Data Quality Assessment

Yuri A.W. Shardt

Data quality assessment seeks to provide a framework for determining the quality of stored data for use in process system identification. The benefit of using historical data for identification is that the potential exists for developing models without needing new (open-loop) tests. This can reduce the costs associated with process identification, as there are no costs associated with loss of productivity or safety concerns. Furthermore, data quality assessment can provide guidance for why the resulting models are not sufficiently good when tested on new data. This often is the result of poor data quality in the initial, modelling data set.

In any data quality assessment framework, the following topics must be considered:

1) **Motive**: For the purposes of this report, the motive is system identification from routine operating data. Therefore, there is a need to specify the model structure and time delay in order to obtain a measure of the data quality.
2) **Outliers**: Outliers are data points that are somehow different from the expected value. If it is assumed that the number of outliers is small in comparison to overall data length, the impact on the overall metric should not be sever. If it is desired to remove outliers beforehand, then recourse can be had to any standard methods, such as 3σ-edit rules or box-plot analysis.
3) **Data Length**: For routine operating data and due to the fact that certain modes may not be excited often, the data length is especially important. Empirically, it has been determined that less than about 1,000 data points can provide unsatisfactory parameter estimates.
4) **Model Changes**: When examining data from the historian, it is possible that the process conditions and hence the underlying model has changed. This implies that there is a need to partition the data into its constituent regions in order to avoid fitting a single model to a multimodel region. As well, model changes can inflate the quality of the data with respect to a single model, but provide unreasonable parameter estimates.

Proposed Data Quality Framework

Based on the above discussion, a 2-step data quality framework presented in Figure 9, is adopted. Assume that routine operating data of length N has been extracted from a data historian.

The data will 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. As shown in Figure 9, the proposed framework for data quality assessment consists of two steps:

1) **Data 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 for either system identification for forecasting.

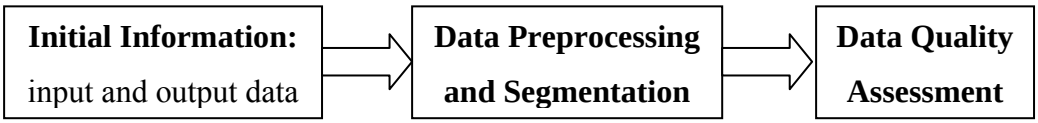

Figure 1: Data Quality Assessment Framework

Data segmentation

Data segmentation, or clustering, must be performed for the extracted data. The primary requirement for the data segmentation is that there should not be a need to fit a model to the extracted data in order to determine whether or not the model changed. Many different methods are available, including principle component methods, autocorrelation-based methods (Badwe, Patwardhan, Shah, Patwardhan, & Gudi, 2010), the local approach fault detection (Basseville, 1988), and entropy-based methods (Denis & Crémoux, 2002). In the proposed framework, the entropy-based approach will be used due to its simplicity. For the given input and output signals compute the differential entropy, ΔH, of the process using the following equation

$$\Delta H_t = \log\left(\frac{L_{output_t}}{L_{input_t}}\right) \tag{1}$$

where $L(t)$ is the length of the curve and can be calculated as

$$L_t = \frac{\left|\left(1 - z^{-1}\right) y_t\right|}{1 - z^{-1}} = \sum_{i=0}^{t} \left|y_t - y_{t-1}\right| \tag{2}$$

The changes in the differential entropy can be used to monitor the changes in the process. At the end of this step, the data will have been partitioned into N_l regions, each of which have the same

model. It can be noted that the reason for the changes in model cannot be determined using this method and if it were desired further analysis would be required in order to determine the cause. Regions with similar entropy can be considered to come from the same underlying model (up to a difference in the gain).

