Systematic Analysis of Parameter Identifiability for Improved Fitting of a Biological Wastewater Model to Experimental Data

Cristina Sarmiento Ferrero, Qian Chai
Telemark University College, Porsgrunn, Norway

Marta Dueñas Díez
Hydro Corporate Research Center, Porsgrunn, Norway

Sverre H. Amrani
BioTek AS, Porsgrunn, Norway

Bernt Lie
Telemark University College, P.O. Box 203, N-3901 Porsgrunn, Norway

Abstract

In this paper, a method for experimental parameter identifiability analysis is described. The method is applied to a model of a wastewater treatment plant at Duvbacken in Gävle, Sweden. Out of the 45 original parameters in the model, the analysis indicates that with the given experimental conditions, 12 parameters can be identified.

1 Introduction

Water is one of our most precious resources, and wastewater treatment is becoming more and more important in a highly populated, industrialized world. Wastewater contains organic matter and microorganisms, and the microorganisms thus deplete the water for oxygen. Furthermore, nitrogen and phosphorous content leads to algae growth. It is thus of importance to reduce the content of organic matter, as well as nitrogen and phosphorous. There are also safety regulations on the content of microorganisms, as well as on other content in the wastewater. See e.g. Metcalf & Eddy (2003) for an overview of characteristics of wastewater, and treatment of wastewater.

Both for design and operation of wastewater treatment plants, it is of interest to develop models of how the plants transform the feed. One characteristic of wastewater feed to the plant is its highly varying amount and composition, (Olsson & Newell 1999). This implies that dynamic models are highly relevant for wastewater treatment plants. One such series of models are the Activated Sludge Models (ASM), (Henze, Harremoës, la Cour Jansen & Arvin 1996), which have been developed to include a description of organic content and decomposition, microorganisms, nitrogen processes, and phosphorous processes. In particular, ASM2d attempts to describe the important processes involved in the transformation of organic matter, nitrogen, and phosphorous, (Henze, Gujer, van Loosdrecht & Mino 2000).

The ASM models are parametric, and nominal parameter values are suggested in publications. However, the models are based on averages of microorganism and organic matter populations, and wastewater characteristics varies from plant to plant. It is thus necessary to fit the model parameters to experimental data from a specific wastewater stream in order to ensure good predictive properties of a model, (Jeppson 1993), (Brun, Kühni, Siegrist, Gujer & Reichert 2002).

In this paper, we consider the Duvbacken waste water treatment plant in Gävle, Sweden, which is designed to treat the municipal wastewater from 100.000 people (100.000 pe)[1]. This plant has recently been modified to remove phosphorous, nitrogen, and organic matter, primarily using bacteria. In situations where the biology (represented by the bacteria) may struggle to perform according to governmental effluent criteria, chemicals are added as a supplement. The particular emphasis of this paper is to study the identifiability of parameters in a dynamic model of the plant, based on available experimental data.

In section 2 of the paper, a biological wastewater treatment plant at Duvbacken, Gävle, is presented, and a model of the plant is verified. In section 3, the basic principles of parameter identifiability are discussed. In section 4, identifiability analysis for the Duvbacken model is carried out, and some results of model fitting are given. Finally, some conclusions are drawn.

[1] pe = person equivalents

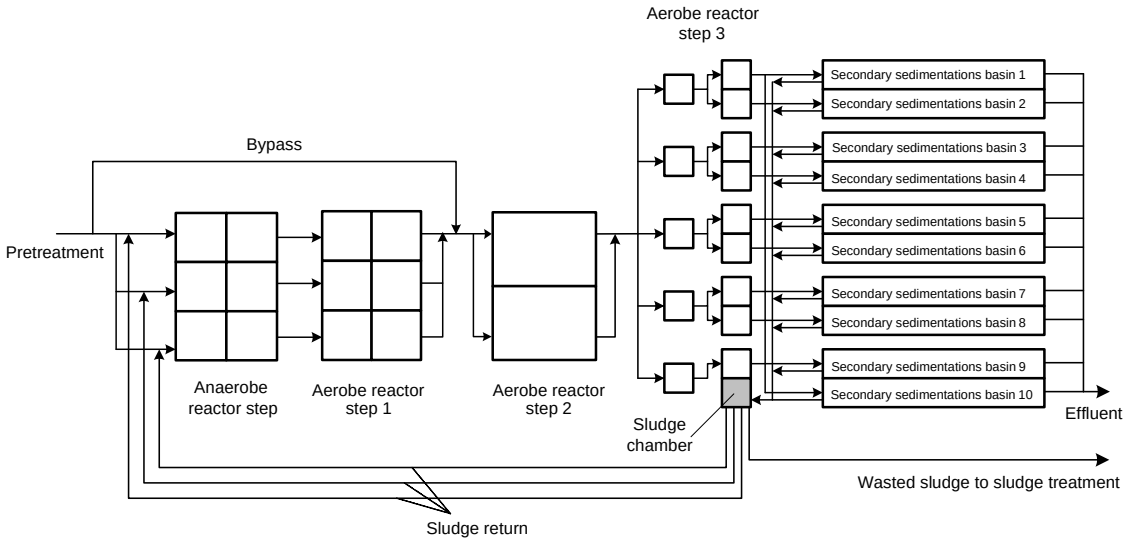

Figure 1: Flowsheet of the Duvbacken Waste Water Treatment Plant.

2 The Duvbacken Plant Model

2.1 Process flow description

Wastewater first undergoes pretreatment involving removal of coarse particles and primary sedimentation of particles. Referring to fig. 1, the pretreated wastewater together with sludge return enters the anaerobe reactor step consisting of three parallel lines, each line consisting of two separate volumes in series, i.e. the anaerobe reactor step consists of a total of 6 separate volumes — each of which is supplied with mixers. The main purpose of the anaerobic reactor step, is to make conditions suitable for biological phosphorous removal.

