DATA QUALITY ASSESSMENT FOR SYSTEM IDENTIFCATION: A PROPOSED NEW FRAMEWORK

Yuri A.W. Shardt and Biao Huang*
Department of Chemical and Materials Engineering
University of Alberta
Edmonton, AB, Canada, T6G 2V4

Abstract

In many industrial plants, data is sampled quickly which leads to the collection of gigabytes of process values that are stored in the data historian, most of which is routine operating data. However, the usefulness of this routine operating data for such applications as system identification, performance assessment, and fault detection, is still largely unknown. Therefore, in this paper, a framework is proposed for analysing the data quality of routine operating data. This framework consists of two main components: model segmentation and data quality assessment for each identified region. Model segmentation is performed using signal entropy, while the data quality is assessed using a ratio between the extreme eigenvalues of the inverse Fisher information matrix. As an example of the proposed data quality assessment method, a process controlled by a proportional controller is analysed theoretically. As well, the data quality assessment framework is applied to an experimental, heated tank example with multiple states. It is shown that the framework is able to effectively assess the quality of the data.

Keywords

Data Quality Assessment, System Identification

Introduction

Data, data, every where,
Not a drop of knowledge

Much like the mariners in Coleridge's poem *The Rime of the Ancient Mariner* who were surrounded by undrinkable water and had no means of making it useful, in today's industrial plants, control engineers are inundated by routine operating data most of which is discarded since its usefulness is unknown. Thus, there is a need to develop techniques that can be used to assess the quality of the stored data, specifically for control purposes.

Routine operating data, which can be defined as closed-loop data without any setpoint changes, is the most common type of data collected in industry. This data is often used for process monitoring or control performance assessment. However, it is rarely used for system identification or controller tuning, as there currently do not exist any methods to determine whether or not the data can actually be used for the given purpose.

Therefore, the objective of this paper is to present a framework for data quality assessment that can be used to give guidance on the usefulness of the routine operating data, with a special focus on the data quality measure to be used. As a sample application of the proposed framework, a detailed analysis of the proposed measure for proportional, as well as proportional and integral (PI), controllers will be presented. An experimental example will also be presented.

Data Quality Assessment Framework

Consider the case that routine operating data of length N has been extracted from a data historian. This data will

* To whom all correspondence should be addressed

consist of two components, the output vector, $\vec{y}_t$, and the input vector, $\vec{u}_t$, that is, $(\vec{y}_t, \vec{u}_t) \in \mathbb{R}^{N \times N}$. The objective is to determine whether the extracted data can be used to identify a model with a given structure. Auxiliarily, it is necessary to determine whether or not all the extracted data belongs to the same model. Thus, the proposed framework for data quality assessment consists of two steps:

1) **Model Segmentation**, which determines which sections of the extracted data series belong to which model, and
2) **Data Quality Assessment**, which assesses the quality of the data for each of the extracted regions.

Model segmentation is required as a preprocessing step to separate the data coming from different models for three reasons. Firstly, it is unlikely that the extracted data can be modelled by a single model and doing this may introduce errors in the model. Secondly, separating the different models in the extracted data can allow preliminary detection of faults, such as sensor failure and actuator problems, as these will cause the model to change from that which is expected. Thirdly, during the data quality assessment step, the presence of multiple models in the extracted data set may increase the perceived information content of the data with respect to the assumption of a single model.

Model Segmentation

The purpose of the model segmentation preprocessing step is to partition the extracted data series into regions that can be assumed to be from the same model without first identifying a model for each region. Additionally, it would be useful if the technique could identify which of the partitioned regions belong to the same model.

It should be noted that model segmentation does not assign a cause or reason for the change in the model. In addition to a switching system, the observed model of the system could change due to sensor failures, changes in the actuator, disturbance model changes, or slow changes in the actual process model, such as fouling in a heat exchanger.

Many different methods to partition the input and output data can be encountered in the literature. The methods can be classified into two broad categories: offline, where the complete data set is obtained and then processed, and online, where data processing occurs in tandem with data collection and the segmentation points are determined immediately (Keogh, Chu, Hart, & Pazzani, 2004). Most segmentation methods can be classified into the following three groups: sliding window, where to a given segment subsequent values are added until some error bound is exceeded; top-down, where the data is partitioned starting from the initial data set until some stopping criterion is reached; and bottom-up, where the data is first partitioned into the smallest number of segments and each of the segments are then merged until some stopping criterion is reached (Keogh, Chu, Hart, & Pazzani, 2004). For all these approaches, a model structure, often linear, is assumed for the data to follow.

