Preprints, 10th IFAC International Symposium on
Advanced Control of Chemical Processes
Shenyang, Liaoning, China, July 25-27, 2018

Automated System Identification in Mineral Processing Industries: A Case Study using the Zinc Flotation Cell

Yuri A.W. Shardt*, Kevin Brooks†

**Technical University of Ilmenau, Ilmenau, Canada*
(e-mail: yuri.shardt@tu-ilmenau.de)
† BluESP, 53 Platina St, Randburg, South Africa, +27 11 251 5900
(e-mail: kevin.brooks@bluesp.co.za)

Abstract: In many industries, including the mineral processing industry, process modelling can be improved by mining the data historian. However, the data in the historian is often contaminated with missing values, unknown operating conditions, and other imperfections. Furthermore, manual segmentation of the data is difficult due to the large number of data points and variables. Thus, there is a need to develop and implement methods that can automatically segment the data set into viable components for identification purposes. One approach uses Laguerre models to segment the data set. However, when used in a multivariate situation, such as in the zinc flotation cell, various issues, such as collinearity, arise. Therefore, the data segmentation algorithm needs to take this into consideration when examining a data set. Using the zinc flotation cell, it is shown that for the multivariate case preselecting the data variables to consider improves the data segmentation.

Keywords: **system identification, data mining, zinc flotation cell**

1. INTRODUCTION

In process industries, when implementing control strategies, models of varying accuracy are required. This is especially the case with model predictive control (MPC). MPC has become the standard in the refining and petrochemical industries (Qin & Badgwell, 2003). Furthermore, this technology is seeing some application in the mining, metals and minerals area (Olivier & Craig, 2017).

Commercial MPC technology makes use of linear or nonlinear models that are obtained from performing planned experiments on the plant. Since this step testing is expensive from an engineering hours perspective, it has led to the development by various companies of automated stepping tools (Kalafatis, et al., 2006; Darby & Nikolaou, 2014), which require bootstrapping through the generation of a "seed" matrix. This is a response matrix for the system that expresses the key relationships, while not necessarily being extremely precise. This matrix is normally generated by performing some manual steps, which runs counter to the aim of reducing or eliminating the need for engineering supervision during testing.

Since the majority of MPCs are installed after the plant has been operational for some time, the question arises as to whether historical data, collected possibly over years of plant operation, could be used to generate these seed models. Experience in attempting this has led to the conclusion that historical data can be used, but that there are practical difficulties in doing so. These include: periods of data where the base level control system is not in the mode required to identify the models; saturation of PID loops; correlation of inputs leading to poor models; poor excitation of the inputs; and dad data, which is extremely common for analysers.

In principle, all of these issues could be addressed by having a large period of data available. The challenge then becomes investigating this data for periods that can and cannot be used for identification. For a large dataset this cannot be done manually.

Therefore, there is a need to investigate the use of algorithms that can calculate periods of data for which model identification is likely to succeed.

Recently, based on the previous work of detecting transients (Horch, 2000), data impact analysis (Carrette, et al., 1996), and segmentation for inferential controllers (Amirthalingam, et al., 2000), two approaches for determining the suitability of a given data segment for control purposes, especially identification, have been developed. The first method developed by Peretzki *et al.* (2011) uses Laguerre models as the basis for extracting the desired model conditions. The key advantage of this approach is that the process time delay is not required. However, 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. 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. Recent work has shown that, since the Laguerre approach does not require knowledge of the time delay, it can be useful in extracting data from industrial historians (Shardt & Shah, 2014; Bittencourt, et al., 2015). The application of these methods to open-loop,

multivariate processes has recently been considered (Patel, 2016). However, since many processes are already running in closed-loop operation, it is necessary to extend the results to such cases.

Therefore, this paper proposes to analyse the Laguerre-based approach for application in a multivariate industrial system to determine the challenges of using this approach for identifying data for system identification with a view of generating the seed matrices for MPC identification. A case study using the zinc flotation cell will be presented to show some of the key results.

2. DATA SEGMENTATION FOR SYSTEM IDENTIFICATION

When processing historical data with an eye on extracting regions that can be used for identification it is necessary to consider not only the theoretical foundations, but also the impact of various tuning parameters on the system.

