Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Data Quality Assessment for System Identification in the Age of Big Data and Industry 4.0

Yuri A.W. Shardt*, Xu Yang, Kevin Brooks***, Andrei Torgashov******

**Technical University of Ilmenau, Ilmenau, Thüringia, Germany*
***University of Science and Technology Beijing, Peking, China*
****BlueSP and University of the Witwatersrand, South Africa*
*****Institute of Automation and Control Processes FEB RAS, Vladivostok, Russia*
(e-mail: yuri.shardt@uni-due.de, yangxu@ustb.edu.cn; kevin.brooks@bluesp.co.za; torgashov@iacp.dvo.ru).

Abstract: As the amount of data stored from industrial processes increases with the demands of Industry 4.0, there is an increasing interest in finding uses for the stored data. However, before the data can be used its quality must be determined and appropriate regions extracted. Initially, such testing was done manually using graphs or basic rules, such as the value of a variable. With large data sets, such an approach will not work, since the amount of data to tested and the number of potential rules is too large. Therefore, there is a need for automated segmentation of the data set into different components. Such an approach has recently been proposed and tested using various types of industrial data. Although the industrial results are promising, there still remain many unanswered questions including how to handle *a priori* knowledge, over- or undersegmentation of the data set, and setting the appropriate thresholds for a given application. Solving these problems will provide a robust and reliable method for determining the data quality of a given data set.

Keywords: **data quality assessment, system identification, big data, Industry 4.0, soft sensors**

1. INTRODUCTION

In many industry plants, process information is continually stored in a data historian for future reference. Given the increasing demands on industry driven by environmental, governmental, and economic considerations, the ability to use the historical data has increased in importance. These data can be used in many different applications including system identification (Sha'aban, 2019; Khatisbisepehr & Huang, 2008; Mehrkanoon, et al., 2012; Arengas & Kroll, 2017; Yang, et al., 2019), fault detection and diagnosis (Ding, 2014; Ding, et al., 2013), control, especially model predictive control (Shardt & Brooks, 2018; Sha'aban, 2019; Klimchenko, et al., 2019), and process monitoring (Shardt, et al., 2012; Huang, 2003). However, not all of the stored data can be used for each task. In fact, it is imperative to determine the quality of the data before using them. This will avoid using bad data to provide meaningless results.

Data quality assessment, that is, determining which parts of the given data set are useful and which ones are not, has often and, historically speaking, solely, been performed using manual methods. These manual methods include such methods as checking variables against thresholds or using graphs. However, such approaches are only useful for relatively small data sets and off-line use. Since current data sets can contain thousands if not millions of data points spread out over multiple variables, manually verification may not be an effective strategy. Furthermore, such an approach cannot be used for online checking of data quality before using the data for online modelling.

Therefore, there is a need to develop and implement methods for automatic data quality assessment. The first such approaches focused on determining the quality of the data for use in system identification. Two different approaches were considered: the Laguerre-model based method (Bittencourt, et al., 2015; Peretzki, et al., 2011) and the autoregressive model with exogenous input method (Shardt & Huang, 2013). Both methods used the invertibility of the Fisher information matrix as the primary metric to assess the data quality. The difference lies in the models assumed for the data set. The Laguerre-model based method as its name suggests uses the Laguerre model as its basis. The main advantage of this approach is that the time delay need not be known before hand (Bittencourt, et al., 2015). On the other hand, the autoregressive model with exogenous input method uses an autoregressive model with exogenous input (ARX) to assess the data quality. Here, the time delay for the process must be known before hand. However, the model used for assessment is close enough to the real process and hence better represents the final model that will be considered (Shardt, 2012). As well, both approaches consider additional metrics, such as the variability of the input and output signals and the current controller modes (manual, automatic, and cascade). These additional metrics can help segment the data set better and more cleanly.

Nevertheless, the segmentation methods are often too aggressive in splitting the data into separate segments (Shardt & Shah, 2014) and there is a need to develop methods that can combine adjacent regions that could be modelled by similar models. Furthermore, it would be useful to know which regions could be represented by similar models so that large data sets for modelling can be obtained. Various approaches involving signal entropy (Shardt & Huang, 2013; Basseville, 1988; Basseville, 1998; Keogh, et al., 2004; Basseville & Nikiforov, 1993) have been proposed.

Another issue is how to handle multivariate data sets (Arengas & Kroll, 2017; Shardt & Brooks, 2018; Arengas & Kroll, 2019). Although the initial data quality assessment methods considered, univariate data sets, most, if not all, industrial data are better treated as multivariate data. This means that multiple variables need to be considered when implementing the assessment. It can be noted that selecting the appropriate set of variables is one of the key challenges, since some of the variable may well be correlated and thus cause the data quality assessment method to fail.

Data quality assessment has been applied in various industrial settings leading to new challenges and perspectives. Such industrial case studies include the floatation cell in a ore separation process (Shardt & Brooks, 2018), modelling of coal-fired power plants (Li, et al., 2019), large-scale thermal plants (Wang, et al., 2018), and various univariate control loops typically found in a chemical plant (Peretzki, 2010). One of the main challenges from an industrial perspective is the development of appropriate thresholds and values for the tuning parameters so that the approach can apply to the largest number of different cases.

Therefore, this paper seeks to present a comprehensive review of the data quality assessment method including a summative review of the different guidelines and suggestions for setting the thresholds and tuning parameters. As well, areas requiring further work will be proposed. Finally, some examples showing the different aspects of the data quality assessment will be presented.

2. THEORY

Before getting into the practical aspects of data quality assessment, it would be useful to examine the theoretical basis. Consider the general closed-loop system shown in Figure 1, where G_c is the controller transfer function, G_p is the process transfer function, G_l is the disturbance transfer function, y_t is the output signal, r_t is the reference signal, u_t is the input signal, and e_t is the white noise disturbance signal. The theoretical results presented will be considered for both open-loop, that is, without a controller, and closed-loop, with a controller, conditions.