Aerobe reactor step 1 has the same physical configuration as has the anaerobe reactor step, that is, three parallel lines, each line consisting of two separate volumes. All six volumes are aerated, either by continuous aeration or by intermittent aeration. The air is supplied from three blowers, serving a common air line distributing air to each of the six volumes — which in principle can be controlled independently — where the air enters each volume through about 140 aeration membrane dishes. The outlets from the three parallel lines converge into a common effluent line.

As can be seen from fig. 1, it is also possible to bypass a certain portion of the pretreated wastewater directly to aerobe reactor step 2. The reason for this is that effluent from the pre-treatment is rich in readily biodegradable organic matter. Hence, as to enhance denitrification in aerobe reactor step 2 (if biological nitrogen removal is the goal), bypass of the pretreated wastewater directly to reactor step 2 might be a useful strategy. Aerobe reactor step 2 consists of two rectangular and parallel basins. Each basin is aerated, either continuously or intermittently, and the air is supplied using four ejectors in each basin, distributed equally along the length of the basins.

From aerobe reactor step 2, the outlet lines converge into one effluent line which leads to aerobe reactor step 3. Aerobe reactor step 3 consists of fourteen small volumes, configured as shown in fig. 1. Each volume is aerated, either continuously or intermittently. The air is supplied from blowers, and the air is dispersed in each volume through aeration dishes.

From aerobe reactor step 3, the wastewater flows into ten parallel secondary sedimentation basins. The sludge from these basins, is pumped into a sludge chamber. The sludge chamber has two main purposes. Firstly, sludge is returned back to the inlet of the anaerobe reactor step — one sludge return line for each of the three parallel lines. Secondly, sludge is wasted and pumped to the sludge treatment which involves addition of polymer, thickening etc. The reject water from the thickening process is returned back to the plant.

2.2 Flow descriptions and sensors

The component concentrations in the influent to the plant is typically characterized as follows, Table 1.

Table 1: Description of typical influent concentrations to the Duvbacken Waste Water Treatment Plant.

Parameters	**Description**	**Units**	**Influent to plant**		**After pretreatment**	
			Unfiltered	Filtered	Unfiltered	Filtered
P_{tot}	Total phosphorous	mg/l	6.0	3.2	3.4	1.2
PO_4-P	Phosphate phosphorous	mg/l		3.4		2.0
NO_3-N	Nitrate nitrogen	mg/l		1.0		1.4
NH_4-N	Ammonia nitrogen	mg/l		27.2		
BOD_7	Biological oxygen demand	mg/l	205	69	99	55

Table 2: On-line sensor at the Duvbacken Waste Water Treatment Plant.

Sensor type	**Location**	**Sensors**
Dissolved oxygen S_{O_2}	Aerobe reactor step 1, one sensor in each parallel line.	–
	Aerobe reactor step 2, two sensors in each basin.	y_6
	Aerobe reactor step 3, 5 sensors in 5 different volumes.	–
Nitrate nitrogen S_{NO_3}	Outlet of aerobe reactor step 2, one sensor	y_4
	Sludge chamber, one sensor.	y_8
Ammonia nitrogen S_{NH_4}	Outlet of aerobe reactor 1, one sensor.	y_1
	Effluent of the plant, one sensor.	y_9
Phosphate phosphorous	Effluent of the plant, one sensor.	–
Total phosphorous	Effluent of the plant, one sensor.	–
Dissolved phosphorous S_{PO_4}	Outlet of aerobe reactor step 3, one sensor.	y_{10}
Total solids $X_A + X_H + X_{PAO}$	Outlet of aerobe reactor step 1, one sensor.	y_3
	Outlet of aerobe reactor step 2, one sensor.	
	Sludge chamber, one sensor	y_7
	Effluent of plant, one sensor.	

With reference to Table 1, the following comments should be made:

- The pretreatment reduces the total phosphorous content considerably due to sludge removal from the primary sedimentation basins.
- The pre-treatment reduces the unfiltered biological oxygen demand with 50. The average volumetric influent flow rate is 1400 – 1500 m^3/h.

The plant has to obey the following governmental effluent requirements:

- $BOD_7 < 10\,mg/l$.
- $P_{tot} < 0.5\,mg/l$.

As can be seen, there are currently no governmental requirements regarding the nitrogen contents in the effluent. Table 2 gives an overview of the available on-line sensors of the plant.

In addition, inorganic soluble phosphorous S_{PO_4} is measured by off-line laboratory analysis in aerobe reactors 1, 2, and 3; y_2, y_5, and y_{10}, respectively.

2.3 Model and model verification

A dynamic model is developed, based on the standard kinetics of the ASM2d model, (Henze, Gujer, Mino, Matsuo, Wentzel, Marais & van Loosdrecht 1999), (Metcalf & Eddy 2003), (Henze et al. 2000). In the dynamic model, each of the 4 steps in the plant (anaerobe, aerobe 1–3) is modeled as a perfectly stirred reactor. This leads to 17 states in each of 4 stirred volumes, i.e. 68 states. The model has a total of 45 parameters. The states of the model are grouped into the concentration of soluble species S_j and particulate species X_j. The model has 18 potential input variables: total volumetric flow rate q and the composition of the 17 species. In the model fitting, 10 measurements are available. Examples of measurements are the concentration of suspended solids SS (i.e. the sum of the microorganisms), the concentration of soluble nitrates S_{NO_3}, the concentration of soluble ammonium S_{NH_4}, and the concentration of inorganic soluble phosphorous S_{PO_4}. The model has been implemented in Matlab, and the calling syntax for running the simulation code and compute the outputs for the plant is

