Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Segmentation Methods for Model Identification from Historical Process Data

Yuri A.W. Shardt[*†], **Sirish L. Shah**[*]

**University of Alberta, Edmonton, Alberta, Canada, T6G 2V4*
(e-mail: {yuri.shardt, sirish.shah}@ ualberta.ca)
†Currently, Yuri A.W. Shardt is an Alexander von Humboldt Fellow at the University of Duisburg-Essen,
Bismarkstraße, 81, Duisburg, North Rhine-Westphalia, Germany (e-mail: yuri.shardt@uni-due.de).

Abstract: In industry, system identification is a time-consuming exercise that impacts the profitability and safety of the plant. One way to avoid this problem is to use stored historical process data in estimating the required process models. Given the large amount of data available, which is often corrupted due to process disruptions, loss of information, and poor data quality, automated segmentation of the data set would be an invaluable asset. Recently, two different methods have been proposed to accomplish this task: one based on Laguerre models and one based on autoregressive with exogenous input (ARX) models. In this paper, the Laguerre approach will be analysed and it will be shown that the results are dependent on selecting appropriate Laguerre model parameters and input signals, while relatively insensitive to the variance thresholds. Furthermore, this approach has a tendency to overpartition the data set based on the smallest changes in the process. Therefore, in order to decrease the number of models identified it is proposed to couple this method with an entropy-based metric for determining similar models. Based on simulations that include this entropy metric, it is shown that a reduction in model partitions is obtained.

Keywords: **system identification, data mining, segment reduction**

1. INTRODUCTION

In many chemical engineering plants, when implementing process control, system models need to be found. Although system identification experiments are often necessary, from the perspective of the plant engineers, they can lead to undesirable deviations from normal operation that have economic and safety consequences. Furthermore, given the relatively wide availability of historical data, it can also be unnecessary. Therefore, determining the usefulness of the stored data is an important consideration for industrial applications. It is generally known that an entire time series may include periods of abnormal activity or process or sensor malfunction. Determining the windows of informative or useful data is known as data segmentation. Segmentation methods also depend on the end use of the data, for example, in this study we are concerned with estimation of Laguerre models from the segmented data.

Initial approaches to resolving this problem considered detecting transients (Horch, 2000), analysing the impact of selected data regions (Carrette, et al., 1996), or determining segments suitable for the design of inferential controllers (Amirthalingam, et al., 2000). More recently, two new approaches have been proposed for determining the suitability of a given data segment for control purposes, especially identification. The first method developed by Peretzki *et al.* (2011) uses Laguerre models as the model basis for extracting the desired model conditions. The data quality is assessed based on the validity of the process model and the significance of the model parameters. The key advantage of this approach is that the process time delay is not required. Unfortunately, this method only works with data obtained under open-loop or closed-loop conditions where the reference signal changes. The second approach developed by Shardt and Huang (2013) uses a condition number based on fitting an autoregressive model with exogenous input (ARX) to the data to determine the quality. The key advantage of this approach is that it can be applied to any operating conditions, including closed-loop without any excitations in the reference signal, but excitations in the disturbance signal, that is, it can use routine operating data. Such data sets are plentiful in many industrial applications. On the other hand, it does require knowledge of both the process orders and time delay in order to estimate the condition number of the data matrix.

Since processes exhibit different characteristics depending on operating conditions and therefore require different models for each regimen, one recurring problem in many data segmentation methods is determining the appropriate number of models in a given data set. Three different situations can be identified: oversegmentation, undersegmentation, and exact. In oversegmentation, too many models are identified compared to the true number, while in undersegmentation too few models are identified. Finally, in the exact case, the true number of models is determined. Not only must the true number of models be correct, but the points at which the models change should also be correctly identified. Such a point will be called a segmentation point. If a method can correctly determine the segmentation points, but overpartitions the data set, then model reduction can achieve the desired model segmentation.