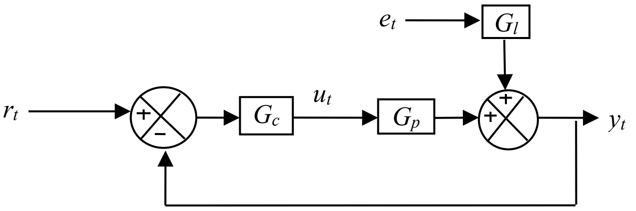

Figure 1: Generic Closed-loop Process

2.1. System Identification Background

For the process shown in Figure 1, we can consider the following situations:

1) **Open-loop data**: In this case, G_c and r_t are ignored and there is no feedback from the output to the input. The input signal itself is manipulated and the process observed.
2) **Closed-loop data**: Here there two subcases to consider:
   a. **Externally Excited Data**: Here the reference signal is continuously changing, that is, exciting the process.
   b. **Routine Operating Data**: Here the reference signal is constant and the only excitations come from the disturbance.

System identification using open-loop data is relatively straightforward (Ljung, 1999). In closed-loop identification, there are three different approaches that can be taken: direct, indirect, or joint identification. In direct identification, the closed-loop input, u_t, and the output, y_t, are used to model the process, while in indirect identification, the reference signal and the output are first used to obtain a model of the closed-loop process that includes the controller. Subsequently, the plant model is determined using the controller model. In joint identification, both the plant and controller are simultaneously identified. In general, since direct identification is very similar to open-loop identification, it is often preferred. In certain cases, such as routine operating data, direct identification is the only approach to take.

2.2. Data Quality Assessment and System Identification

When assessing the quality of a data set for system identification, the primary objective is to determine if the data is sufficiently excited for identifying the true parameter estimates.

For the purposes of the developing the theory, let us assume that the data set comes from a single model given as $\left(\vec{y}_t, \vec{u}_t\right)_i$ of length N, where i is the i^{th} sampling point and the subscript arrows denote a vector. Should there be any reason to suspect that the data does not come from a single model, then the data should first be partitioned into regions with similar characteristics and then each region separately analysed for data quality. Finally, assume that the model of interest for the data set is a single-input, single-output (SISO) model with the following form

$$y_t = f\left(u_t, \vec{\theta}\right) \tag{1}$$

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

$$\vec{\theta} = \left\langle \theta_1, \quad \theta_2, \quad \cdots \quad \theta_r \right\rangle \tag{2}$$

Taking the derivative of Equation (1) with respect to the parameters gives the Jacobian, $\mathcal{J}$, which can be written as

$$\begin{aligned} \mathcal{J} &= \left.\frac{\partial f\left(u_t, \vec{\theta}\right)}{\partial \vec{\theta}}\right|_{\substack{\vec{\theta}=\hat{\vec{\theta}} \\ u_t = u}} \\ &= \left.\left[\frac{\partial f\left(u_t, \vec{\theta}\right)}{\partial \theta_1} \quad \frac{\partial f\left(u_t, \vec{\theta}\right)}{\partial \theta_2} \quad \cdots \quad \frac{\partial f\left(u_t, \vec{\theta}\right)}{\partial \theta_p}\right]\right|_{\substack{\vec{\theta}=\hat{\vec{\theta}} \\ u_t = u}} \end{aligned} \tag{3}$$

Evaluating the Jacobian matrix, $\mathcal{J}$, given as Equation (3), for each of the inputs will give the regression matrix, $\mathcal{M}$, that is,

$$\mathcal{M} = \begin{bmatrix} \frac{\partial f\left(u_1,\vec{\theta}\right)}{\partial \theta_1} & \frac{\partial f\left(u_1,\vec{\theta}\right)}{\partial \theta_2} & \cdots & \frac{\partial f\left(u_1,\vec{\theta}\right)}{\partial \theta_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f\left(u_i,\vec{\theta}\right)}{\partial \theta_1} & \frac{\partial f\left(u_i,\vec{\theta}\right)}{\partial \theta_2} & \cdots & \frac{\partial f\left(u_i,\vec{\theta}\right)}{\partial \theta_p} \end{bmatrix} \tag{4}$$

For a linear system, that is, a system where the Jacobian matrix is independent of the parameters, we can write the identification problem as

$$\mathcal{M}\vec{\theta} = \vec{y} \tag{5}$$

The least-squares solution can be obtained by first multiplying Equation (5) by $\mathcal{M}^T$ to give

$$\mathcal{M}^T\mathcal{M}\vec{\theta} = \mathcal{M}^T\vec{y} \tag{6}$$

where $\mathcal{M}^T\mathcal{M}$ is a square matrix, and hence, satisfying one of the requirements for invertibility. Furthermore, it can be noted that $\mathcal{M}^T\mathcal{M}$ is the Fisher information matrix, $\mathcal{F}$, that is,

$$\mathcal{F} = \mathcal{M}^T\mathcal{M} \tag{7}$$

For the nonlinear case, $\mathcal{F}$ can still be calculated, however its value will depend on the parameter values.

In order to obtain unique parameter estimates, Equation (6) needs to be solved. In general, this implies that the inverse of the Fisher information matrix must be found. Therefore, the invertibility of this matrix will determine the uniqueness of the solution.

For an arbitrary $n\times n$ matrix, $\mathcal{L}$, to be invertible, any one of the following conditions must hold (Anton, 2000):

1) $\det(\mathcal{L}) \neq 0$;
2) The eigenvalues of $\mathcal{L}$ cannot be zero; and
3) $\operatorname{rank}(\mathcal{L}) = n$.