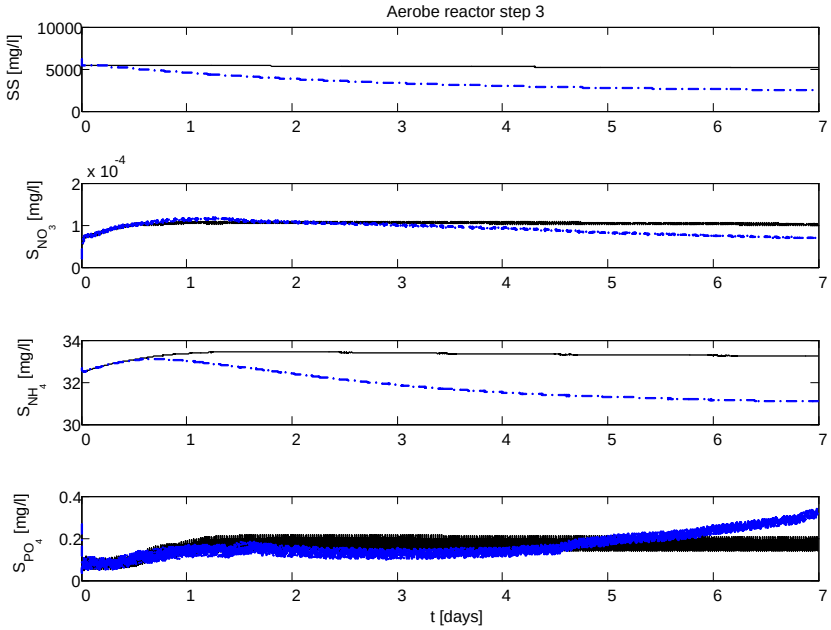

Figure 2: Simulated outputs SS $= X_{\mathrm{A}} + X_{\mathrm{H}} + X_{\mathrm{PAO}}$, $S_{\mathrm{NO_3}}$, $S_{\mathrm{NH_4}}$, and $S_{\mathrm{PO_4}}$ over 7 days, with sludge ages $T_{\mathrm{sludge}} =$ 6 d (solid, black) and $T_{\mathrm{sludge}} = 2$ d (dashed, blue).

```
[T,Y]=bioout(s0,x0,q,Sf,Xf,P0)
```

Here, `s0` and `x0` are the initial state vectors, `q` is the feed flow rate, `Sf` and `Xf` are the feed concentrations of the state variables, and `P0` is the parameter vector. The response from the simulator is `T` which is a vector of time instants where the solution is available, and `Y` which is the matrix of output responses.

Before analyzing the identifiability and estimating parameters, it is a good idea to verify that the model gives reasonable responses, (Olsson & Newell 1999). Figure 2 depicts the concentrations of SS, $S_{\mathrm{NO_3}}$, $S_{\mathrm{NH_4}}$, and $S_{\mathrm{PO_4}}$ for two different sludge ages in the plant — $T_{\mathrm{sludge}} = 6$ d and $T_{\mathrm{sludge}} = 2$ d,[2] during 7 days of operation. The results are reasonable:

- Figure 2 shows that the suspended solids concentration SS is reduced when the sludge age is reduced. This makes sense: less residence time means that less solids is precipitated.
- To observe bio-N operation (nitrogen removal), in general the sludge age should be no less than 6 d. To observe bio-P operation (phosphorous removal), the sludge age should be no less than 2 d. Thus, with a sludge age of 6 d in fig. 2, the system should be on the verge of dropping out of bio-N operation, and with a sludge age of 2 d, the system should be on the verge of dropping out of bio-P operation.
- Although bio-N operation may be obtained with a sludge age of 6 d, in the scenarios simulated in fig. 2, the feed contains no *autotrophic biomass* X_{AUT}, and the concentration of X_{AUT} in the effluent stays below 10^{-2} mg/l for the entire period of 7 d (not shown). To observe nitrification, X_{AUT} should have a value around, say, 300 mg/l. Nitrification would then manifest itself in e.g. a marked decrease in dissolved ammonium $S_{\mathrm{NH_4}}$.
- Phosphorous removal is observed from e.g. the amount of dissolved phosphorous $S_{\mathrm{PO_4}}$. In fig. 2, the concentration of $S_{\mathrm{PO_4}}$ starts to increase (after 4 d) in the case of the 2 d sludge age, while it stays low for the higher sludge age. This indicates bio-P operation at a sludge age of 6 d, while the bio-P operation is failing at a sludge age of 2 d.

There are some oscillations in some outputs in fig. 2 — these are due to oscillating oxygen feed in the aerobe reactors.

[2] The sludge age is the residence time of the sludge in the reactor system.

3 Parameter Identifiability

3.1 System, model, and parameter

For a system $\mathcal{S}$, assume that we have developed a model structure $\mathcal{M}$ with parameter $\theta \in \mathbb{R}^{n_\theta}$; the corresponding model is denoted $\mathcal{M}(\theta)$. The (parametric) model may be based on a mechanistic understanding of the system $\mathcal{S}$, or be a generic mathematical mapping. By varying θ over a feasible set Θ, we get a set of models $\mathcal{M}(\Theta)$, i.e. a model structure $\mathcal{M}$. In reality, the model behavior depends on the experimental conditions $\mathcal{K}$; $\mathcal{K}$ is a description of how an experiment has been carried out, including a sequence of inputs $u_t \in \mathbb{R}^{n_u}$, $t \in \{1,\ldots,T\}$ to, and outputs $y_t \in \mathbb{R}^{n_y}$, $t \in \{1,\ldots,T\}$ from the real system $\mathcal{S}$. With a given model $\mathcal{M}(\theta)$ and experimental conditions $\mathcal{K}$, we can compute a model output $y_t^m(\mathcal{M}(\theta), \mathcal{K})$. We will simplify the notation for the model output to $y_t^m(\theta)$.