In the control literature, a commonly used partition method is based on the local approach for fault detection developed by Basseville *et al*. (Basseville, 1998). Similar to the preceding methods, an assumed model for the extracted data to follow is required.

More novel approaches include a signal entropy-based approach (Denis & Crémoux, 2002; Micó, Mora, Cuesta-Frau, & Aboy, 2010), which has the advantage that no model of the system need be provided in order to segment the data. This approach is useful for online performance assessment of a given model to represent the system, since there is no need to waste computational resources on determining the best model for the current data. For data quality assessment, this approach has the benefit that since no model need be assumed for the data, there is no need to be concerned with whether the extracted data is well represented by the given model. In fact, assuming a model would lead to circular reasoning.

In this vein, consider the method originally proposed by Denis and Crémoux (Denis & Crémoux, 2002) to segment a deterministic geophysical signal by computing the changes in signal entropy using the following quantities

$$H(t) = \log\left(\frac{L(t)}{t}\right) \tag{1}$$

where $L(t)$ is defined as the "length" or the degree of tortuosity in the signal, which for a time series can be written as

$$L(t) = \sum_{l=1}^{t} \left| X(l) - X(l-1) \right| \tag{2}$$

This has been extended to stochastic signals by rewriting Eq. (1) as the following difference equation (Shardt & Huang, 2011)

$$H(t) = \log\left(\left(1 - z^{-1}\right) L(t)\right) \tag{3}$$

where z is the forward shift operator. By taking the difference between the input and output entropies, that is,

$$\Delta H(t) = H_{output}(t) - H_{input}(t) = \log\left(\frac{L_{output}(t)}{L_{input}(t)}\right) \tag{4}$$

the resulting difference can be shown to be independent of the input into the process. A region can be said to be the same if the entropy change for the given region is similar to the value for another region.

After applying model segmentation, a total of N_k regions will be identified. Each identified region will have an entropy value H_i, length N_i, and a subset of the original data points $\left(\vec{y}_t, \vec{u}_t\right)_i \in \mathbb{R}^{N_i \times N_i}$. Once model segmentation has been completed, it is then necessary to assess the data quality of each part.

Data Quality Assessment: Initial Results

The purpose of data quality assessment is to determine whether or not the data contains sufficient excitation or information to identify a model. This step would be performed for each segment separately.

Consider a single region with data given as $\left(\vec{y}_t, \vec{u}_t\right)_i$ of length N_i. Assume that the model of interest for this region has the form given as

$$\vec{y}_t = f\left(u_t, \vec{\beta}\right) \tag{5}$$

where $\vec{\beta}$ is a vector of p-parameters, that is,

$$\vec{\beta} = \left\langle \beta_1, \quad \beta_2, \quad \cdots \quad \beta_p \right\rangle \tag{6}$$

It is preferably that p be much smaller than N_k as the performance of identification using routine operating data is strongly influenced by small data sets.

Linearising the model about the true parameter values gives the process Jacobian, J, with respect to the input variables,

$$\begin{aligned} J &= \left.\frac{\partial f\left(u_t, \vec{\beta}\right)}{\partial \vec{\beta}}\right|_{\substack{\vec{\beta}=\hat{\vec{\beta}} \\ u_t = u}} \\ &= \left[\begin{array}{cccc} \dfrac{\partial f\left(u_t, \vec{\beta}\right)}{\partial \beta_1} & \dfrac{\partial f\left(u_t, \vec{\beta}\right)}{\partial \beta_2} & \cdots & \dfrac{\partial f\left(u_t, \vec{\beta}\right)}{\partial \beta_p} \end{array}\right]\Bigg|_{\substack{\vec{\beta}=\hat{\vec{\beta}} \\ u_t = u}} \end{aligned} \tag{7}$$