All three conditions for invertibility given above are equivalent, that is, if one holds, then the others hold as well (Anton, 2000). Therefore, from a theoretical perspective, it is necessary and sufficient to check either the eigenvalues of $\mathcal{L}$ or the determinant to determine invertibility. However, in addition to the theoretical constraints on invertibility, when dealing with a numerical problem, there is also a need to consider the numerical stability of the matrix, that is, how will small perturbations in the values effect the overall result. Such small perturbations often arise from such factors as measurement noise or unexpected disturbances in the system. Therefore, in addition to checking the theoretical invertibility of the matrix, it would be useful to check the numerical stability of the matrix. One such approach is the condition number of a matrix, $K(\mathcal{L})$, which is defined as

$$K_p\left(\mathcal{L}\right) = \left\|\mathcal{L}\right\|_p \left\|\mathcal{L}^{-1}\right\|_p \tag{8}$$

where $\|\cdot\|_p$ is some matrix norm (Quarteroni, et al., 2000; Quarteroni & Saleri, 2003). Practically, there are three choices for p, namely, $p = 1, 2$, or ∞. Selecting $p = \infty$ tends to produce too conservative bounds for the condition of the matrix (Quarteroni, et al., 2000), although the calculation is rather straightforward. Selecting $p = 2$ is preferred (Quarteroni, et al., 2000). In this case,

$$K_2\left(\mathcal{L}\right) = \frac{\sigma_{\max}\left(\mathcal{L}\right)}{\sigma_{\min}\left(\mathcal{L}\right)} \tag{9}$$

where $\sigma(\mathcal{L})$ are the singular values of the matrix $\mathcal{L}$ (Quarteroni, et al., 2000). If $\mathcal{L}$ is a symmetric positive definite matrix, then, by noting that $\sigma(\mathcal{L}) = |\lambda(\mathcal{L})|$, where $\lambda(\mathcal{L})$ is an eigenvalue of $\mathcal{L}$, Equation (9) can be rewritten as (Quarteroni, et al., 2000)

$$K_2\left(\mathcal{L}\right) = \frac{\max\left(\left|\lambda\left(\mathcal{L}\right)\right|\right)}{\min\left(\left|\lambda\left(\mathcal{L}\right)\right|\right)} \tag{10}$$

A matrix is said to be well-conditioned if K_2 is less than a given threshold, ε. The lower bound for K_2 is 1, which will be achieved when both the maximum and minimum absolute eigenvalues are equal. The upper bound for K_2 is $+\infty$, which is achieved when the smallest eigenvalue is zero, and hence the matrix is uninvertible.

The threshold for the condition number is normally set to be 10^4 (Shardt & Huang, 2013).

Since $\mathcal{F}$ is a symmetric, positive definite matrix, which implies that its 2-norm can be calculated using Equation (10), it follows that the condition number given by Equation (10) can be used to assess the data quality. Therefore, we can define the data quality index, η_{data}, as

$$\eta_{data} = \frac{\max\left(\left|\lambda\left(\mathcal{F}\right)\right|\right)}{\min\left(\left|\lambda\left(\mathcal{F}\right)\right|\right)} = \frac{\max\left(\left|\lambda\left(\mathcal{M}^T\mathcal{M}\right)\right|\right)}{\min\left(\left|\lambda\left(\mathcal{M}^T\mathcal{M}\right)\right|\right)} < \varepsilon \tag{11}$$

A data set 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. Furthermore, it can be noted that a well-conditioned $\mathcal{F}$-matrix implies that the variances obtained for the parameters will be reasonable and hence the results obtained will be significant.

Practically speaking, a threshold value of 10^4 works well. However, the thresholds can be changed depending on the desired properties of the model. Factors such as the desired accuracy of the model, the measurement noise, or model structure can be taken into consideration when selecting the threshold.

2.3. Data Partitioning

It has so far been assumed that the given data set comes from a single operating region so that the assumption of a single (linear) model holds. However, in practice, most data sets contain multiple different regions with varying data structures.

In fact, it is possible to use the above data quality index to partition a given data set. Basically, assume that initial we have k data points, where k is some arbitrary, small number. This number represents the smallest number of data points that we believe is necessary to obtain a good model. Note that this value can depend on the type of data being used, for example, for routine operating data, k could be larger than for open-loop data. For these first k values, compute the value of the data quality index and compare it against the threshold. If the index is below the threshold, add another point and repeat until it fails. The region until failure can be considered to be a single region. If the index is above the threshold, take the next k points and repeat.

2.4. Model Considerations

When implementing the data quality assessment procedure, it can be seen that the type of model selected could have an impact on the assessed value of a given data set. In general, it may not be known which model structure fits the data set the best and there is a need to use a generic model for assessing the data quality. In practice, there exist two different approaches that can be taken (Bittencourt, et al., 2015):

1) **ARX Models**: ARX models are of the form

$$Ay_t = Bu_{t-k} + e_t \tag{12}$$

where A and B are polynomials in z^{-1} of order n_a and n_b respectively and k is the time delay. In order to implement this method, it is necessary to know the time delay k. Since it is known that any prediction error model can be approximated by a high-order ARX model, by selecting high orders for n_a and n_b, the data quality for arbitrary prediction error models can be assessed. The main drawback of this approach is that the time delay must be known.

2) **Laguerre Model**: A Laguerre model is based on the orthogonal Laguerre polynomials, which allows for easy removal of unnecessary model components without affecting the rest of the parameters. The i^{th} order Laguerre basis function, L_i, is

$$L_i\left(z^{-1},\alpha\right) = \frac{\sqrt{1-\alpha^2}}{z^{-1}-\alpha}\left(\frac{1-\alpha z^{-1}}{z^{-1}-\alpha}\right)^{i-1} \tag{13}$$

where α is the time constant, and z^{-1} is the backshift operator. The resulting model can then be written as

$$y_t = \sum_{i=1}^{N_g} \theta_i L_i\left(z^{-1},\alpha\right) u_t + e_t \tag{14}$$

where N_g is the Laguerre order of the process. The advantage of the Laguerre approach is that the time delay needs not be known in order to perform the partitioning. However, the final model that will be fit (often some type of prediction error model) is different from the model used for data quality assessment. This mismatch may lead to issues with the accuracy of the assessment. In practice, the advantage of not needing a time delay often overrides other considerations.