With available model structure $\mathcal{M}(\Theta)$ and experimental conditions $\mathcal{K}$, we can attempt to find a specific parameter $\hat{\theta}$ which is such that the model output $y_t^m(\hat{\theta})$ is close to the system output y_t in some sense. To achieve this, we can choose from a set of identification methods I; the estimate $\hat{\theta}$ will also depend on the chosen identification method I. We will simplify the notation and write $y_t^m(\hat{\theta})$ as $\hat{y}_t$.

Usually, it is assumed that the real system $\mathcal{S}$ is an element of the model structure $\mathcal{M}(\Theta)$ and has a true parameter θ^* such that $\mathcal{S}$ is (input-output) identical to $\mathcal{M}(\theta^*)$. It is thus of interest to study whether and to what degree it is possible to find the true system parameters θ^* from the given model structure $\mathcal{M}(\Theta)$, experimental conditions $\mathcal{K}$, and identification method I. We will denote this study an *experimental* parameter identifiability study, since the result depends on the experimental conditions $\mathcal{K}$ (and to some degree on I). Another important study could be whether and to what degree θ^* can be found from $\mathcal{M}(\Theta)$ if we allow *any* (conceivable) experimental conditions $\mathcal{K}$; this we will denote a *theoretical* parameter identifiability study.

Clearly, the set of experimentally identifiable parameters will be a subset of the set of theoretically identifiable parameters. Examples of methods for studying theoretical parameter identifiability, are given in e.g. Pohjanpalo (1978) and Holmberg (1982), and in Ljung & Glad (1994); see also Walter & Pronzato (1997). However, for many industrial processes it will be difficult to get acceptance for carrying out active experiments, and it may be necessary to rely on the natural perturbation in the system. Thus, we may not be able to identify all the theoretically identifiable parameters in practice, and the experimental identifiability may give a better measure of the parameters which can be found.

After a set of identifiable parameters have been found, these can be estimated using some identification technique I. Then, the statistical quality of the model parameters $\hat{\theta}$ and prediction model $\hat{y}_t$ should be studied.

Parameter identifiability and parameter estimation in mechanistic models has been widely studied in in engineering (Bard (1974), Beck & Arnold (1977), Walter & Pronzato (1997), Rawlings & Ekerdt (2002), (Ljung 1999), (Söderström & Stoica 1989),). The systematic approach described in Brun, Reichert & Künsch (2001) seems well suited to Biological Waste Water Models, since it handles large simulation models, and provides identifiability diagnosis for parameter subsets.

3.2 Practical identifiability analysis

We assume that the output $y_t \in \mathbb{R}^{n_y \times 1}$ and the model output $y_t^m(\theta \mid \mathcal{K})$ are related by $y_t = y_t^m(\theta \mid \mathcal{K}) + e_t$, where e_t describes model error/uncertainty. Furthermore, both y_t and u_t are available for $t \in \{1,\ldots,T\}$ (experimental conditions $\mathcal{K}$) and we introduce the notation $y = y^m(\theta \mid \mathcal{K}) + e$ to describe all the resulting equations; y, y^m, $e \in \mathbb{R}^{n_y \cdot T \times 1}$. With a perfect model structure and the correct parameter vector θ^*, the error e will be $e \equiv 0$ and $y \equiv y^m(\theta^* \mid \mathcal{K})$. In the real world, e is unknown. A possible strategy is to set $e = 0$, and hope for a solution θ such that $y = y^m(\theta \mid \mathcal{K})$. We base the analysis on a linearized approximation of this equation, giving $T \cdot n_y$ linear equations in n_θ unknowns:

$$S \cdot \Delta\hat{\theta} = \Delta y, \tag{1}$$

where $\Delta y \triangleq y - y^m(\theta^i \mid \mathcal{K})$, $\Delta\hat{\theta} \triangleq \hat{\theta} - \theta^i$, $S \triangleq \left.\frac{\partial y^m}{\partial \theta}\right|_{\theta=\theta^i}$. S is denoted the model *sensitivity*. If no solution exists, it is common to project Δy on the column space of S, $\Delta y_{\mathcal{R}(S)}$, and solve $S \cdot \Delta\hat{\theta} = \Delta y_{\mathcal{R}(S)}$ — this equation always has a solution, which coincides with the solution of the corresponding normal equation from using the least squares method:

$$S^T S \cdot \Delta\hat{\theta} = S^T \cdot \Delta y. \tag{2}$$

The chosen method for analyzing the identifiability of parameters, is based on Brun et al. (2001), see also Brun et al. (2002), Dueñas Díez, Fjeld, Andersen & Lie (2005). This method is rooted in the linear equation of either eq. 1, or 2. Essentially, these equations have a solution if sensitivity matrix S has full rank. Clearly, if the jth column s_j of matrix S is a zero vector, then matrix S exhibits rank loss and parameter θ_j can not be found. Similarly, if $\|s_j\|$ is "small", we may expect problems in finding θ_j. However, even when all columns of S are significantly different from zero vectors, we may have problems in finding θ if some column is linearly dependent on the other columns. This linear dependence is termed *collinearity*. Two possible measures of collinearity are the condition number κ_j, and the smallest singular value $\sigma_{\min}(S)$.

