Chapter 8

State Estimation of the Anaerobic Digestion Process

8.1 Introduction

The anaerobic digestion process depends on the population and vitality of different biomass species. Therefore, almost all dynamic models define the concentration of at least one biomass population in their state vector (at least all dynamic models reviewed in Gerber (2009) and Wolf (2013)). Although there are approaches to measure biomass concentration (Davey et al., 1993, Ferreira et al., 2005) on biogas plants it usually is not measured online yet. Therefore it has to be estimated. More complex models such as the ADM1 define many more state vector components (see Table 7.1) where most of them cannot be measured online as well or where measuring them on- or offline is too expensive, cf. Spanjers and van Lier (2006).

In this chapter (Sec. 8.2) the state estimator introduced in Section 4.1 is applied to the simulation model developed in Section 7.4. Similar results were also published in Gaida et al. (2012b) in the course of this thesis. The developed state estimator is needed in the optimal feed control introduced in Chapter 9. Simulation results of the control using the state estimator can be found in Section 9.3.6.

In the past various different state estimation methods were applied to anaerobic digestion processes. Among them are an observer based estimator based on a variable structure model (Morel et al., 2006a,b), a mass balance based estimator (Bernard et al., 2000), extended Kalman filter (Jones et al., 1989, 1992, Polster, 2009), robust interval observer (Montiel-Escobar et al., 2012), fuzzy estimator (Polit et al., 2001, Carlos-Hernandez et al., 2009), adaptive observer (Rodriguez et al., 2011) and recurrent neural networks observer (Urrego-Patarroyo et al., 2008). In Alcaraz-González and González-Álvarez (2007) an excellent review on observer design for anaerobic digestion processes is given.

8.2 State Estimation using Software Sensors

The state estimation approach introduced in Section 4.1, proposed to only use input values $\boldsymbol{u}$ and output values $\boldsymbol{y}$ of the anaerobic digestion process to estimate its state $\hat{\boldsymbol{x}}$ (see eq. (4.10)). Here, the time t dependent input vector function $\boldsymbol{u}$ is defined by the volumetric flow rates $Q_{\mathrm{substrate}}$ of the $n_{\mathrm{u}} = 4$ available substrates, which are measured in $\frac{\mathrm{m}^3}{\mathrm{d}}$, that is

$$\boldsymbol{u} := \left(Q_{\mathrm{maize}}, Q_{\mathrm{manure}}, Q_{\mathrm{grass}}, Q_{\mathrm{ccm}}\right)^T \tag{8.1}$$

It shall be assumed that the volumetric flow rates of these substrates are measured with a sampling rate of $\delta_{\mathrm{u}} = 6$ h. The physical and chemical parameters of the substrates are assumed to be constant, so that the developed estimator only yields reliable results for substrate characteristics the estimator has learned during training.

The output vector function $\boldsymbol{y}$ is composed of the simulated pH values inside the two digesters $(\mathrm{pH}_1, \mathrm{pH}_2)$, the produced biogas volumetric flow rates $(Q_{\mathrm{gas},1}, Q_{\mathrm{gas},2})$ and the relative amount of methane and carbon dioxide $(r_{\mathrm{ch}_4,1}, r_{\mathrm{co}_2,1}, r_{\mathrm{ch}_4,2}, r_{\mathrm{co}_2,2})$ in the produced biogas (eq. (7.4)). Thus, in total there are $n_{\mathrm{y}} = 8$ measurement variables, four for each digester:

$$\boldsymbol{y} := \left(\underbrace{\mathrm{pH}_1, Q_{\mathrm{gas},1}, r_{\mathrm{ch}_4,1}, r_{\mathrm{co}_2,1}}_{\text{primary digester}}, \underbrace{\mathrm{pH}_2, Q_{\mathrm{gas},2}, r_{\mathrm{ch}_4,2}, r_{\mathrm{co}_2,2}}_{\text{secondary digester}}\right)^T \tag{8.2}$$

These measurements are assumed to be measured with a sampling rate of $\delta_{\mathrm{y}} = 6$ h as well.