Assessing Data Quality

Data quality assessment for system identification is focused on determining whether or not the data is sufficiently informative or rich in order to identify a model from it. This analysis can be achieved by considering the properties of the Fisher information matrix. It will be assumed in this example that all the data used comes from a single model. 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

$$\vec{y}_t = f\left(u_t, \vec{\theta}\right) \tag{3}$$

where f is an arbitrary function and $\vec{\theta}$ is a vector of r-parameters, that is,

$$\vec{\theta} = \left\langle \theta_1, \quad \theta_2, \quad \cdots \quad \theta_r \right\rangle \tag{4}$$

It is preferably that r be much smaller than N_i as the performance of identification using routine operating data is strongly influenced by small data sets. Consider the identification matrix, $\mathcal{M}$,

$$\mathcal{M} = \begin{bmatrix} \dfrac{\partial f\left(u_1, \vec{\theta}\right)}{\partial \theta_1} & \dfrac{\partial f\left(u_1, \vec{\theta}\right)}{\partial \theta_2} & \cdots & \dfrac{\partial f\left(u_1, \vec{\theta}\right)}{\partial \theta_p} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial f\left(u_i, \vec{\theta}\right)}{\partial \theta_1} & \dfrac{\partial f\left(u_i, \vec{\theta}\right)}{\partial \theta_2} & \cdots & \dfrac{\partial f\left(u_i, \vec{\theta}\right)}{\partial \theta_p} \end{bmatrix} \tag{5}$$

The inverse of the Fisher information matrix, $\mathcal{Q}$, can be obtained as follows:

$$\mathcal{F} = \mathcal{Q}^{-1} = \mathcal{M}^T \mathcal{M} \tag{6}$$

Therefore, analysis of the invertibility of $\mathcal{F}$ can be used to assess the quality of the data. The easiest approach is to consider the condition number of the matrix of interest, $\mathcal{F}$, to determine whether or not the data collected is rich enough. As will be shown later, the condition number of

the matrix can be computed using a ratio of the eigenvalues of $\mathcal{F}$, namely, define the data quality index, η_{data} as

$$\eta_{data} = \frac{\max\left(\left|\lambda\left(\mathcal{M}^T\mathcal{M}\right)\right|\right)}{\min\left(\left|\lambda\left(\mathcal{M}^T\mathcal{M}\right)\right|\right)} \tag{7}$$

where λ represents the eigenvalues of $\mathcal{M}^T\mathcal{M}$. 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. Similarly, the data is said to be **informative enough** with respect to the given model structure if $\eta_{data} < \varepsilon$, that is, the $\mathcal{F}$-matrix is sufficiently well-conditioned for the taking of an inverse. As well, a well-conditioned $\mathcal{F}$-matrix will imply that the variances obtained for the parameters will be reasonable and hence the results obtained will be significant. Practically, a threshold value of 10^4 works well.

Case Study

Consider the data shown in Figure 2, obtained from a single-heated tank experiment with changes in the hand valve and level setpoint. The information about the different regions is shown in Table 1. The temperature setpoint was 40.2°C for the first part of the experiment and then set to 50.6°C for the remainder. The cold water temperature fluctuated around 8.8˚C. Different operating conditions were produced by changing the hand valve in the system. The same proportional and integral (PID) controller was used for all cases with values $K_c = 2.5$ (normalised) and $\tau_I = 100$ s. A schematic of the system is shown in Figure 118.

The system was assumed to be well-described by a FOPDT model. Therefore, step tests were used to obtain the open-loop process models for the different operating points. The results are also shown in Table 1.

Table 1: Parameters for the three regions

Regions	Hand Valve (°)	Level Setpoint (cm)	K (°C·h/kg)	τ (s)	θ (s)
A	50	20	1.46	80	30
B and C	65	20	2.34	94	30