Therefore, this paper proposes to analyse the Laguerre-based approach in order to understand the key parameters affecting the data segmentation methods. This analysis will show that one of the main issues is oversegmentation. In order to resolve this problem, an entropy-based measure of the data set is proposed to reduce the model partitions.

2. LAGUERRE MODEL-BASED DATA SEGMENTATION

2.1 Background Information

The Laguerre model-based data segmentation method uses, as its name suggests, Laguerre models. Such models have the advantage that they can implicitly incorporate the time delay into the form of the Laguerre order selected. As well, a Laguerre model basis set is orthogonal to each other, which implies that the required models can be easily removed from the analysis without affecting the rest of the model parameters. The i^{th} order Laguerre model can be written as

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

where L_i is the i^{th} order Laguerre basis function, α is a time constant, and z^{-1} is the backshift operator. The model identified in this approach can then be written as

$$y(t)=\sum_{i=1}^{N_g}\theta_i L_i\left(z^{-1},\alpha\right)u(t)+e(t) \tag{2}$$

where $y(t)$ is the output signal, $u(t)$ is the input signal, $e(t)$ is the error, θ_i is the to-be-determined coefficient, and N_g is the Laguerre order of the process. The model given by Equation (2) can be solved using standard least-squares analysis.

The Laguerre-based procedure can be summarised as the following series of steps (Peretzki, et al., 2011):

1) Load, scale, and centre the data.
2) Determine any changes in operating points.
3) For each operating point, perform the following steps:
    a. If the process is an integrator, integrate the input. Compute the Laguerre basis functions, variances, and regressor matrix.
    b. Initialise the region counter to the current data point, $k_{init} = k$.
    c. Compare the variances, the condition number of the regressor matrix, and the significance of the parameters against the thresholds. 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.b. Otherwise, set $k = k + 1$, and go to Step 3.c. The “good” data region is then $[k_{init}, k]$.
    d. The procedure should be stopped once k equals N, the number of data points in the given operating region.

The required variances are obtained using the following update rule:

$$\begin{aligned} m_{y_t} &= \lambda_{m_y} y_t+\left(1-\lambda_{m_y}\right)m_{y_{t-1}} \\ \sigma_{y_t}^2 &= \frac{2-\lambda_{m_y}}{2}\left(\lambda_{\sigma_y}\left(y_t-m_y\right)\right)^2+\left(1-\lambda_{\sigma_y}\right)\sigma_{y_{t-1}}^2 \end{aligned} \tag{3}$$

It can be seen from Equation (3) that for each variance update rule, there are two tunable forgetting parameters: λ_{m_y} and λ_{σ_y}. Furthermore, it can be seen from Equation (3) that there are 3 threshold parameters that check the value of the variance. Since the behaviour of these 3 parameters is predictable, namely, increasing the threshold will decrease the number of quality regions and increase the number of rejected regions, they will be excluded from the analysis. It should be noted that setting these parameters can be quite difficult as the variability of the signals, even after normalisation can vary drastically. Therefore, for the purposes of this paper, the following parameters are of interest:

1) Laguerre Model Parameters, of which there are 2: α and N_g. For the cases considered, α will be taken from the interval [0.30, 0.95] and N_g from [1, 10].
2) Forgetting Parameters, of which there are 6, two each for the 3 variances, u, y, and R. Each forgetting parameter will be taken from the interval [0.85, 0.99].

To illustrate the usefulness of this method, first-order, second-order, including inverse response and underdamped, zero-order, and a nonlinear tank system simulation will be considered. The tank system simulation is based on the results presented in (Shardt, 2012). For each case, a range of different parameters and input signals was considered. For the input signal, white Gaussian noise, step changes, and pseudorandom binary signals (PRBS) were used. Both open-loop and closed-loop cases with excitation in the reference signal were considered. Given the large number of different combinations, all of the simulations were performed automatically in MATLAB and appropriate summary figures produced. Summary graphs are given in this paper to illustrate the results obtained.