Evaluating the Jacobian matrix, J, given as Equation (7) for each of the inputs in the k^{th} region will give the regression matrix, A, that is,

$$A = \left[\begin{array}{cccc} \dfrac{\partial f\left(u_1, \vec{\beta}\right)}{\partial \beta_1} & \dfrac{\partial f\left(u_1, \vec{\beta}\right)}{\partial \beta_2} & \cdots & \dfrac{\partial f\left(u_1, \vec{\beta}\right)}{\partial \beta_p} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial f\left(u_k, \vec{\beta}\right)}{\partial \beta_1} & \dfrac{\partial f\left(u_k, \vec{\beta}\right)}{\partial \beta_2} & \cdots & \dfrac{\partial f\left(u_k, \vec{\beta}\right)}{\partial \beta_p} \end{array}\right] \tag{8}$$

The inverse of the Fisher information matrix, Q, can be obtained as follows:

$$Q = A^T A \tag{9}$$

For the case of linear, least-squares regression, this represents the matrix that would need to be inverted in order to determine the parameter estimates.

In numerical methods, the condition of a matrix is used to determine whether a given system is invertible. Many equivalent definitions of condition exist. In this paper, the definition using eigenvalues is selected. It can be noted that an uninvertible matrix contains at least one eigenvalue with a value of zero (Anton, 2000). Therefore, the data quality index, η_{data}, can be defined as

$$\eta_{data} = \frac{\max\left(\left|\text{eig}\left(A^T A\right)\right|\right)}{\min\left(\left|\text{eig}\left(A^T A\right)\right|\right)} \tag{10}$$

where *eig* represents the eigenvalues of A^TA. A matrix is said to be well-conditioned if the ratio of largest to smallest eigenvalues in absolute value is less than a given threshold, ε.[1] Similarly, the data is said to be **informative enough** with respect to the given model structure if $\eta_{data} < \varepsilon$, that is, the Q-matrix is sufficiently well-conditioned for the taking of an inverse. As well, a well-conditioned Q-matrix will imply that the variances obtained for the parameters will be reasonable and hence the results obtained will be significant.

Theoretical Analysis of Data Quality for a Proportional Controller and an Autoregressive Model with Exogenous Input

As an example of the proposed data quality assessment framework, consider the discrete-time, autoregressive model with exogenous input (ARX)

$$y_t = \vec{y}_{t-1}\vec{\beta} + \vec{u}_{t-1-d}\vec{\alpha} \tag{11}$$

where

$$\begin{aligned} \vec{y}_{t-1} &= \left\langle y_{t-1}, y_{t-2}, \ldots y_{t-n_\beta} \right\rangle \\ \vec{\beta} &= \left\langle \beta_1, \beta_2, \ldots \beta_{n_\beta} \right\rangle^T \\ \vec{u}_{t-d-1} &= \left\langle u_{t-1-d}, u_{t-2-d}, \ldots u_{t-n_\alpha-d} \right\rangle \\ \vec{\alpha} &= \left\langle \alpha_1, \alpha_2, \ldots \alpha_{n_\alpha} \right\rangle^T \end{aligned} \tag{12}$$

where α and β are the parameters to be fitted and d is the time delay. For ARX models, linear, least-squares analysis can be directly used to obtain parameter estimates (Ljung, 1999). The Q-matrix can be written as (Söderström & Stoica, 1989)

[1] A threshold, ε, of 10^4 is often used in practice.

$$Q = \begin{bmatrix} E\left(\vec{y}_{t-1}^{T}\vec{y}_{t-1}\right) & E\left(\vec{y}_{t-1}^{T}\vec{u}_{t-d-1}\right) \\ E\left(\vec{u}_{t-d-1}^{T}\vec{y}_{t-1}\right) & E\left(\vec{u}_{t-1}^{T}\vec{u}_{t-d-1}\right) \end{bmatrix} \quad (13)$$

It can be noted that with a proportional controller, u_t and y_t are not independent and can be given as

$$u_t = -Ky_t \quad (14)$$

This implies that the Q-matrix given by Eq. (13) can be rewritten as

$$Q = \begin{bmatrix} E\left(\vec{y}_{t-1}^{T}\vec{y}_{t-1}\right) & -KE\left(\vec{y}_{t-1}^{T}\vec{y}_{t-d-1}\right) \\ -KE\left(\vec{y}_{t-d-1}^{T}\vec{y}_{t-d}\right) & K^2E\left(\vec{y}_{t-d-1}^{T}\vec{y}_{t-d-1}\right) \end{bmatrix} \quad (15)$$