Scaling of outputs and parameters is essential, and it is recommended to use dimensionless quantities $\tilde{y}_{t,i} = y_{t,i}/y_i^\circ$ and $\tilde{\theta}_i = \theta_i/\theta_i^\circ$. In the sequel, it is assumed that such a scaling has been introduced as part of the preprocessing of the data.

In their method, Brun et al. (2001) consider the *sensitivity measure* δ_j^{msqr} defined as

$$\delta_j^{\text{msqr}} \triangleq \frac{1}{\sqrt{n_y T}} \left\| s_j \right\|_2 . \tag{3}$$

Their strategy is then to sort the various parameters according to the value of δ_j^{msqr}. If there is a marked drop in the value of δ_j^{msqr} for some j, then they propose to assume that those parameters with the smallest sensitivity measure can not be found; hence these parameters are removed from the set of tunable parameters. It should be noted, however, that if the sensitivity measure decreases without marked jumps, and if it is within a couple of decades from the largest value, it may be unjustified to remove a parameter. In typical applications, a set of 10–20 parameters are often retained in the parameter set *after* the sensitivity screening described above.

Their chosen measure of collinearity index $\gamma(\theta)$ is

$$\gamma(\theta) \triangleq \frac{1}{\sigma_{\min}\left(\tilde{S}\right)} = \frac{1}{\sqrt{\sigma_{\min}\left(\tilde{S}^T \tilde{S}\right)}}, \tag{4}$$

where $\tilde{S}$ is a slightly modified sensitivity matrix where the columns have been normalized to have unit norm. It should be noted that this measure is not necessarily a good measure of collinearity, see Stewart (1987) and Belsley (1991).

Collinearity is a measure of how combinations of parameters interact, thus they propose to consider all possible subsets $\mathcal{P}$ of parameters of θ, and compute the collinearity index for all these possible subsets. Let θ_p be one of these subsets, thus $\gamma(\theta_p)$ is to be computed for all possible $p \in \mathcal{P}$.

4 Identifiability Analysis for the Duvbacken Plant Model

4.1 Experimental conditions

With our state space based model, we need a set of experimental input signals u_t in order to carry out the identifiability analysis. Identifiability analysis as discussed in this paper, does not require knowledge of the experimental response y_t from the real system. However, if we later want to estimate the parameters, then the system response is needed.

Here, we use a simulation model implemented in Matlab, with constant influent flow rate and compositions; the main transients in the system are due to initial values that are not at steady state, and an oscillating strategy for oxygen feed. Although the relatively small perturbation of the system is far from ideal for parameter identification, it will serve to illustrate the method for experimental parameter identifiability analysis.

4.2 Model sensitivity

The sensitivity was found by numerically perturbing the simulation model. First, a nominal output was computed, $y_t^m\left(\theta^0\right)$, and reshaped into $y^m\left(\theta^0\right) \in \mathbb{R}^{n_y T \times 1}$. Then the perturbed output $y_t^m\left(\theta^0 + e_j \cdot \delta\theta_j\right)$ was computed and reshaped into $y^m\left(\theta^0 + e_j \cdot \delta\theta_j\right)$; e_j is column j of the identity matrix $I \in \mathbb{R}^{n_\theta \times n_\theta}$. Then finally column s_j of the sensitivity S was computed as

$$s_j \approx \frac{y^m\left(\theta^0 + e_j \cdot \delta\theta_j\right) - y^m\left(\theta^0\right)}{\delta\theta_j}.$$

The outputs and parameters were scaled according to recommendations in Brun et al. (2002).

4.3 Sensitivity index

The sensitivity measure $\delta_j^{\text{msqr}} = \left\| s_j \right\|_2 / \sqrt{n_y T}$ was computed. Out of the 45 original parameters, the 20 most sensitive parameters are shown in Table 3.

Although the decrease in δ_j^{msqr} is gradual, we thus make the deliberate choice of assuming that parameters 21–45 are not identifiable. We thus assign the nominal parameter value to these parameters, and then continue with collinearity analysis to see which of the 20 parameters with highest sensitivity that we can estimate.

4.4 Collinearity indices and identifiable parameters

Similarly as to in Brun et al. (2002), the 20 parameters of interest from the parameter sensitivity ranking can be classified in different groups according to the biological processes to which they belong. The parameters are divided into 4 different groups according to the kind of biological processes they describe: 4 parameters are related to hydrolysis of particulate substrate, 6 parameters are related to heterotrophic organisms, 10 parameters are related to phosphorus-accumulating organisms, and none are related to autotrophic organisms. The distribution of the 20 most sensitive parameters is thus:

Table 3: Parameter importance ranking according to the sensitivity measures. Parameters labeled by an asterisk (*) are considered identifiable from the available data.