2.2 Selecting the Laguerre-Model Parameters

Since the Laguerre model is important for performing the data partitions, selecting appropriate parameter values is crucial. There are two parameters to consider: the time constant α and the order of the system, N_g. According to (Peretzki, 2010)

$$N_g \geq -\frac{\theta\log(\alpha)}{2\tau_s}+1 \tag{4}$$

where θ is the continuous time delay and τ_s is the sampling time. For the example, consider 4 different first-order systems as shown in Table 1, whose generic transfer function can be written as

$$G=\frac{K}{\tau s+1}e^{-\theta s} \tag{5}$$

Each model was simulated for 300 seconds and then concatenated to form a single large data set containing 1200 samples. The expected partitions should occur at sample instants 300, 600, and 900 s.

These simulations were all performed in open-loop conditions using three different input signals: white Gaussian noise, step, and pseudorandom binary signal (PRBS). The parameters vary as previously described. From Equation (4), this implies that the required order should lie between 3 and 22. Since the values stop at 10, it is expected that at low values of α, segmentation will be poor.

Table 1: First-Order Model Parameters

Model	Process Gain, K	Time Constant, τ	Deadtime, θ
I	1.54	60	5
II	1.54	20	40
III	1.54	60	40
IV	1.10	60	40

For both the Gaussian noise and step test results (not shown), an incorrect number of models is identified. In the Gaussian case, too many models are found, while in the step test case, too few are found. On the other hand, for the pseudorandom binary signal, the results are shown in Figure 1. The figures show model partition number as a function of both α and N_g. The model partition number assigns a number to each data point at each time stamp. If two adjacent points have the same model partition number (*y-axis*), then the method considered the given region to be from the same underlying model. It can be seen that the partition points are at the correct values (to within the time delay) at 300, 600, and 900 s. Furthermore, it would seem that irrespective of N_g, the results are similar. On the other hand, there seem to be a few lines that veer away from the main blocks. Secondly, it would seem that for the region between 300 and 600 s, the method cannot partition the data accurately. It is interesting to note that in this case, both the time constant and time delay changed abruptly. This suggests that potentially the method is unable to deal with multiple simultaneous changes. Figure 2 shows the number of models identified for a given N_g and α. A model was defined as a region of at least 100 samples whose model partition number were all the same. From this figure, it can be seen that for α close to 1, the results are better given the constraints on N_g. This shows that in open loop Inequality (4) holds. Selecting a sufficiently large α implies that the effect of time delay can be minimised and a relatively small order used.

Second-order systems behave similarly to the first-order systems shown in Figure 1 and Figure 2, even in the presence of an inverse response. The main difference is that the effect of α on the system is much more pronounced, so that at low values, the mismatch is large and over 100 partitions can be determined. Both inverse response and oscillatory response do not tremendously impact the ability of the system to partition the set.

For integrating processes, before the method can be used, the input signal must be integrated. Figure 3 shows the total number of partitions with size greater than 100 for the integrating case. This figure shows that the behaviour is quite different. Firstly, none of the combinations come even close to giving the correct number. It can be noted that those regions at zero represent cases where for the whole length there was no partition with size greater than 100. Secondly, the optimal values that produce the lowest number of partitions occur around $\alpha = 0.55$, which is a small value. Given that the model had a time delay of either 5 s (for the first partition) or 40 s (for the other two), Inequality (4) is not satisfied. As well, if integrating processes are present, for example, a level loop, then the identification algorithm needs to use different calibration values.

Finally, consider a heated tank as described in (Shardt, 2012), where it is desired to develop models between the temperature and steam flow rate, with the level being a disturbance. It should be noted that although the overall system is nonlinear, at any given operating point, its behaviour is very well modelled by a first-order plus deadtime model. Furthermore, process changes are introduced by changing the height in the tank, which in turn causes changes in both the gain and time constant. As was previously noted with simultaneous changes in the first-order system and the lack of segmentation, we see a similar, but less pronounced, situation. Instead of having no partitions identified, about 10 different partitions are determined.