3. DATA QUALITY ASSESSMENT PROCEDURE

Figure 2 shows a schematic overview of the general data quality assessment framework. The details regarding the steps are (Peretzki, et al., 2011; Bittencourt, et al., 2015; Shardt & Brooks, 2018):

1) **Preprocessing**: Load and preprocess the data set. This will often mean scaling and centring the data set.
2) **Mode Changes**: In many industrial systems, the overall system may change its behaviour in a known fashion, for example, operating points may change, faults may occur, or controller setting may be changed. In such cases, it makes sense to incorporate this information into the data quality assessment algorithm. Separating the known changes will mean that the final results will be better. It can be noted that, for example, the number of initial data points required can depend on the control conditions. Therefore, detecting the changes will improve the results.
3) **Partitioning**: For each identified mode, perform the following steps:
    a. **Initialisation**: If the length of the unanalysed data for the given mode is greater than the minimum required length r, set the model counter to the current data point, $k_{init} = k$ and then set $k = k + r$. Otherwise, go to the next identified mode.
    b. **Preprocessing**: For certain types of processes, it may be necessary to perform additional manipulations, for example, for an integrating process, it is necessary to integrate the input.
    c. **Computation**: Compute the required values. In most cases, this will include the variances of the signals and the condition number of the information matrix.
    d. **Comparison**: Compare the variances, the condition number of the regressor matrix, and the significance of the parameters against the thresholds.
        i. **Failure**: If any of the thresholds fail to be met go to the next data point, that is, $k = k + 1$, and go to Step 3.a.
        ii. **Success**: Otherwise, set $k = k + 1$, and go to Step 3.b. The "good" data region is then $[k_{init}, k]$.
    e. **Termination**: The procedure stops once k equals N, the total number of data points in the given mode. Repeat Step 3 for any remaining modes.
4) **Simplification**: It may be desirable to compare adjacent regions and determine if they could be considered to come from a single model. Often the segmentation algorithm will be a bit too strict and provide too many segments (Shardt & Shah, 2014).

In general, a recursive method can be used to compute the required variances, that is, the following update rule is used:

$$\begin{aligned} m_{y_t} &= \lambda_{m_y} y_t + \left(1-\lambda_{m_y}\right) m_{y_{t-1}} \\ \sigma^2_{y_t} &= \frac{2-\lambda_{m_y}}{2}\left(\lambda_{\sigma_y}\left(y_t - m_y\right)\right)^2 + \left(1-\lambda_{\sigma_y}\right)\sigma^2_{y_{t-1}} \end{aligned} \tag{15}$$

where λ is the forgetting factor and σ^2 is the variance of the given signal. The two forgetting factors, λ_{m_y} and λ_{σ_y}, need to be tuned. The variance is updated using the above formulae for

3 different signals, the inputs, outputs, and the regression matrix.

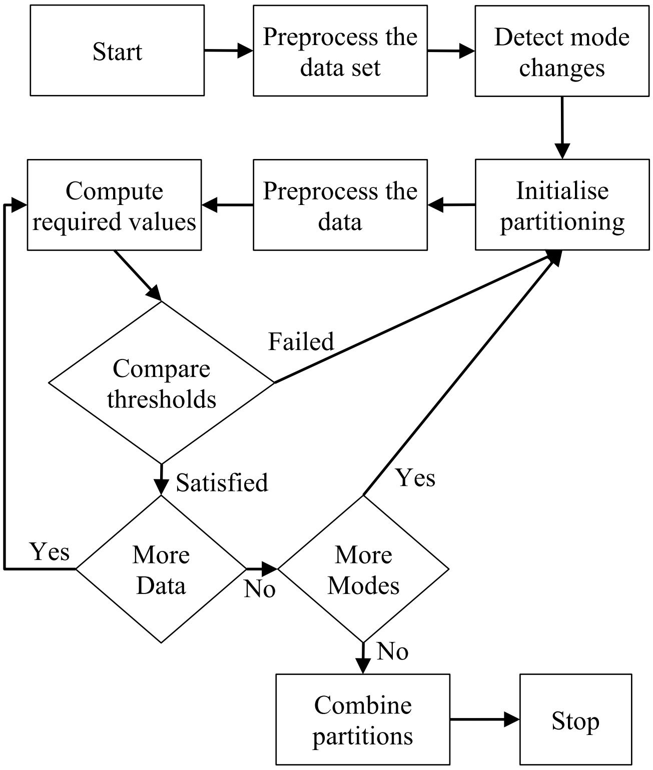

Figure 2: Data Quality Assessment Framework

3.1. Setting the Parameters for Partitioning

As with any method, there are a series of parameters that must be set in order for the method to work. Since the Laguerre-based approach is more commonly used, the primary focus will be on setting the required parameters for this approach.

The Laguerre model parameters, α and N_g, are the two model parameters whose value needs to be set. From (Peretzki, 2010), we have that

$$N_g \geq -\frac{\theta \log(\alpha)}{2\tau_s}+1 \tag{16}$$

where θ is the continuous time delay and τ_s is the sampling time. Previous investigations have shown that α should be set between 0.80 and 0.95 (Shardt & Shah, 2014). When the exact time delay is not known, then an estimated upper bound can be used.

The forgetting factors in Equation (15) also need to be set. Previous investigations suggest that selecting a value of 0.99 for all the forgetting factors works well (Shardt & Shah, 2014).

The minimum required number of data points for identification r can be set based on experience and the required accuracy of the model. For open-loop or externally excited closed-loop data, a value of 20 will suffice. For routine operating data, the value will need to be much larger. In many cases setting $r = 100$ will work.

3.2. Setting the Thresholds for Partitioning

The second important aspect is setting the appropriate thresholds for the partitioning. The success of failure of the assessment can strongly depend on the values selected for the different thresholds.

Firstly, it can be noted that many of the thresholds depend on the properties of the signal and the system at hand. Noisy or systems with large normal variation will require larger thresholds then systems with less noise or small variation. Therefore, it is important that the user take the time to understand the process and implement appropriate bounds.

Secondly, setting conservative thresholds that result in overpartitioning of the data set are probably better than overly loose thresholds that fail to detect such changes. The reason for this is that it is always easier to combine partitions then it is to try and split a partition into multiple partitions.

Often, the thresholds for variances for closed-loop control must be rather small (often on the order of 10^{-7}) due to the fact that a good controller will eliminate most variation in the signal. On the other hand, in open-loop data, the variances can be higher, but even then, consideration needs to be made for such cases as step changes, which could be potentially used for system identification, but whose variance will be small, especially for the input signal.

The threshold for the condition number can be set to the standard value of 10^4. Selecting a different threshold can be based on the desire to vary the quality of the model obtained, for example, a large threshold will decrease the quality of the model, but could allow for identification of more difficult processes.

