Left plot*: $y = \frac{1}{\nu k_1} x + \frac{1}{\alpha}$. *Right plot*: $y = \frac{1}{(1-\nu)k_6} x + \frac{1}{\alpha}$.

Table 11: Estimates of the yield coefficients k_1 and k_6 from the assumption of a constant ratio between acidogenic and methanogenic bacteria, for the LBE-INRA experimental data (UASB pilot plant).

Yield coefficient	Estimated value		95% confidence interval
k_1	18.3	$\mathrm{g_{COD}.g_{cell}^{-1}}$	$[11.3; 40.2]$
k_6	200.2	$\mathrm{mmol_{CH_4}.g_{cell}^{-1}}$	$[170.9; 238.6]$

The use of the linear relationship (19) yields a far better estimate for parameter k_1. However, one should keep in mind that the assumption of a constant ratio between acidogenic and methanogenic bacteria might be completely wrong in practice due to either the inherent fluctuations of the influent concentrations or the change in the operating conditions (flow rate, temperature, *etc.*. Anyway, it can be checked that the estimated values of k_1 and k_6 computed from this assumption are compatible with those provided in Tab. 9 (p. 18).

5 Conclusions

In this contribution, the estimation of the yield coefficients matrix in mass-balance based models has been investigated for the anaerobic digestion process. The methodology is based on the former work of Bernard and Bastin [3]. Provided that the reaction scheme is known, the estimation procedure is twofold: (*i*) the reaction scheme is first validated from a mass-balance point of vue for a given data set; (*ii*) then, the yield coefficients (or parameter combinations) being structurally identifiable for the available measurements are estimated by deriving linear relationships.

The procedure has been applied to several real data sets provided by the EC-TELEMAC project partners, and two models of increasing complexity have been considered, namely the AMH1 and AM2 models. In any case, accurate estimates could be obtained for some of the parameters. The results are however faily dependent on the input/output data set.

- The LBE-INRA experimental data have been considered for the estimation of both AMH1 and AM2 models. Since this data set is of high quality, accurate estimates were obtained for all the yield coefficients (except for parameter α, the value of which could not be computed reliably from this procedure). For model AM2 in particular, the estimated coefficients were found in agreement with those obtained by Bernard *et al.* [4] from steady-state considerations.

- The experimental data from Sauza have been considered in the scope of model AMH1 only (the information being insufficient to run model AM2). Again, accurate estimates could be obtained for the yield coefficients, whereas the estimate for parameter α were found completely unreliable.

- At last, several experimental data sets produced at USC for both UASB and CSTR digesters have been considered in the scope of model AMH1. Since no information is available on the biomass concentration for these data, the set of parameters being identifiable is strongly restricted. However, much insight could be deduced from the comparison between (*i*) data sets at different times and, (*ii*) data sets corresponding to different digester configurations. Also note that the estimation of model AM2 was investigated for these data, but no relevant results were obtained in this case. This is mainly due to large inconsistencies between the measurements of pH, partial and total alkalinities and VFAs.

A great deal of research is currently devoted to the design of robust observers to monitor and control the anaerobic digestion process (software sensors), *e.g.*, asymtotic observers, interval observers, *etc...* Since most of these observers rely on mass-balance considerations, the ability to compute accurate values for the yield coefficients (with tight confidence intervals) is obviously of great importance. The accuracy of the estimated state variable is indeed strongly related to the uncertainty on the yield coefficients. In this context, the estimation procedure considered in this study appears as a useful tool to improve the robustness of such observers.

References

[1] G. Bastin and D. Dochain. *On-line estimation and adaptive control of bioreactors.* Elsevier, Amsterdam, The Netherlands, 1990.

[2] O. Bernard, V. Alcazar, J.-L. Gouzé, L. Mailleret, A. Puñal, J. Rodriguez, R. Salazar, J.-P. Steyer, P. Vanrolleghem, and U. Zaher. Design of models for normal working conditions. Technical report, TELEMAC IST 2000-28156, "Telemonitoring and Advanced Telecontrol of High-Yield Wastewater Treatment", Deliverable 3.1a, 2002.

[3] O. Bernard and Bastin G. First step in mass balance modeling: estimation of the yield matrix. Technical report, INRIA-COMORE and UCL-CESAME, 2002.

[4] O. Bernard, Z. Hadj-Sadok, D. Dochain, A. Genovesi, and J.-P. Steyer. Dynamical model development and parameter identification for an anaerobic wastewater treatment process. *Biotechnol. Bioeng.*, 75(4):424–438, 2001.

[5] L. Chen and G. Bastin. Structural identifiability of the yield coefficients in bioprocess model when the reaction rates are unknown. *Math. Biosciences*, 132:35–67, 1996.

[6] J. Sanchez, S. Arijo, M. Muñoz, M. Moriñigo, and J. Borrego. Microbial colonization of different support materials used to enhance the methanogenic process. *Appl. Microbial. Microtechnol.*, 41:480–486, 1994.