$E\left(\vec{y}_{t-1}^{T}\vec{y}_{t-1}\right)$ can be rewritten as

$$\begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n_\beta - 1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(n_\beta - 2) \\ \vdots & & \ddots & \\ \gamma(n_\beta - 1) & \gamma(n_\beta - 2) & \cdots & \gamma(0) \end{bmatrix} \quad (16)$$

while $E\left(\vec{y}_{t-1}^{T}\vec{y}_{t-d-1}\right)$ can be rewritten as

$$\begin{bmatrix} \gamma(d) & \gamma(d+1) & \cdots & \gamma(d+n_\alpha - 1) \\ \gamma(|d-1|) & \gamma(d) & \cdots & \gamma(d+n_\alpha - 2) \\ \vdots & & \ddots & \vdots \\ \gamma(|d+1-n_\beta|) & \gamma(|d+2-n_\beta|) & \cdots & \gamma(|d+n_\alpha - n_\beta|) \end{bmatrix} \quad (17)$$

It can be noted that $E\left(\vec{y}_{t-d-1}^{T}\vec{y}_{t-1}\right) = E\left(\vec{y}_{t-1}^{T}\vec{y}_{t-d-1}\right)^T$. Finally, $E\left(\vec{y}_{t-d-1}^{T}\vec{y}_{t-d-1}\right)$ can be rewritten as

$$\begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n_\alpha - 1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(n_\alpha - 2) \\ \vdots & & \ddots & \vdots \\ \gamma(n_\alpha - 1) & \gamma(n_\alpha - 2) & \cdots & \gamma(0) \end{bmatrix} \quad (18)$$

From Eqs. (15), (16), (17), and (18), it can be seen that if $d = 0$, then $-K$ times the first column of Q will always be equal to the second column of Q. This implies that the matrix is not invertible and hence in this situation the model should not be identifiable. This result is in agreement with the previously published results in system identification for the conditions for identifying a process using closed-loop data without external excitation and a proportional controller (Söderström & Stoica, 1989; Gevers, Bazanella, & Ljubiša, 2008; Shardt & Huang, 2011), which state that for identifying a system without external excitation, the following relationship must hold

$$\max\left(d - n_\beta, -n_\alpha\right) \geq 0 \quad (19)$$

Examining the Q-matrix, it can likewise be seen that Eq. (19) represents one of the conditions for Q to be in general invertible.

Theoretical Analysis of Data Quality for a Proportional, Integral Controller and a First-Order Autoregressive Model with Exogenous Input

As the complexity of the model and controller increase, the ability to analyse the results exactly is made more difficult. For the case of a proportional and integral (PI) controller and a first-order autoregressive model with exogenous input (ARX) with $d = 0$, an analytical solution can be obtained for identifiability as a function of the parameter values (Shardt & Huang, 2010). It should be noted that an arbitrary first-order ARX model cannot be identified in this case. It can be shown that the regions of unidentifiability are linked with the invertibility of the Q-matrix. The details of this proof are omitted in interests of length.

Performing a simulation for this case and calculating the data quality index for different values of α_1, where $\beta_1 = 1$ and a PI controller of the form

$$G\left(z^{-1}\right) = \frac{1 - \left(-0.9 + 0.25\alpha_1\right)z^{-1}}{1 - 0.9z^{-1}} \quad (20)$$

is used. A total of 10,000 data point was used for the simulation.

Figure 1 shows the data quality index, η_{data}, and the difference between the estimated and true values for α_1 as a function of the true value of α_1. It can be seen that as η_{data} increases in value, the difference between the estimated and true values also increases in value. Furthermore, it can be theoretically shown that at $\alpha_1 = -13/18$ (Shardt & Huang, 2010), the Q-matrix is uninvertible, and hence the system is unidentifiable.

Experimental Example of the Framework

Consider the heated-tank system shown in Figure 2. It can be noted that the hand valve on the drain can be used to change the exit resistance coefficient and create different models. The level controller was set in cascade with the cold water flow controller. The temperature setpoint was set to 43°C throughout the experiment. The cold water temperature fluctuated around 23˚C. Different operating conditions were produced by changing the hand valve and level setpoints in the system. Open-loop step tests were used to determine the expected open-loop

process models for the different operating points, which had the form

$$G_p(s) = \frac{K}{1+\tau_p s} e^{-\tau_d s} \qquad (21)$$

where K is the gain, τ_p is the process time constant, and τ_d is the deadtime. The results are shown given in Table 1. The same proportional, integral, and derivative (PID) controller was used with values $K_c = 1.5$ (normalised), $\tau_I = 57.5$ s, and $\tau_D = 5.8$ s. All the experimental data is shown in Figure 3. The process started in Model A and at 1017 s, it was changed to Model B. At 2498 s, the model was changed from Model B to Model C. Finally, at 4044 s, the model was changed from Model C to Model D.