3.3. Partition Simplification

The proposed data quality assessment procedure tends to overpartition the data set, that is, even if two adjacent regions actually belong to the same model, the procedure will consider them to be different (Shardt & Brooks, 2018).

Furthermore, it would be useful to identify which partitions, even if widely separated, are potentially the same, since these could then be used together for system identification, for example, one region could be the validation and the other the modelling data set.

One of the challenges of this step are that it should be more or less implemented without finding models for the system and comparing them.

One potentially interesting approach is to use an entropy-based metric. It has been shown that the signal entropy value of the difference between the input and output signals can be used to monitor a process and determine if it changes (Shardt & Huang, 2013). The entropy of a signal, which measures the amount of information in a signal, is given as

$$H = \log\left(\frac{\sum_{k=1}^{N}\left|x_k - x_{k-1}\right|}{N}\right) \tag{17}$$

where H is the entropy, N is the signal length, and x is the signal of interest. The difference in entropy would then be calculated as

$$\Delta H = H_y - H_u \tag{18}$$

where H_y is the entropy of the output signal and H_u the entropy of the input signal. Assuming that the input signal is always a pseudorandom signal or a white, Gaussian noise signal, then the difference between the input signal entropy and output signal entropy will be constant and equal to the model entropy. The advantage of this approach is that it simply requires the computation of a difference of values for the two signals. Instead of monitoring the complete signal, it is also possible to use a moving window approach where only the last N values are considered.

Another approach to solving this challenge is to consider various clustering algorithms, which can be used to compare the partitions and determine which ones are similar.

3.4. Multivariate Considerations

The last area of consideration is multivariate data sets. Although all of the above methods easily generalise to the multivariate case, there are some additional challenges.

First, determining which of the variables should be used for data partitioning is a large question (Arengas & Kroll, 2019; Shardt & Brooks, 2018). If the wrong set of variables is used, then it is possible that the method will fail or give an incorrect result. It should be noted that selecting all available input variables may not be efficient, since some of these inputs could be correlated with each, which will mean that the resulting Fisher information matrix is uninvertible (as it should be given the circumstances). However, the correlations and relationships between the variables can change depending on the mode or circumstances, so that it is now necessary to bear this in mind.

4. INDUSTRIAL EXAMPLE

4.1. Process Description

Before considering the actual implementation of the data segmentation system, it will be useful to briefly examine the actual system considered.

The data used was obtained from a section of the lead zinc concentrator at the Mount Isa Mines in Queensland, Australia. The concentrator is a complex operation, recovering both lead and zinc from a feed sourced from three different mines. The ore is milled and is then fed to a lead removal circuit. The lead is recovered in the form of a concentrate. The reject stream from this unit, termed the tailings, is fed to a zinc flotation unit. In this circuit, a number of banks of flotation cells, are used to recover the zinc. As shown in Figure 3, these banks are named the roughers, scavengers and recleaners.

The section of the circuit covered here is the zinc roughers (Brooks & Koorts, 2017). The rougher tails from the upstream lead circuit are the feed to the zinc roughers. As shown in Figure 4, this bank consists of four cells (FC23, FC24, FC25, FC26). The objective of this bank is to perform a rough separation of zinc from the waste material. Copper sulphate (activator) and naphthalene sulphate (depressant) are added upstream. Ethyl xanthate, a collector, is added to cells FC23 and FC25. The tails of the rougher (unfloated material) report downstream to the scavengers where the majority of the remaining zinc is floated. The concentrate (floated material) from the roughers reports to the recleaners.

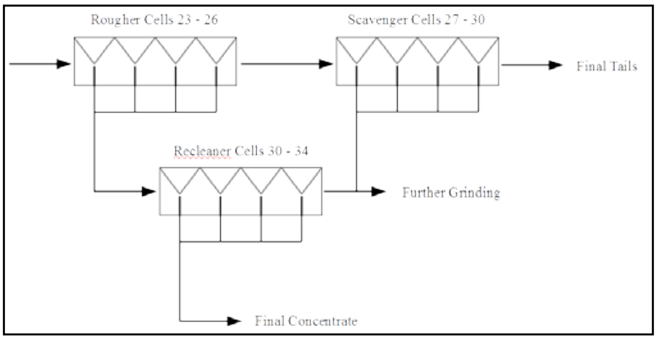

Figure 3: Zinc rougher, scavenger and recleaner circuit.

In the rougher bank, levels are controlled per pair of cells. The flowrate of air can be varied on a per cell basis. Composition measurement by X-ray fluorescence (XRF) is used on all concentrate and tails streams. In Figure 4, LC1 and LC2 are level PID controllers on pairs of cells, FC1 to FC4 are flow PID controllers on air flowrates and FC5 to FC8 are reagent flow PID controllers. FI1 is the volumetric feed flowrate. Analysers AI1 to AI3 measure zinc percentages in the feed, concentrate, and tails respectively.

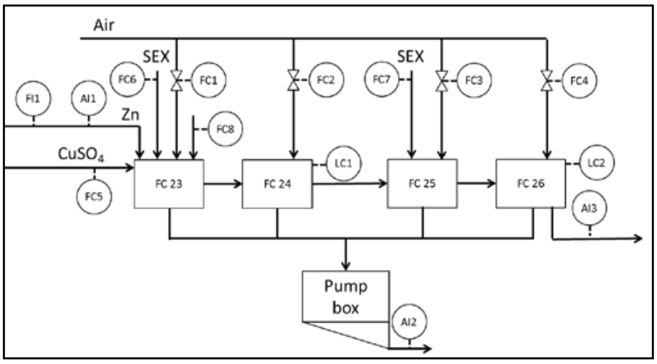

Figure 4: Rougher Bank Showing Control Loops and Analysers

4.2. Data Set Characteristics

The data collected for this investigation consists of two months of plant operation, collected at a frequency of one minute. The historian's interpolation routine is used to ensure the data is aligned. No special care was used to ensure that the data had any particular characteristics, other than that the plant was running. It was reported that during this period some step tests had been conducted. The completed data set can be downloaded from https://doi.org/10.5281/zenodo.3701260.