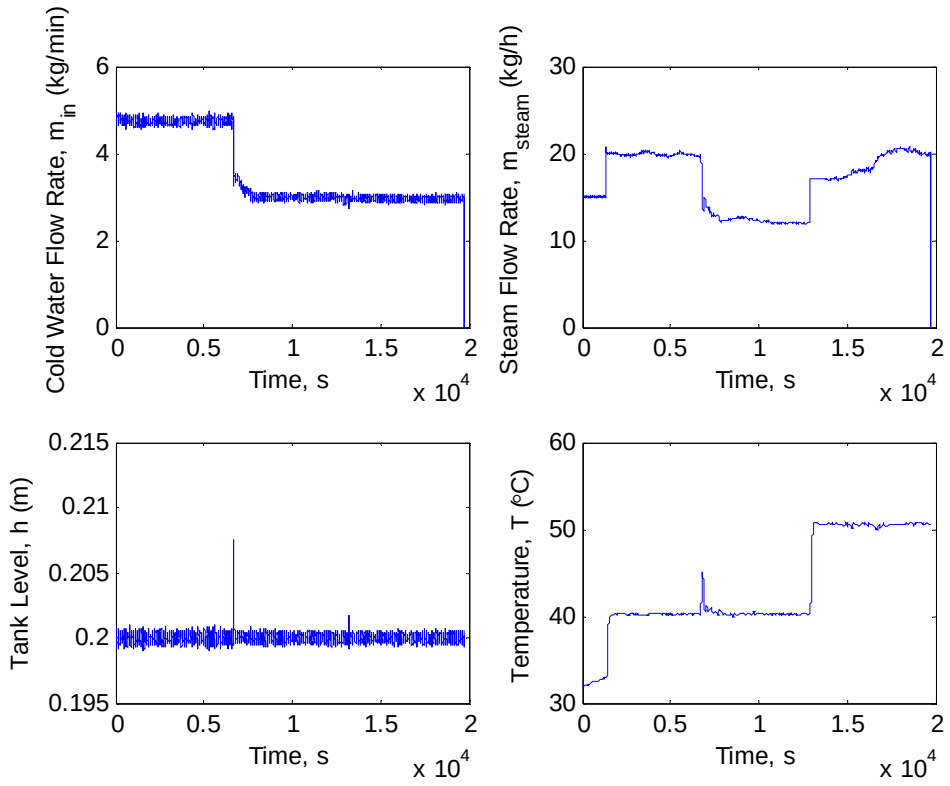

Figure 2: Cold water flow rate (top right), steam flow rate (top right), level (bottom left), and temperature (bottom right) as a function of time for the duration of the experiment

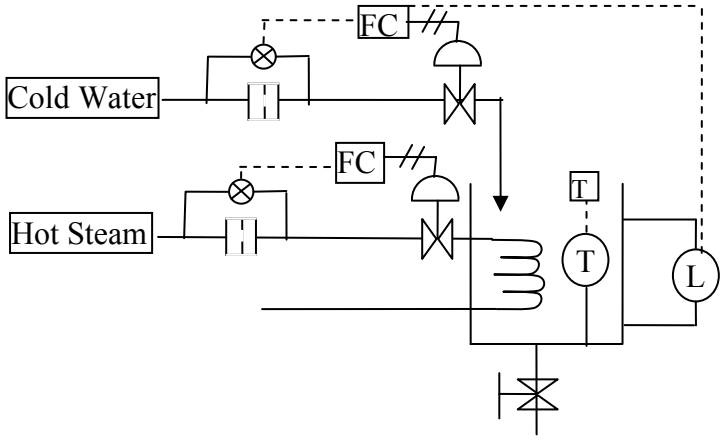

Figure 3: Schematic of the Process

Results

The process started in Model A and at 6684 s, it was changed to Model B. At 12,830 s, the system was set to open-loop and a step test was performed in order to determine the FOPDT parameters for Model B. At 14,150 s, the system reached a new steady-state temperature of

50.6°C. The loop was closed once more and the process ran in closed-loop with a temperature setpoint of 50.6°C. However, for the later part of Model B, the temperature of the cold water increased from 8.8°C to 10.5°C. After staying relatively constant for the duration of the step test and the first part of Model C, it fell back down to 7.0°C. It should be noted that the cold water temperature is not recorded automatically and must be read manually and tracked.