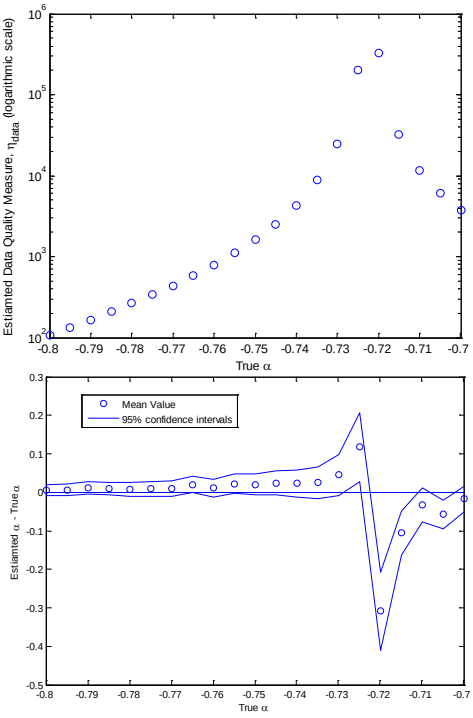

Figure 1. (top) Data quality index, η_{data}, and (bottom) the difference between the estimated and true values of α_1 for the controller given as Eq. (20). The bottom figure is taken from (Shardt & Huang, 2010).

Table 1. Parameters for the four models

Model	Hand Valve (°)	Level Setpoint (cm)	K (°C·h/kg)	τ (s)	τ_d (s)
A	50	20	0.74	52	40
B	65	20	1.68	73	40
C	65	35	2.02	133	40
D	55	20	1.42	56	40

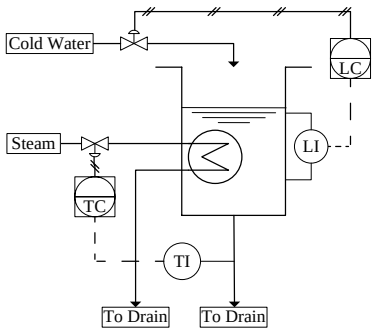

Figure 2: Process schematic

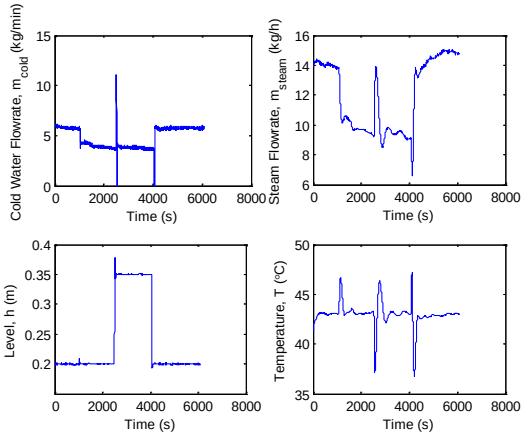

Figure 3: Cold water flow rate (top right), steam flow rate (top right), level (bottom left), and temperature (bottom right) as a function of time for the duration of the experiment

Model Segmentation

Figure 4 shows the entropy difference between the output (temperature) and input (steam flow rate) as a function of time calculated using Eq. (4). It should be noted that the drift in the values implies that the given process had not yet necessarily settled from the previous change or that unexpected changes, such as large, unmeasured changes in the cold water temperature may be present. This is especially the case for the first and last regions. The regions which are more or less straight lines represent the constant model areas, while the abrupt changes in the model are reflected by the spikes. Finally, it can be noted due to deadtime in the system, as well as the fact that the entropy sum was taken over all the data, the identification of changes is delayed. The identified regions with their corresponding entropies are given in Table 2.

An autoregressive model of the form

$$y_t = \frac{\beta_1 z^{-40}}{1+\alpha_1 z^{-1}} u_t + \frac{1}{1+\alpha_1 z^{-1}} e_t \qquad (22)$$

was fit to each of the segmented regions. All models passed the appropriate regression analysis tests (Ljung, 1999). The results are presented in Table 3.