Forty-three variables were collected: for each of the PID controllers, setpoint, process value and output (SV/PV/MV) were recorded. The three analysers provide measure of iron, lead and zinc percentages. Variables collected are listed in Table 1. The process was assumed to be running under control throughout the period of investigation.

The ultimate goal of the models is to design a model predictive controller (MPC) for the unit. The manipulated variables (MVs) are the air flows, levels, and the flows of the reagents. The outputs or controlled variables (CVs) are the zinc percentages in the concentrate and tailing streams. The primary focus of this investigation is on the zinc percentage in the concentrate stream. Similar results are expected for the other situations. The focus is on the multivariate nature of the data set.

Table 1: Test Variables

Tag	Attributes	Description
FC1	SV/PV/MV	Air flow to FC23
FC2	SV/PV/MV	Air flow to FC24
FC3	SV/PV/MV	Air flow to FC25
FC4	SV/PV/MV	Air flow to FC26
FC6	SV/PV/MV	EX (reagent) to FC23
FC7	SV/PV/MV	EX (reagent) to FC25
LC1	SV/PV/MV	FC24 Level
LC2	SV/PV/MV	FC26 Level
FC8	SV/PV/MV	NS (reagent) to FC3
FC5	SV/PV/MV	CuSO4 (reagent) to FC22
AI2	Fe/Pb/Zn	Primary Rougher Concentrate Compositions
AI3	Fe/Pb/Zn	Primary Rougher Tailings Compositions

The following 3 situations will be considered:

1) **Variable Selection**: which of the input variables should be selected for data partitioning, since not all of the selected variables may be independent.
2) **Reduction of Partitions**: determining of some or all of the regions could be modelled by similar models.
3) **Model Validation**: Using some of the suggested partitions, models will be fit and compared.

4.3. Variable Selection

One of the most important issues in multivariate data quality assessment is determining which of the potential variables can or should be used for partitioning the data set. One of the main issues is the selecting a set of independent variables. In order to examine the situation, the following 3 cases will be considered:

1) **Case 1**: using all the input variables to partition the data set.
2) **Case 2**: using a subset of variables based on correlation analysis.
3) **Case 3**: using a subset of variables based on user selection.

For Case 1, the results are shown in Figure 5. In Figure 5, the top figure shows the actual measured zinc concentration in the concentrate stream. It should be noted that for a series of values around 6.2×10^4 min, the value went to $-10{,}000$, which is an impossible value for concentration, suggesting that the process was not running at this point. Therefore, these extreme values have been replaced by -1 in the top figure for ease of display. The original values were used for the data partitioning part. The bottom figure shows the partitioned data. The programme assigns the same partition number to adjacent points if they are assumed to belong together. A separate number implies that the points do not belong together. The jumps in the value arise from the way the programme reacts to values going to zero. It is assumed that since the process at these points is not working properly it resets the counter. The goal is to find plateaus in the partitioning graph that represent the regions of sufficient excitation. From Figure 5, it can be seen that there are few if any plateaus. This implies that either the data itself is not sufficiently excited or that some of the variables used are correlated with each other. If we examine the correlation plot shown in Figure 6, we can quickly see that many of the variables are strongly correlated with each other. The variables are ordered the same way as in Table 1, so that the first variable is the air flow to FC23 and the last variable is $CuSO_4$ to FC22. It should be noted that all the variables are strongly correlated with each other. However, some are much more strongly related than others, for example, variables 5 to 10 are all correlated with a value close to 1. This suggests a very strong relationship between the variables. As well, note that variables 3 and 4 are also strongly correlated.

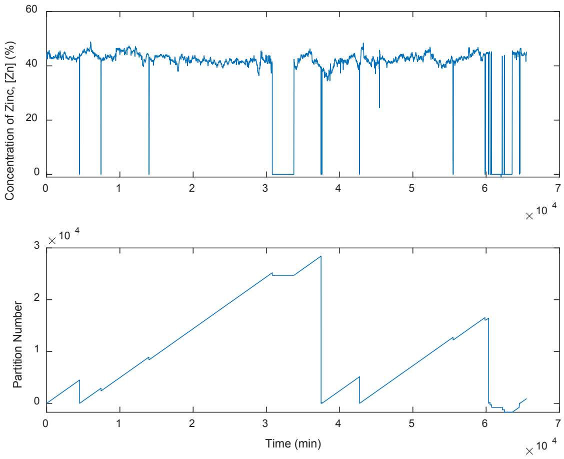

Figure 5: Data Partitioning for Case 1: Using all Available Variables

Using the results obtained from Figure 6, the variables for Case 2 will be defined as variables 1, 2, and 4, that is the first two flow rates and the reagent to FC23. The results are shown in Figure 7. It can now be seen that more additional regions can be found and that the partitioning seems to align better with the actual results.

Finally, Case 3 will consider the case of simply using the first three flow rates, that is, the first three variables, for partitioning the data. From Figure 6, we can see that these three variables are also independent of each other raising the question if they too can provide good results. Figure 8 shows the results. Comparing with the previous case, we can see that the two results are similar. This suggests that at least for the example considered that the variables selected for partitioning do not matter as long as the variables are independent of each other for the given data set.