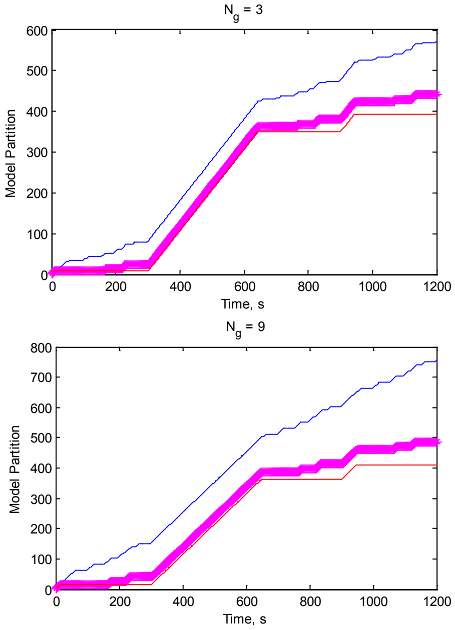

Figure 1: Effect of α and N_g for first-order processes excited by a PRBS signal. Only some of the results are shown due to space constraints (dashed, blue is α = 0.3, thick pink α = 0.5, and thin, red α = 0.95). Adjacent points with the same model partition number were assigned to the same model

Finally, under closed-loop conditions with external excitation, the results are similar to those previously obtained, except that more partitions are determined. The correct change time (to within time delay) is also found.

2.3 Tuning the Forgetting Parameters

The final set of tunable parameters is the forgetting factors, which by definition vary between 0 and 1. Taking the best results for the Laguerre parameter values of $\alpha = 0.9$ and

$N_g = 10$, each of the two tunable parameters will be swept from 0.85 to 0.99 in increments of 0.01 to determine if these parameters have any effect on the segmentation obtained. In general, in open-loop, these parameters do not have any effect on the segmentation obtained. In the worst case, at the extremes of the selected regions, there may be a slight spike to either a larger or smaller number. The largest changes are seen in the computing the variance for the input signal, while the smallest changes are seen with computing the variance of the output signal. In the closed-loop case with excitation of the reference signal, the results are the same as for the open-loop case.

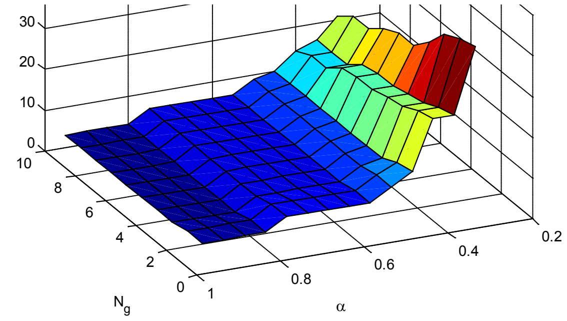

Figure 2: The number of models identified as a function of α and N_g for first-order models and a PRBS input.

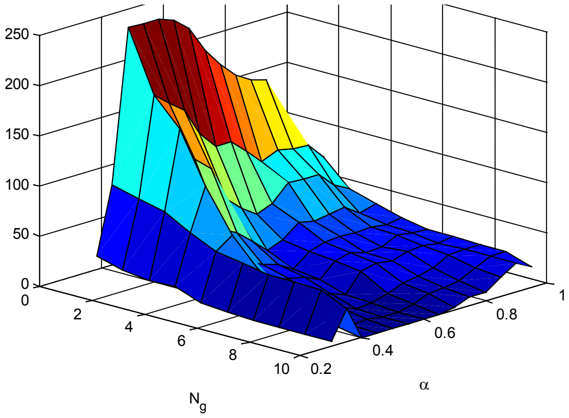

Figure 3: The number of models identified as a function of α and N_g for integrating models and a PRBS input

2.4 Tuning the Thresholds

The selection of appropriate thresholds can have an impact on the quality of the segmentation results, even after normalising the signals. The reason for this impact is that normalising assumes that the underlying signal is at least seminormally distributed. However, in practice, such signals could have regions of abnormal operations that could negatively impact the results. If there are a significant number of extreme values, these can skew the mean and variance, which could give very small values for the main component of the process. These thresholds need to be set based on the process values and conditions that are actually present in the system. Step tests, as will be shown in the experimental section, need much lower variances and higher condition number thresholds than for other methods.