δ_j^{msqr}	Parameters		Parameter #
14.812	μ^*_{PAO}	Maximum growth rate of PAO	30
3.9867	q^*_{PP}	Rate constant for storage of X_{PP}	29
3.2385	b_{PAO}	Rate for Lysis of X_{PAO}	32
3.105	K^*_{X}	Saturation coefficient for particulate COD	8
2.9954	K_{h}	Hydrolysis rate constant	4
2.9412	μ^*_{H}	Maximum growth rate on substrate	16
2.9324	q^*_{fe}	Maximum rate for fermentation	17
2.5858	$K^*_{\text{A,PAO}}$	Saturation coefficient for acetate, S_{A}	36
2.2768	b_{PHA}	Rate for Lysis of X_{PHA}	33
2.2196	q^*_{PHA}	Rate constant for storage of X_{PHA} (base X_{PP})	28
1.9906	$K_{\text{A,HET}}$	Saturation coefficient for growth on acetate S_{A}	23
1.9526	K^*_{PHA}	Saturation coefficient for PHA	44
1.326	$K_{\text{O}_2\text{,HYD}}$	Saturation / inhibition coefficient for oxygen	7
1.2238	b^*_{HET}	Rate for Lysis of X_{H}	19
1.1272	$K_{\text{O}_2\text{,HET}}$	Saturation / inhibition coefficient for oxygen	20
1.0583	η^*_{fe}	Anoxic hydrolysis reduction factor	6
0.99162	$K_{\text{O}_2\text{,PAO}}$	Saturation / inhibition coefficient for oxygen	34
0.95771	K_{F}	Saturation coefficient for growth on S_{F}	21
0.52842	$K^*_{\text{PS,PAO}}$	Saturation coefficient for phosphorus in storage of PP	38
0.46989	b^*_{PP}	Rate for Lysis of X_{PP}	45

- Hydrolysis of particulate substrate: K_{X}, K_{h}, $K_{\text{O}_2\text{,HYD}}$, η_{fe}
- Heterotrophic organisms: μ_{H}, q_{fe}, $K_{\text{A,HET}}$, b_{HET}, $K_{\text{O}_2\text{,HET}}$, K_{F}
- Phosphorus-accumulating organisms: μ_{PAO}, q_{PP}, b_{PAO}, $K_{\text{A,PAO}}$, b_{PHA}, q_{PHA}, K_{PHA}, $K_{\text{O}_2\text{,PAO}}$, $K_{\text{PS,PAO}}$, b_{PP}
- Autotrophic organisms: —

Next, the collinearity index $\gamma(\theta_p)$ is calculated for all possible subsets of the top 20 parameters. Values for $\gamma(\theta_p)$ lie in the range 1–11. This is almost within the range of $\gamma(\theta_p) \leq 10$ as proposed by Brun et al. (2002), so our parameters are hardly collinear. However, to illustrate the concept of collinearity, we instead consider parameters to be collinear if $\gamma(\theta_p) > 5$. Here it is found that there are parameter subsets with up to 13 elements which fulfill $\gamma(\theta_p) \leq 5$, whereas subsets with 14 and more elements all have $\gamma(\theta_p) > 5$. Therefore, we consider a maximum of 13 parameters as potentially identifiable from the available data. The subset containing the largest number of parameters with the smallest collinearity index $\gamma(\theta_p) \leq 5$ is selected as the best identifiable; and has collinearity index $\gamma(\theta_p) = 4.27$.

Experience reported in the literature suggest that parameter b_{PAO} is difficult to estimate, even though the model is sensitive to this parameter, (Brun et al. 2002): typically, an unrealistic value of b_{PAO} is found in the parameter estimation. We thus choose to take b_{PAO} out of the set of parameters to estimate. With this simplification, $\gamma(\theta_p)$ drops to 4.20. The chosen final subset of 12 parameters consists of 2 parameters related to hydrolysis of particulate substrate, 3 parameters related to heterotrophic organisms, and 7 related to phosphorus-accumulating organisms; these parameters are indicated with an asterisk in Table 3.

4.5 Parameter estimation

After finding an identifiable parameter subset based on knowledge of $y_t^m\left(\theta^0 \mid \mathcal{K}\right)$, we are ready to estimate the parameters. To do so, we need responses y_t, $t \in \{1,\ldots,T\}$ from the plant. As an initial study before doing parameter estimation based on real data from the Gävle plant, we choose to check whether suitable parameter estimates can be obtained based on responses from the simulation model. Thus, the nominal parameters in the ASM2d plant are chosen as the "correct" parameters θ^*, and then some outputs $y_t\left(\mathcal{K}\right) = y_t^m\left(\theta^* \mid \mathcal{K}\right)$ are computed. The initial parameter guess θ^0 was chosen as a 10% increase of θ^*. Some responses $y_t\left(\mathcal{K}\right)$ are displayed in fig. 2 with a sludge age of 6 d. Clearly, the output indicates a relatively stiff system (compare e.g. the slow variation of S_{NH_4} to the rapid variation of S_{PO_4}), with relatively little excitation. Hence, difficulties may be expected in finding all parameters. Also, as we have seen, most of the identifiable parameters are related

to the phosphorous removal, so we would expect that states related to phosphorous can be fitted better to the data.

In order to find parameter estimates, a weighted least squares criterion is applied

$$\hat{\theta} = \arg\min_{\theta} J = \arg\min_{\theta} e^T W e \tag{5}$$

where $\hat{\theta}$ contains the optimal parameter values of the parameters marked with an asterisk in Table 3 and the other parameters are chosen as θ^0, W is block diagonal with W_t in the diagonal blocks,

$$W_t = \text{diag}\left(1/\left(y_1^\circ\right)^2, ..., 1/\left(y_{n_y}^\circ\right)^2\right),$$

and e is the observation error $y - y^m$, (Dueñas Díez et al. 2005).

To compute the parameter estimates, the nonlinear least squares algorithm `lsqnonolin` of the Optimization Toolbox in Matlab is used. Table 4 shows the initial parameter values θ^0, the "correct" values θ^*, and the estimated values $\hat{\theta}$ obtained after minimizing J.

Table 4: Comparison between initial and estimated parameter values.