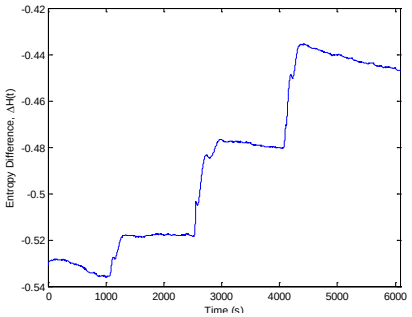

Figure 4: Length of the curve as a function of time for the experimental data

Table 2. Information about the segmented data set

Region	Start (s)	End (s)	Data Points	Entropy	η_{data}
A	162	1077	915	-0.5307	3.5
B	1296	2530	1234	-0.5173	5.1
C	2966	4066	1100	-0.4793	2.0
D	4367	6100	1733	-0.4451	29.3

Table 3. Fitted model parameters

Region	α_1 (standard deviation)	β_1 (standard deviation)
A	-0.987±0.005	0.0038±0.003
B	-0.995±0.02	0.0037±0.0009
C	-0.997±0.001	0.0085±0.001
D	-0.990±0.003	0.00312±0.0009

From the results, it can be seen that the identified models as expected are different. Converting the discrete time values to continuous time values shows that the identified models are close to the step test values, as shown in Table 1. It should be noted that the step tests values themselves may have some inherent error.

Conclusions

This paper has presented a framework for assessing the quality of routine-operating data for system identification. In this framework, there are two steps: model segmentation and data quality assessment. Signal entropy is used to partition the models and a condition number based on the inverse of the Fisher information matrix is used to determine the data quality. It was shown that the results based on the condition number correspond well with previously developed complex system identification criteria. Finally, the proposed framework was tested on an pilot-scale example, where the framework was able to partition the data into separate models and accurately determine the reliability of the model parameters obtained.

Future work will focus on testing the framework for more general multivariate cases, as well as for more complex industrial systems.

Acknowledgments

The authors would like to thank the National Science and Engineering Research Council (NSERC) of Canada for funding.

References

Anton, H. (2000). *Elementary Linear Algebra* (8th ed.). Hoboken, New Jersey, United States of America: John Wiley & Sons, Inc.

Basseville, M. (1998). On-board Component Fault Detection and Isolation Using the Statistical Local Approach. *Automatica , 34* (11), 1391-1415.

Denis, A., & Crémoux, F. (2002). Using the Entropy of Curves to Segment a Time or Spatial Series. *Mathematical Geology , 34* (8), 899-913.

Gevers, M., Bazanella, A., & Ljubiša, M. (2008). Informative data: how to get just sufficiently rich? *Proceedings of the 47th IEEE Conference on Decision and Control*, (pp. 1962-1967). Cancun.

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2004). Segmenting Time Series: A Survey and Novel Approach. In M. Last, A. Kandel, & H. Bunke, *Data mining in time series databases* (pp. 1-22). Singapore, Singapore: World Scientific Publishing Co. Pte. Ltd.

Ljung, L. (1999). *System Indentification: Theory for the User* (2nd ed.). Upper Saddle River, New Jersey, United States of America: Prentice-Hill, Inc.

Micó, P., Mora, M., Cuesta-Frau, D., & Aboy, M. (2010). Automatic segmentation of long-term ECG signals corrupted with broadband noise based on sample entropy. *Computer Methods and Programs in Biomedicine , 98*, 118-129.

Shardt, Y. A., & Huang, B. (2011). Assessing Plant-Model Mismatch Using Signal Entropy. *Proceedings of the Canadian Chemical Engineering Conference.* London, Ontario, Canada.

Shardt, Y. A., & Huang, B. (2011). Closed-Loop Identification with Routine Operating Data: Effect of Time Delay and Sampling Time. *Journal of Process Control , 21* (7), 997-1010.

Shardt, Y. A., & Huang, B. (2010). Conditions for Identifiability Using Routine Operating Data for a First-Order ARX Process Regulated by a Lead-Lag Controller. *DYCOPS 2010.* Leuven, Belgium.

Söderström, T., & Stoica, P. (1989). *System Identification.* New York, New York, United States of America: Prentice Hall.