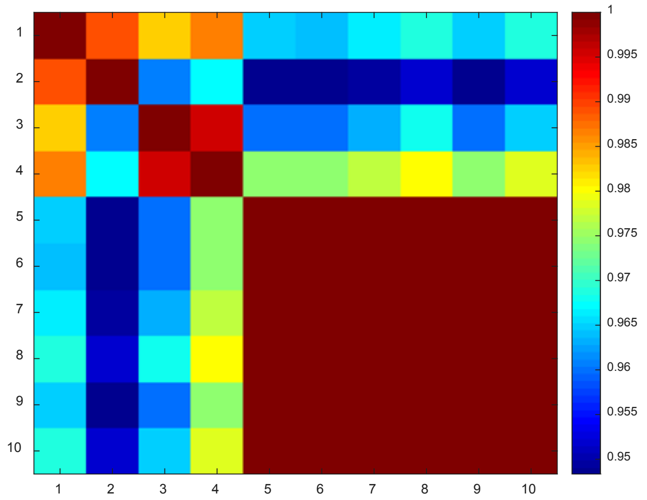

Figure 6: Correlation Plot for the Variables of Interest

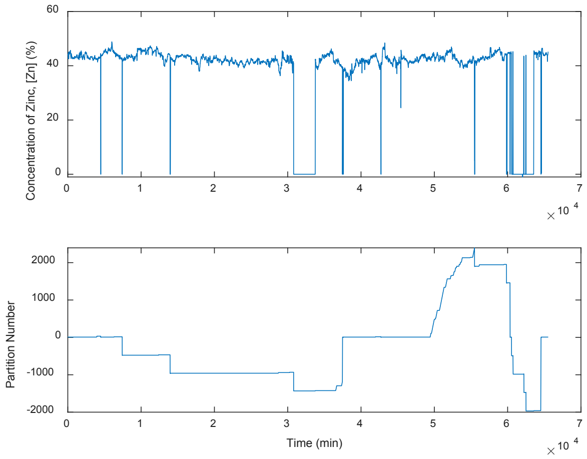

Figure 7: Data Partitioning for Case 2: Using Uncorrelated Variables

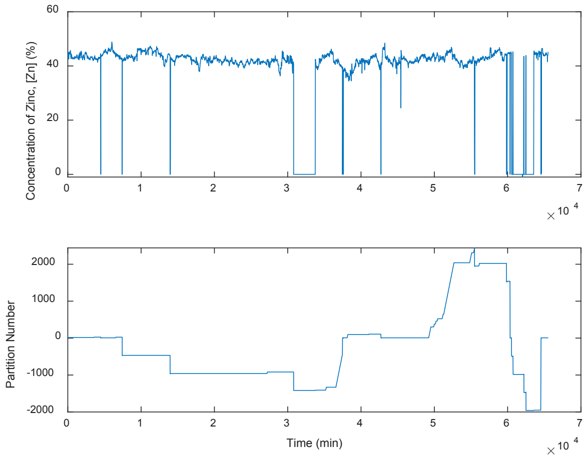

Figure 8: Data Partitioning for Case 3: Using Only the First Three Air Flow Rates

4.4. Reduction of Partitions

Taking Case 2 from the variable selection situation, it now desired to investigate the impact of the reduction of partitions on the overall results. In the previous results, the number of partitions was reduced using the entropy-based method. Here the results with and without partition reduction will be compared.

Figure 9 shows the partitioning results for Case 2 but without any reduction of partitions. It can be seen that there are now more partitions and some of the partitions are separated as belong to different potential models. By combining adjacent partitions, it is possible to increase the amount of available data and create potentially better models. Therefore, it makes sense to determine if adjacent partitions could belong to the same overall model.

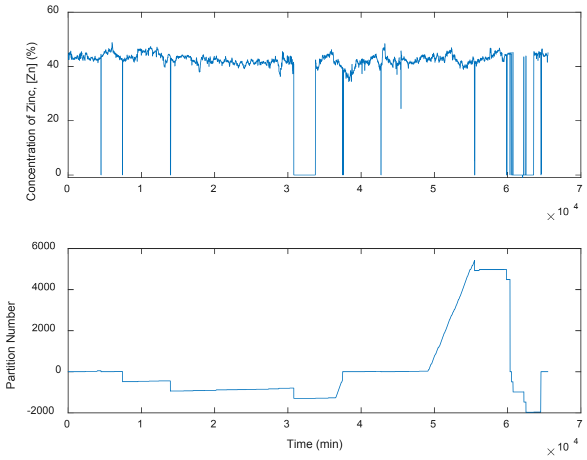

Figure 9: Data Partitioning for Case 2: Using Uncorrelated Variables and No Reduction of Partitions

4.5. Model Validation

Using the data segmentation results from above, the different, the following regions will be considered:

- S1: 17,079−18,631
- S2: 18,645−20192
- S3: 20,2017−21,650
- S4: 21,670−23,112
- S1′: 14,038−28,738
- S2′: 17,079−23,112

S1, S2, S3, and S4 are the initial subpartitions of S1′ and S2′. The difference between S1′ and S2′ lie in exactly which endpoints are considered and the exact reduction values are used.

The data for each section was modelled using a commercial package that uses canonical variate analysis (CVA) (Larimore, 1990; Zhao, et al., 2006). Multiple input, single output models with a settling time of 60 min were obtained for the zinc percentage in the concentrate.

Table 2 shows the results for the different sections. It can be noted that in the development of models for use in model predictor control, it is the gains that are considered to be important. Therefore, the focus is on the accuracy of the gains. Furthermore, it should be noted that these models are to be used as seed models for providing the initial model parameters for subsequent online identification.

From Table 2, it can be seen that the models for the larger partitions S1′ and S2′ compare well, although the fits are not particularly good. Partitions S3 and S3 have the same signs for the gains, although the negative gain for the feed zinc is physically unrealistic. Partition S1, although having the highest correlation coefficient, has a positive gain for total xanthate, which is not what is found in practice. The practical issue is that the sample sets lengths of around 24 hours for S1 to S4 are too short to derive reliable linear time invariant models. Set S1′ is 240 hours, and S2′ 100 hours, so that the CVA produces better models in these cases. It is very encouraging for the method that the set S2′, which is a section of S1′, produces very similar models. It would appear that for the purposes of dynamic model identification heuristics need to be added to the algorithm specifying a minimum dataset length. Thus, the ability to combine adjacent partitions is an important aspect of any data segmentation method.

Table 2: Comparison of Models for Different Partitions

Partition		S1	S2	S3	S4	S1′	S2′
Gain	Total xanthate	0.83	0.01	−0.29	−1.05	−0.34	−0.31
	Feed flow	−0.02	0.001	−0.01	−0.01	−0.002	−0.004
	Feed zinc	0.34	−0.31	−0.15	−0.10	0.11	0.11
	Feed percent solids	−0.16	−0.21	−0.03	−0.07	−0.10	−0.11
Mean Squared Error		1.11	0.76	1.30	0.62	1.26	1.04
R^2		0.84	0.19	0.69	0.49	0.06	0.29

5. CONCLUSIONS

This paper has examined the field of data quality assessment and its recent successes. A general data quality assessment algorithm was proposed and the details of setting its parameters examined. Previous work has shown that selecting appropriate thresholds can impact the accuracy and speed of the resulting algorithm. Furthermore, extending the results to them multivariate situation introduces new challenges including how best to select the variables for segmentation. Selecting the wrong subset of variables can lead to issues with collinearity between the variables.