Symbol	Description	θ^0	$\hat{\theta}$	θ^*	Unit
μ_{PAO}	Maximum growth rate of PAO	0.0322	0.03156	0.0293	h^{-1}
q_{PP}	Rate constant for storage of X_{PP}	0.0385	0.03876	0.0350	g X_{PP} g^{-1} X_{PAO} h^{-1}
K_{X}	Saturation coefficient for particulate COD	0.1100	0.11537	0.1000	g X_S g^{-1} X_H h^{-1}
μ_{H}	Maximum growth rate on substrate	0.1513	0.16832	0.1375	g X_S g^{-1} X_H h^{-1}
q_{fe}	Maximum rate for fermentation	0.0756	0.03970	0.0688	g S_F g^{-1} X_H h^{-1}
$K_{\text{A,PAO}}$	Saturation coefficient for acetate, S_{A}	4.4000	4.42080	4.0000	g COD m^{-3}
q_{PHA}	Rate constant for storage of X_{PHA} (base X_{PP})	0.0963	0.10698	0.0875	g X_{PHA} g^{-1} X_{PAO} h^{-1}
K_{PHA}	Saturation coefficient for PHA	0.0110	0.01125	0.0100	g X_{PHA} g^{-1} X_{PAO}
b_{HET}	Rate for Lysis of X_{H}	0.0101	0.01049	0.0092	h^{-1}
η_{fe}	Anoxic hydrolysis reduction factor	0.4400	0.45286	0.4000	–
$K_{\text{PS,PAO}}$	Saturation coefficient for phosphorus in storage of PP	0.2200	0.21772	0.2000	g P m^{-3}
b_{PP}	Rate for Lysis of X_{PP}	0.0050	0.00513	0.0046	h^{-1}

Note that for several parameters, $\left|\theta_j^* - \theta_j^0\right| < \left|\theta_j^* - \hat{\theta}_j\right|$. The reason for this is that we are estimating only a subset of the parameters.

4.6 Model validation

Validation is the comparison of model output $y_t^m\left(\hat{\theta} \mid \mathcal{K}'\right)$ with the real output $y_t\left(\mathcal{K}'\right)$ based on training data $\mathcal{K}'$, while parameter estimate $\hat{\theta}$ is based on training data $\mathcal{K}$. The main reason why the validation data $\mathcal{K}'$ should differ from the training data $\mathcal{K}$ is to avoid fitting the model to noise. In this paper, we constructed the training data from an assumed perfect model, $y_t = y_t^m\left(\theta^* \mid \mathcal{K}\right)$, and we choose to use the training data for validation, too. The evolution of some key outputs are displayed in fig. 3: the "real" output $y_t = y_t^m\left(\theta^*\right)$ from the biological reactor, and the prediction outputs $\hat{y}_t = y_t^m\left(\hat{\theta}\right)$. For comparison, we also include the model outputs $y_t^m\left(\theta^0\right)$, which will indicate how much the model has been improved through parameter estimation.

The concentration of S_{PO_4} after parameter estimation (dashed red curve in fig. 3) is quite close to the "real" process output (black, solid line). As argued previously, the model simulates bio-P operation, which can be observed from the time evolution of S_{PO_4}. We also argued that due to the bio-P operation, most of the identifiable parameters are related

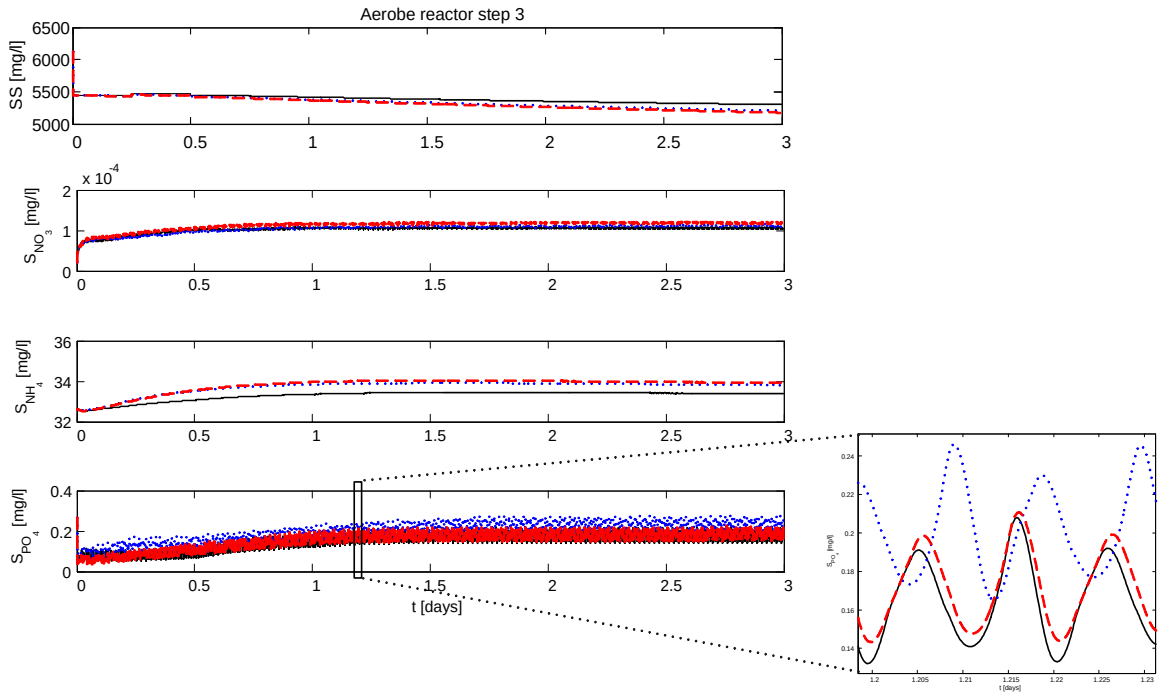

Figure 3: Comparison of model outputs from aerobe reactor step 3: $y_t = y_t^m(\theta^*)$ (solid, black), $\hat{y}_t = y_t^m(\hat{\theta})$ (dashed, red), $y_t^m(\theta^0)$ (dotted, blue).

to phosphorous removal. Furthermore, more weight has been put on phosphorous removal outputs because of the governmental restrictions on total phosphorous in the effluent. Thus, the model has mainly been adjusted to give good prediction of those measurements which are related to phosphorous.