3. IMPROVED METHOD

Based on the above analysis of the system, it can be seen that one of the issue is with the number of segments obtained. Since it has been shown that the method can accurately determine when the model truly changed, it remains to somehow determine a method that can reduce the number of segments between the transitions. One potentially interesting approach is to use an entropy-based metric. Recently, Shardt and Huang (2013) showed 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. The entropy of a signal, which measures the amount of information in a signal, is given as

$$H = \log\left(\frac{\sum_{k=1}^{N}\left|x_k - x_{k-1}\right|}{N}\right) \tag{6}$$

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

$$\Delta H = H_y - H_u \tag{7}$$

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. Therefore, it is suggested that instead of simply segmenting based on the above results, an extra step be added at the end of the procedure, so that the entropy of each segment is computed and compared against adjacent values. If the entropy is similar, then the models can be combined, and if they are different then it can be considered that the partitions indeed do represent regions with different underlying dynamics. The modification to the general segmentation procedure involves adding the following step:

4) After all the regions have been determined, compute the entropy for each region. If the entropy of two adjacent regions is within a threshold, treat the two regions as one.

4. EXPERIMENTAL COMPARISON

The newly proposed modifications to the method will be tested using a temperature-steam control loop in both open and closed-loop. The data used in this experiment were extracted from the DeltaV historian without taking into consideration any preconceived time frames, that is, the data was extracted for some amount of time irrespective of whether or not the process was working or even in a given mode. Two different cases will be considered: open-loop and closed-loop experiments. A schematic of the process is shown in Figure 4.

4.1 Open-Loop Results

For the open-loop experiment, 3 hours of data were extracted from the data historian from a region in which it was known that step tests were being made. Both the original and refined methods were tested on this data set. The values of the Laguerre model parameters were varied to verify whether or not the same results hold. Manual analysis of the data shows that a single operating point had been selected on which multiple step tests were performed. Therefore, it is expected that a single region should be identified, as all the data could be used for identification. The thresholds were modified appropriately given the initial data set. The results are presented in Figure 5 for the original method and in Figure 6 for the proposed refined method. It can be noted that the previously observed conditions hold for this process. As well, the proposed method is able to reduce greatly the number of identified models even for improperly selected Laguerre model parameters. Using $\alpha = 0.85$ and $N_g = 10$ and comparing the proposed partition with the actual temperature values, which is shown in Figure 7, demonstrates that the method after refinement is able to correctly assign most of the data range. The change centred on 3,000 s is the transient behaviour caused by system start up. Therefore, no model should be identifiable from this particular region.

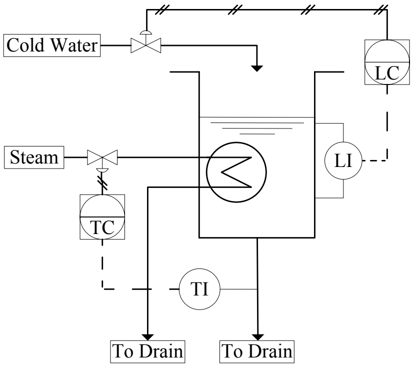

Figure 4: Process schematic

4.2 Closed-Loop Results

Once again, a 3-hour data period was extracted in which it was known that closed-loop tests were being performed. In this particular data set, manual analysis showed that there are 2 regions (at the start before 1834 s and the end after 5934 s) of manual operation and the remainder was a series of closed-loop step changes in the reference signal at different process conditions. Both the original and proposed refined method were run on the same set of underlying parameter conditions. The results are shown in Figure 8 for the original method and Figure 9 for the refined method. Figure 10 shows a comparison between the actual data and the partitions. Both figures show that the previously obtained result hold and that the refined method can reduce the number of partitions.