Data segmentation

Figure 4 shows the differential entropy between the output (temperature) and input (steam flow rate) as a function of time calculated using Equation (1). The red dashed lines are the 3σ confidence intervals for the differential entropy using the mean value of the entropy in the given region. The Latin letters represent the three closed-loop regions with different conditions, while the Greek letters represent the transitional regions. A window of 1,000 samples was used to determine the entropy values.

Firstly, it can be noted that both regions α and γ represent open-loop step tests in the system, while region β represents the change from Model A to Model C. The size of these regions is determined by the windowing used.

Secondly, for Region B, it can be seen that the entropy difference suggests that there are two main regions in the model. This can indeed be confirmed by noting that the cold water temperature did indeed increase over the course of this test. Furthermore, it can be seen that the steam flow rate decreased over this time period. These changes can have an impact both on the identifiability of the model as well as on the entropy difference results, since the cold water temperature is not being monitored.

Finally, for Region C, a similar situation to that observed for Region B occurred, but this time the cold water temperature feed by 3°C and the steam flow rate correspondingly had to increase.

The entropy averages for each of the models is given in Table 2, as well as the start and end points of the data used for identification and the detection change times. The proposed entropy-based method can identify changes almost as quickly as the changes in the signals under consideration. Since there is a time delay of about 30 s in the system, the changes in the temperatures are not noticed until about 30 s have passed. Therefore, there is no way of detecting the changes any faster. For the first region, the change is detected in about 40 s, while for the

second region, due to the change in cold water temperature, the method detects the change too quickly, since there are other confounding issues.

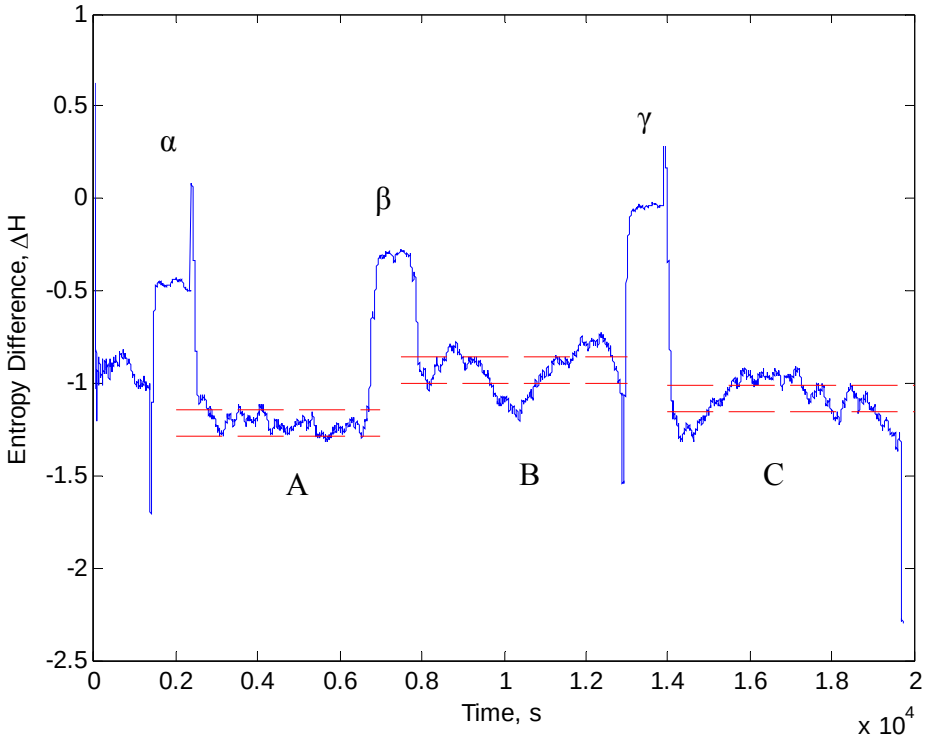

Figure 4: Entropy of the curve as a function of time for the experimental data. The red dashed lines are the 3σ confidence intervals for the entropy about the mean value in the given region. The Latin letters represent the main modelling regions, while the Greek letters represent the transient points.