It is important to note that output vector function $\boldsymbol{y}$ and input vector function $\boldsymbol{u}$ were chosen deliberately so that they contain process parameters, which are measured in practice on almost every biogas plant.

The current state estimate $\hat{\boldsymbol{x}}(t_k)$ is calculated out of the current input and output values as well as their moving averages, see eq. (4.10). The settings for the moving average filters are summarized in Table 8.1. It can be seen that $N_{\mathrm{u}} = 5$ moving average filters for the inputs are used and $N_{\mathrm{y}} = 7$ for the outputs.

Table 8.1: Settings of moving average filters for input and output values. For the definitions of the moving average filters see eqs. (4.6) and (4.8).

$N_{\mathrm{u}} = 5$	$i_{\boldsymbol{\Lambda}_{\mathrm{u}}} = 1, \dots, N_{\mathrm{u}}$	1	2	3	4	5		
	$w_{\mathrm{u}, i_{\boldsymbol{\Lambda}_{\mathrm{u}}}}$	12 h	1 d	3 d	7 d	14 d		
$N_{\mathrm{y}} = 7$	$i_{\boldsymbol{\Lambda}_{\mathrm{y}}} = 1, \dots, N_{\mathrm{y}}$	1	2	3	4	5	6	7
	$w_{\mathrm{y}, i_{\boldsymbol{\Lambda}_{\mathrm{y}}}}$	12 h	1 d	3 d	7 d	14 d	21 d	31 d

To create the measurement matrix $\boldsymbol{Y}$ in total 120 simulations each lasting 950 days were performed with randomly varying substrate mixtures (defined by $\boldsymbol{u}$), leading to $N = 456,000$ samples (see eq. (4.11)). With the above defined numbers for N_u and N_y the second dimension of the matrix $\boldsymbol{Y}$ is given as $D = 88$, see eq. (4.11).

The values of each substrate flow were restricted to remain between a lower and an upper bound as can be seen in the left part of Table 8.2. In the right section of Table 8.2 the resulting ranges of the measurement values $\boldsymbol{y}$ are shown.

Table 8.2: Range of the measurement matrix $\boldsymbol{Y}$.

component	**min**	**max**	**unit**	**component**	**min**	**max**	**unit**
Q_maize	5.00	30.00	$\frac{\mathrm{m}^3}{\mathrm{d}}$	pH_1	7.28	7.72	–
Q_manure	5.00	40.00	$\frac{\mathrm{m}^3}{\mathrm{d}}$	$Q_\mathrm{gas,1}$	1,486.38	9,390.72	$\frac{\mathrm{m}^3}{\mathrm{d}}$
Q_grass	0.00	5.00	$\frac{\mathrm{m}^3}{\mathrm{d}}$	$r_{\mathrm{ch}_4,1}$	45.67	56.59	%
Q_ccm	0.00	5.00	$\frac{\mathrm{m}^3}{\mathrm{d}}$	$r_{\mathrm{co}_2,1}$	43.41	54.33	%
				pH_2	7.64	7.89	–
				$Q_\mathrm{gas,2}$	72.85	2,796.08	$\frac{\mathrm{m}^3}{\mathrm{d}}$
				$r_{\mathrm{ch}_4,2}$	48.63	63.14	%
				$r_{\mathrm{co}_2,2}$	36.86	51.37	%

To train and validate the supervised machine learning methods (see Section 4.1.1) in total five training and five validation datasets are created using 5-fold cross-validation. Each training dataset contains the data from 24 selected simulations and thus each validation dataset contains the data from the remaining 96 simulations.