5. CONCLUSIONS

In this paper, an investigation of a recently proposed method for data segmentation for system identification was made to determine the effect the different parameters have on the segmentation results. It was determined that there are 4 key sets of variables to consider: (1) Laguerre model parameters, (2) forgetting factor values, (3) thresholds, and (4) excitation component of the input signal. Based on both simulation and experimental testing, it was determined that Laguerre model parameters and thresholds are the two most important variables to consider. The forgetting factors do not have much of an influence on the parameter values, while the input signal's influence was only to cause a change in the various thresholds. However, irrespective of the fine tuning done, the number of segments determined was almost always much greater than the true number of segments present. Therefore, in order to correct this problem, a segment reduction procedure was introduced based on the entropy values for adjacent segments. If the two values lay within a certain tolerance, then it was concluded that the two segments should be merged. Implementing an entropy-based step improved the performance markedly.

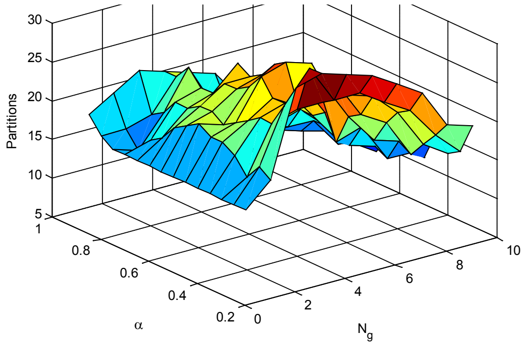

Figure 5: The number of models identified as a function of α and N_g for a pilot scale process using the original method.

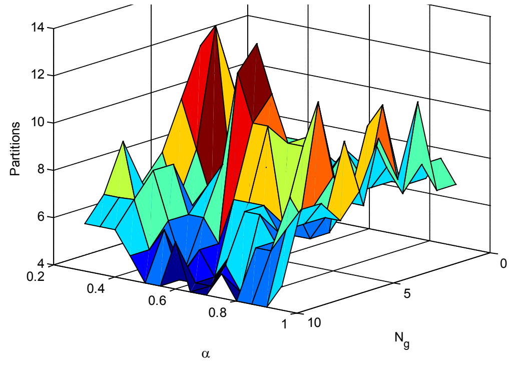

Figure 6: The number of models identified as a function of α and N_g for a pilot scale industrial process using the proposed refined method.

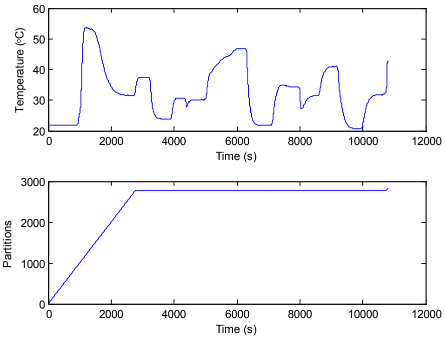

Figure 7: Comparison between the automatic partitions and the actual, open-loop data

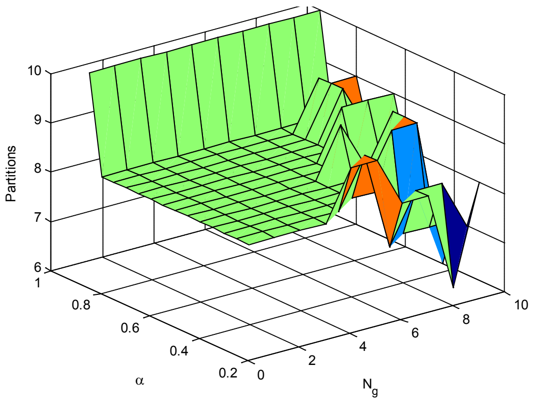

Figure 8: The number of models identified as a function of α and N_g for a pilot scale industrial process using the original method for closed-loop data