Table 2: Information about the segmented data set

Region	Start (s)	End (s)	Change Time (s)	Data Points	Entropy
A	2,011	6,674	6,720	4,663	-1.21
B	8,110	12,830	12,840	4,720	-0.933
C	14,980	16,080	—	1,100	-1.08

Data Quality Assessment

The data quality assessment for this step will be performed using the data from regions A and B at different sampling rates and model orders. Sampling rates of 1, 2, 6, 10, and 30 will be considered, while model orders ($n_A = n_B$) of 1, 2, and 3 will be tested. The results are shown in Table 3. It can be seen that the data quality index for the first order models is much smaller

than the corresponding indices for higher order models. The unexpected trend for decreasing data quality values as the sampling rate increases can be attributed to the decrease in the number of data points present in the matrix. Experimentally, it is difficult to obtain large samples of data that are consistent throughout the region. Therefore, there is a need to not only consider the data quality index itself, but also the number of data points available.

A threshold of 10^4 has been used as the upper bound for the data quality index. In this particular case, it can be seen that all of the data is below the threshold and should provide reasonable parameter estimates.

Table 3: Data quality index for different sampling rates and model orders. Entries in **bold** are greater than the threshold of 10^4, while entries in ***bold italics*** have data samples with fewer than 500 values.

Sampling Rates (s)	($n_A = n_B = 1$)		($n_A = n_B = 2$)		($n_A = n_B = 3$)	
	Region A	Region B	Region A	Region B	Region A	Region B
1	4.61	7.58	1,490	3,310	2,780	6,160
2	4.60	7.57	930	2,110	1,870	4,470
6	4.59	7.61	323	742	941	2,580
10	***4.56***	***7.52***	***163***	***346***	***511***	***1,560***
30	***4.87***	***7.82***	***48.1***	***111***	***156***	***404***

System Identification

Although the above values were determined based on an ARX-model, since in closed-loop system identification the model structure must be correctly specified in order to obtain unbiased parameter estimates and it is difficult to obtain an exact ARX system in a real system, BJ-models will be for both Regions A and B for sampling times of 6 s.

The discrete-time values are presented in Table 4. All the fitted models passed the appropriate regression analysis tests, including the uncorrelatedness of the residuals and that between the input and output (Ljung, 1999).

In order to compare these values with the original continuous-time model, the values in Table 4 are converted using the exact discretisation formulae. The resulting values are shown in Table 5. It should be noted that the values differ a bit from the step test values. This discrepancy can be attributed to the fact that the step test values may not be accurate. Nevertheless, except for the gain for Region A, are similar to the values obtained from closed-

loop analysis. For Region A, the large discrepancy in the gain could be caused by the quantisation error in the temperature readings. Zooming in on the temperature data for Region A, it can be seen that the acceptable temperatures are discrete, which can cause a loss of information especially if like in this case the excitations are small.

Table 4. Fitted, discrete-time, BJ-model parameters at a sampling time of 6 s

Region	Region A	Region B
f_1 (standard deviation)	-0.905±0.05	-0.955±0.009
β_1 (standard deviation)	0.0465±0.01	0.0951±0.009
c_1 (standard deviation)	0.161±0.02	0.159±0.04
d_1 (standard deviation)	-0.982±0.01	-1

Table 5. Fitted, continuous-time, FOPDT-model parameters

Region	K (°C·h/kg)	τ_p (s)
A	0.489 (1.56–0.23)	60 (160–36)
B	2.09 (2.87–1.57)	128 (161–106)

Conclusions

The presented case study shows that the proposed method can indeed segment the data and determine its quality effectively. Therefore, this approach looks quite promising.

References

Badwe, A. S., Patwardhan, R. S., Shah, S. L., Patwardhan, S. C., & Gudi, R. D. (2010). Quantifying the impact of model-plant mismatch on controller performance. *Journal of Process Control , 20*, 408-425.

Basseville, M. (1988). Detecting Changes in Signals and Systems—A Survey. *Automatica , 24* (3), 309-326.

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

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