As explained in Section 4.1.1 the estimation task is solved as classification problem. Therefore, the simulated state vectors $\boldsymbol{X}$ are divided into $C = 10$ classes, see eq. (4.12). To measure the performance of the classification methods on the validation datasets the misclassification rate (MCR) is used as a performance measure. This measure is defined as:

$$\begin{aligned} \mathrm{MCR} &:= \quad 100 \cdot \left(1 - \frac{1}{N_\mathrm{V}} \cdot \sum_{i=1}^{N_\mathrm{V}} \Gamma\left(\boldsymbol{y}_i\right)\right), \qquad \boldsymbol{Y}_\mathrm{V} := \left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_i, \ldots, \boldsymbol{y}_{N_\mathrm{V}}\right)^T \\ \Gamma\left(\boldsymbol{y}_i\right) &:= \begin{cases} 1 & \text{if } \boldsymbol{y}_i \text{ classified correctly} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{8.3}$$

For this application in this thesis the two methods LDA and Random Forests are used (see Subsection 4.1.1.1 and Subsection 4.1.1.3). In the publication Gaida et al. (2012b) also the method GerDA (see Subsection 4.1.1.2) was used with very good results. Out of time and resource issues the method was not applied this time.

For LDA only the dimension of the projected feature space d has to be specified, see Section 4.1.1.1. Here, an LDA transformation into a feature space of $d = C - 1$ dimensions led to the best subsequent linear classification results.

Random Forests was configured with 20 decision trees. Further parameters are set to default values as are given in the implementation of Jaiantilal (2010).

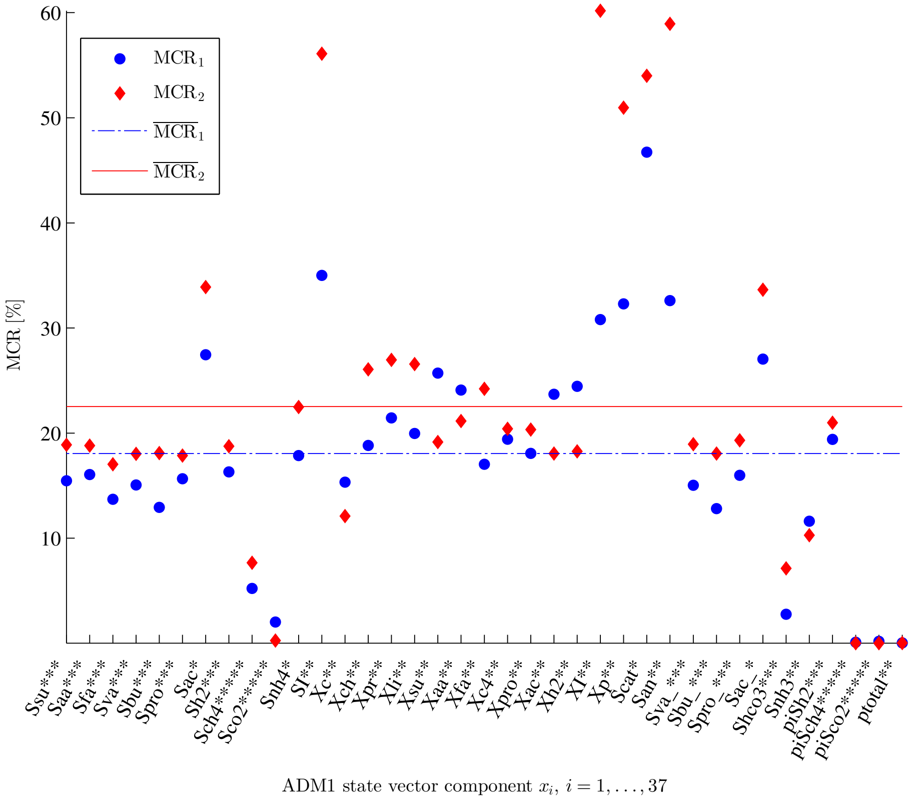

Figure 8.1: Comparison of the MCR of the state estimators for the two digesters using RF. The n_* stars (*) next to the x-axis labels signifies that for these state vector components only a $C - n_*$ classification problem was solved, due to insufficient data support for some of the classes. This was addressed by merging such classes with their neighbor class.