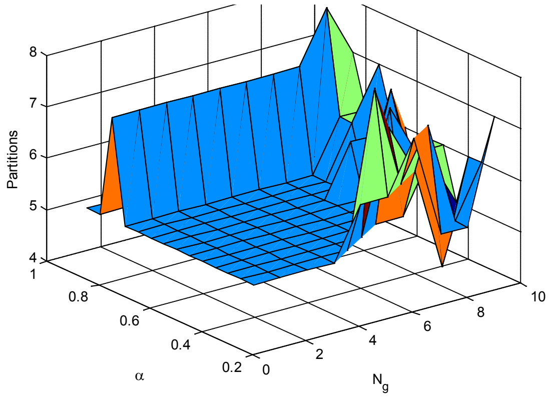

Figure 9: The number of models identified as a function of α and N_g for a pilot scale industrial process using the refined method for closed-loop data.

In the open-loop case, it was determined that the following tuning rules are recommended:

1) For most processes, set *α* between 0.8 and 0.95. For integrating processes, set *α* to be smaller than about 0.6.
2) Based on the selected *α* and the maximal time delay, determine an appropriate number of Laguerre models. It does not hurt to overestimate the number, as this is an orthogonal basis and the additional models will not affect the segmentation.
3) Set the forgetting factors to any value greater than about 0.95. The exact value is immaterial.
4) The thresholds need to be carefully set based on the input signal expected. This can involve some trial and error depending on the data available. For step tests, setting small thresholds for the variance of the input signal is very important. Values as small as 10^{-12} can be required.
5) The entropy threshold can be set to between 0.1 and 0.25 to provide the best results.

In the closed-loop case, in general, the results are the same as for the open-loop case, except that some of the thresholds may need to be set even smaller, especially with step changes in the setpoint. As well, the entropy threshold, depending on the disturbances present, may need to be set much lower than previously (around 0.005) in order to deal with "good" controllers.

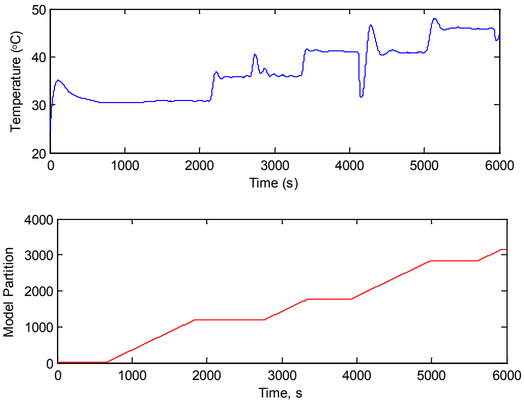

Figure 10: Comparison between the automatic partitions and the actual data for the closed-loop data

6. ACKNOWLEDGEMENT

The authors would like to thank the National Research Council of Canada for funding.

7. REFERENCES

Amirthalingam, R., Sung, S. W. & Lee, J. H., 2000. Two-step procedure for data-based modeling for inferential control applications. *AIChE Journal,* 46(10), pp. 1974-1988.

Carrette, P., Bastin, G., Genin, Y. Y. & Gevers, M., 1996. Discarding Data May Help in System Identification. *IEEE Transactions on Signal Processing,* November, 44(9), pp. 2300-2310.

Horch, A., 2000. *Condition Monitoring of Control Loops (Doctoral Thesis),* Stockholm, Sweden: KTH.

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.

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

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

Shardt, Y. A. W. & Huang, B., 2013. Statistical properties of signal entropy for use. *Journal of Chemometrics,* November, 27(11), p. 394–405.