The prediction capabilities of other outputs (e.g. S_{NH_4} and S_{NO_3}) are poorer. As argued, the lack of autotrophic biomass (X_{AUT}) in the process leads to poor nitrification. Consequently, parameters related to the autotrophic (nitrifying) organism group are poorly identifiable. However, since there are no government restrictions on the nitrogen outputs, the poor prediction capability of nitrogen outputs is acceptable.

5 Conclusions

In this paper, a method for assessing parameter identifiability of large scale models based on experimental data is discussed. The method is applied to a model of a biological wastewater treatment plant. First, an overview of the Duvbacken wastewater treatment plant in Gävle, Sweden, is given. Next, an overview is given of a method for analyzing experimental identifiability which has been proposed in the literature. Finally, the identifiability analysis theory is applied to a model of the Duvbacken plant. It should be emphasized that because real experimental data were unavailable, the study has been based on simulated data assuming a perfect model. However, the procedure can be used without modification on real data.

In the analysis, a model with 45 parameters has been studied. From the experimental identifiability analysis, 12 of the parameters were found to be identifiable. Some details in the analysis are given. Then these 12 parameters are used in a parameter estimation study, and improved parameters are found using a Matlab routine for nonlinear least squares minimization. After finding the parameters, the model is validated against the original experimental data. The model fit is not perfect. However, the results are logical from an understanding of the process operation: the model fit is quite good for those outputs that matter in the operation of the plant, and less good in outputs which are less important to the plant operation.

Some future work may include:

- The current model implementation in Matlab is relatively rigid with respect to possible input signals, and a rewriting of the model in a more flexible modeling language such as Modelica is planned.
- Verifying the model more thoroughly for other operating scenarios.
- Fitting the model to experimental data for larger initial errors in parameters, and comparing different optimization algorithms, as well as multiple shooting vs. single shooting (used here).

- Fitting the model to noisy data/real experimental data.
- Including a statistical analysis of the parameters and prediction capabilities of the model.

References

Bard, Y. (1974), *Nonlinear Parameter Estimation*, Academic Press, New York.

Beck, J. V. & Arnold, K. J. (1977), *Parameter Estimation in Engineering and Science*, Wiley, New York.

Belsley, D. A. (1991), *Conditioning Diagnostics. Collinearity and Weak Data in Regression*, John Wiley & Sons, New York.

Brun, R., Kühni, M., Siegrist, H., Gujer, W. & Reichert, P. (2002), 'Practical identifiability of ASM2d parameters - systematic selection and tuning of parameter subsets', *Water Research* **36**(16), 4113–4127.

Brun, R., Reichert, P. & Künsch, H. R. (2001), 'Practical identifiability analysis of large environmental simulation models', *Water Resour. Res.* **37**(4), 1015–1030.

Dueñas Díez, M., Fjeld, M., Andersen, E. & Lie, B. (2005), 'Validation of a compartmental population balance model of an industrial leaching process: The Silgrain(R) process', *Chem. Engng. Sci.* **X**(Y), xx–yy.

Henze, M., Gujer, W., Mino, T., Matsuo, T., Wentzel, M. C., Marais, G. V. R. & van Loosdrecht, M. C. M. (1999), 'Activated sludge model no. 2d, ASM2D', *Wat. Sci. Tech.* **39**(1), 165–182.

Henze, M., Gujer, W., van Loosdrecht, M. C. M. & Mino, T. (2000), *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3.*, Vol. 9 of *Scientific and Technical Report*, IWA Publishing.

Henze, M., Harremoës, P., la Cour Jansen, J. & Arvin, E. (1996), *Wastewater Treatment. Biological and Chemical Processes*, second edn, Springer, Berlin.

Holmberg, A. (1982), 'On the parameter identifiability of microbial growth models incorporating michaeles-menten type nonlinearities', *Math. Biosci.* **62**, 23–43.

Jeppson, U. (1993), On the verifiability of the activated sludge system dynamics, Tekn. lic. dissertation, Lund Institute of Technology.

Ljung, L. (1999), *System Identification. Theory for the User*, second edn, Prentice Hall, Upper Saddle River, New Jersey.

Ljung, L. & Glad, S. T. (1994), 'On global identifiability for arbitrary model parameterizations', *Automatica* **30**(2), 265–276.

Metcalf & Eddy, I. (2003), *Wastewater Engineering. Treatment and Reuse*, fourth edn, McGraw Hill, New York.

Olsson, G. & Newell, B. (1999), *Wastewater Treatment Systems. Modelling, Diagnosis and Control*, IWA Publishing, London.

Pohjanpalo, H. (1978), 'System Identifiability Based on the Power Series Expansion of the Solution', *Mathematical Biosciences* **41**, 21–33.

Rawlings, J. B. & Ekerdt, J. G. (2002), *Chemical Reactor Analysis and Design Fundamentals*, Nob Hill Publishing, Madison, Wisconsin.

Söderström, T. & Stoica, P. (1989), *System Identification*, Prentice Hall International.

Stewart, G. (1987), 'Collinearity and least squares regression', *Statistical Science* **2**(1), 68–84.

Walter, E. & Pronzato, L. (1997), *Identification of Parametric Models from Experimental Data*, Springer.