The proposed data quality assessment algorithm was validated using data extracted from a historian for a zinc flotation cell. It was shown that the partitioning depended strongly on the variables selected and the methods used to reduce the number of partitions.

Future work will focus on generalising the results to the multivariate case and providing better methods for combining adjacent partitions.

ACKNOWLEDGEMENTS

Yuri A.W. Shardt and Xu Yang would like to thank the National Natural Science Foundation of China under grant #61673053, the Fundamental Research Funds for the Central Universities under Grant #FRF-BD-19-002A, and the National Key R&D Programme of China under grant #2017YFB0306403 for funding.

REFERENCES

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

Arengas, D. & Kroll, A., 2017. *A search method for selecting informative data in predominantly stationary historical records for multivariable system identification.* Sinaia, Romania, IEEE.

Arengas, D. & Kroll, A., 2017. *Searching for informative intervals in predominantly stationary data records to support system identification.* Sarajevo, Bosnia-Herzgovina, IEEE.

Arengas, D. & Kroll, A., 2019. *A Data Selection Method for Large Databases Based on Recursive Instrumental Variables for System Idnetification of MISO Models.* Naples, Italy, IEEE.

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

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

Basseville, M. & Nikiforov, I. V., 1993. *Detection of Abrupt Changes: Theory and Application.* Englewood Cliffs (New Jersey): PTR Prentice-Hill, Inc..

Bittencourt, A. C., Isaksson, A. J., Peretzki, D. & Forsmann, K., 2015. An Algorithm for Finding Process Identification Intervals from Normal Operating Data. *Processes,* 3(2), pp. 357-383.

Brooks, K. S. & Koorts, R., 2017. *Model Predictive Control of a Zinc Flotation Bank Using Online X-ray Fluorescence Analysers.* Toulouse, France, Elsevier.

Ding, S. X., 2014. *Data-driven design of fault diagnosis and fault-tolerant control systems.* 1 ed. London: Springer.

Ding, S. X. et al., 2013. A Novel Scheme for Key Performance Indicator Prediction and Diagnosis With Application to an Industrial Hot Strip Mill. *IEEE Transactions on Industrial Informatics,* November, 9(4), pp. 2339-2247.

Huang, B., 2003. A pragmatic approach towards assessment of control loop performance. *International Journal of Adaptive Control and Signal Processing,* Volume 17, pp. 589-608.

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

Khatisbisepehr, S. & Huang, B., 2008. Dealing with Irregular Data in Soft Sensors: Bayesian Method and Comparative Study. *Industrial & Engineering Chemistry Research,* 47(22), pp. 8713-8723.

Klimchenko, V. V., Samotylova, S. A. & Torgashov, A. Y., 2019. Feedback in a Predictive Model of a Reactive Distllation Process. *Journal of Computer and Systems Sciences International,* Volume 58, pp. 637-647.

Larimore, W. E., 1990. *Canonical variate analysis in indentification, filtering, and adaptive control.* Honolulu, Hawaii, USA, IEEE.

Li, J., Shi, R., Xu, C. & Wang, S., 2019. Process identification of the SCR system of coal-fired power plant for de-NOx based on historical operation data. *Environmental Technology,* 40(25).

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

Mehrkanoon, S., Falck, T. & Suykens, J. A. K., 2012. *Parameter Estimation for Time Varying Dynamical Systems using Least Squares Support Vector Machines.* Brussels, Belgium, s.n., pp. 1300-1305.

Peretzki, D., 2010. *Data mining for process identification (Diploma Thesis),* Cassel, Germany: University of Cassel.

Peretzki, D., Isaksson, A. J., Bittencourt, A. C. & Forsman, K., 2011. *Data Mining of Historic Data for Process Identification.* Minneapolis, Minnesota, United States of America, AIChE.

Quarteroni, A., Sacco, R. & Saleri, F., 2000. *Numerical Mathematics.* Secaucus (New Jersey): Springer.

Quarteroni, A. & Saleri, F., 2003. *Scientific Computing with MATLAB.* Berlin: Springer-Verlag.

Sha'aban, Y. A., 2019. Model predictive control from routine plant data. *IFAC Journal of Systems and Control,* Volume 8.

Shardt, Y. A. W., 2012. *Data Quality Assessment for Closed-Loop System Identification and Forecasting with Application to Soft Sensors,* Edmonton: University of Alberta.

Shardt, Y. A. W. & Brooks, K., 2018. *Automated System Identification in Mineral Processing Industries: A Case Study using the Zinc Flotation Cell.* Shenyang, China, Elsevier, pp. 132-137.

Shardt, Y. A. W. & Huang, B., 2013. Data quality assessment of routine operating data for process identification. *Computer and Chemical Engineering,* Volume 55, p. 19–27.

Shardt, Y. A. W. & Huang, B., 2013. Statistical properties of signal entropy for use in detecting changes in time series data. *Journal of Chemometrics,* 27(11), pp. 394-405.

Shardt, Y. A. W. & Shah, S. L., 2014. *Segmentation Methods for Model Identification from Historical Process Data.* Cape Town, South Africa, Elsevier, pp. 2836-2841.

Shardt, Y. A. W. et al., 2012. Determining the State of a Process Control System: Current Trends and Future Challenges. *Canadian Journal of Chemical Engineering,* April.pp. 217-245.

Wang, J., Su, J., Zhao, Y. & Zhou, D., 2018. Searching historical data segments for process identification in feedback control loops. *Computers & Chemical Engineering,* Volume 112, pp. 6-16.

Yang, X. et al., 2019. A KPI-Based Soft Sensor Development Approach Incorporating Infrequent, Variable Time Delayed Measurements. *IEEE Transactions on Control Systems Technology,* p. (in press).

Zhao, H., Harmse, M., Guiver, J. & Canney, W. M., 2006. *Subspace identification in industrial APC applications−a review of recent progress and industrial experience.* Newcastle, Australia, IFAC.