In Figure 8.1 the mean MCR obtained during the 5-fold cross-validation for both digesters is shown. Next to the results for each state vector component also the mean performance over all state vector components are visualized as straight lines. The state vector of the ADM1 is defined in Table 7.1. The symbols of the state vector components shown in Figure 8.1 are not exactly visualized the same way as they are in Table 7.1. But, as the state vector components are given in the same order the meaning of each symbol can be deduced.

A mean misclassification rate $\overline{\text{MCR}}$ of around 20 % as is visualized in Figure 8.1 is not really satisfying. In Gaida et al. (2012b) better results could be obtained. The reason for the decrease in performance could be because in this thesis a more complex model is used as was used in Gaida et al. (2012b). Especially the connection of some kinetic parameters to the substrates (see Section 7.5) might lead to more severe non-linearities and worse predictability. Nevertheless, this accuracy is seen as good enough for its purpose. As will be seen in the next Chapter 9 it is recommended to perform predictions over 100 days or longer. The obtained state after 100 days depends largely on the substrate feed and only very loosely on the initial state. Therefore, the exact value of the initial state is not that important if one operates with large prediction horizons. As most biogas plants are operated in steady state, to work with a long prediction horizon is good practice. If a biogas plant is operated dynamically as it will be more often the case in the future the initial state becomes more important again. In that case, the question will be whether the state estimator's accuracy will be sufficient for dynamic plant operation. This question will not be answered in this thesis.

The average results for both methods LDA and Random Forests are given in Table 8.3. It can be seen that LDA yields very bad results, therefore it is not used any further for the state estimation task in this thesis. Applying LDA to the first 25 principal components, determined using principal component analysis (PCA), better results can be achieved, see Table 8.3. The application of Random Forests to the first 25 principal components yields worse results than using Random Forests directly on the raw data.

Table 8.3: Performance comparison of the state estimators on the investigated methods. $\overline{\text{MCR}}$ and $\bar{\sigma}_{\hat{x}}$ are the mean MCR (standard deviation, respectively) over all state vector components.

method	$\overline{\text{MCR}}_1$ $(\pm\bar{\sigma}_{\hat{x},1})$ [%]	$\overline{\text{MCR}}_2$ $(\pm\bar{\sigma}_{\hat{x},2})$ [%]
LDA	71.68 (±13.53)	70.13 (±19.33)
LDA & PCA	24.94 (±12.56)	31.35 (±17.54)
Random Forests	18.06 (±10.31)	22.53 (±15.76)

In Gaida et al. (2012b) further experiments were performed regarding number of moving average filters and estimator performance using noisy data.

8.3 Summary and Discussion

In this chapter it could be shown that the state estimation approach originally proposed in Section 4.1 is capable to estimate the state vector of the ADM1 with moderate accuracy. Whether the accuracy is sufficient will be investigated in the simulation studies in Section 9.3.6 of the next chapter. However, it should be mentioned that for practical use of this state estimator two challenges have to be dealt with. On the one hand, the state estimator depends on the simulation model of the biogas plant and on the other hand it depends on the characteristics of the fed substrates.

As the anaerobic digestion process changes and usually the substrates do not have constant parameters as well, the state estimator has to be retrained throughout. If the model is changed or recalibrated, also the machine learning method, here Random Forests, must be learned anew. To avoid spending the time for the full training process online learning methods that update the surrogate model based on new data might be an option. For Random Forests there are algorithms called online Random Forests, e.g. see (Osman, 2008, Saffari et al., 2009, Denil et al., 2013).

As measuring substrate parameters frequently is costly and elaborate, to estimate them instead or additionally is an interesting alternative. Especially for biogas plants operating on the OFMSW the input changes constantly so that substrate parameters must either be measured online or be estimated. There are a couple of publications focusing on input estimation for the anaerobic digestion process, e.g. see (Theilliol et al., 2003, Jáuregui-Medina et al., 2009).

Alternatives to state-based controls are controllers that use directly measurable variables with or without a data-driven model. In this case it is important to measure a combination of process values that lets the control identify the "state" of the process. Examples of that approach can e.g. be found in Boe et al. (2010), Castellano et al. (2007) and Molina et al. (2009).