In data segmentation, the Laguerre polynomial is often used, since it eliminates the need for knowing the process time delay. For these reasons, it makes sense to use this approach when extracting historical data for which the time delays is not accurately known.

Tuning parameters in any method primarily impact on how critically the algorithm scrutinises each of the regions to determine the suitability for identification purposes. In general, the tighter the bounds, the greater the scrutiny and the fewer regions suitable for identification will be found. On the other hand, looser bounds will allow for a greater number of suitable regions.

A final element of consideration is handling multivariate data. This involves the selection and consideration of which subset of the available parameters should be considered for identification purposes. This problem is not necessarily a trivial one and it could easily require substantial considerations.

2.1 Data Segmentation Algorithm

The general data segmentation algorithm can be described as (Peretzki, et al., 2011):

1) **Preprocessing**: Load and preprocess the data set. Most often, this will involve scaling and centring the data set.
2) **Mode Changes**: In order to simplify the detection of suitable regions, it is important to separate the data set into the different modes that are present. Modes can be defined as changes in operating points, faults, controller settings, or other similar known changes. Removing the known changes will improve the ability of the algorithm to detect the changes.
3) **Segmentation**: For each mode, perform the following steps:
   a. **Initialisation**: Set the mode counter to the current data point, $k_{init} = k$.
   b. **Computation**: Compute the required values for the given algorithm. In most cases, this will include the variances of the signals and the condition number of the information matrix.
   c. 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.b.
      ii. **Success**: Otherwise, set $k = k + 1$, and go to Step 3.c. The "good" data region is then $[k_{init}, k]$.
4) **Termination**: The procedure stops once k equals N, the total number of data points in the given operating region.
5) **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).

2.2 Laguerre-Based Segmentation

The Laguerre-based data segmentation uses orthogonal Laguerre polynomials to model the system. This orthogonality allows for easy removal of unnecessary model components without affecting the rest of the parameters. The i^{th} order Laguerre model is given 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 resulting model can then be written as

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

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 parameters for the model given by Equation (2) can be obtained using standard regression analysis.

In this approach, a recursive method is used to compute the required variances, that is, the following update rule is used:

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

where λ is the forgetting factor and σ^2 is the variance of the given signal. It can be noted that two forgetting factors are present λ_{m_y} and λ_{σ_y}, which need to be tuned. The variance is updated using the above formulae for 3 different signals, the inputs, outputs, and the regression matrix. Based on previous experience, the forgetting factors will all be set to 0.99.

The Laguerre model parameters, α and N_g, are the other two model parameters whose value needs to be set. According to (Peretzki, 2010)

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

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. For the purposes of this investigation, α will be selected as 0.80, while the value of N_g will be set to 6, since the actual values of the time delay are not known. However, it is known that it is not greater than about 100 minutes. The sampling time is fixed to 1 minute. These constraints support the value for N_g that has been selected.

Selecting the thresholds can be a bit difficult, especially without considering some of the properties of the signals themselves. For the input signal, in order to be generous and allow for more regions to be identified, the variance threshold was set to 10^{-7}. For the output signal, the variance threshold was set to 10^{-7}. The regression variance was set to 10^{-3}. The condition number threshold was set to the standard value of 1,000 [**cite my thesis**].

2.3 Multivariate Analysis

Since most of the previous approach have only considered univariate input variables, this paper will also examine the implications in terms of the multiple inputs and their impact on finding suitable regions. Different combinations of variables will be taken to determine if it is possible to segment a given data set without necessarily using all the required variables. Clearly, the more variables that are present, the larger the matrices involved, and the greater the computational power required. Since the quality of the model is only one item to consider, it is important to consider the trade-off between speed of segmentation and the accuracy of the results.

As well, when dealing with multivariate data, it may happen that some of the parameters are irrelevant for identification. In such cases, it will be interesting to examine the impact that irrelevant variables have on the ability of the method to determine the identification regions.

3. 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 in this study has been 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 1, these banks are named the roughers, scavengers and recleaners.

The section of the circuit covered here is the zinc roughers. The rougher tails from the upstream lead circuit are the feed to the zinc roughers. As shown in Figure 2, this bank consists of four cells (FC23, FC24, FC25, FC26). Their aim is to do 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.

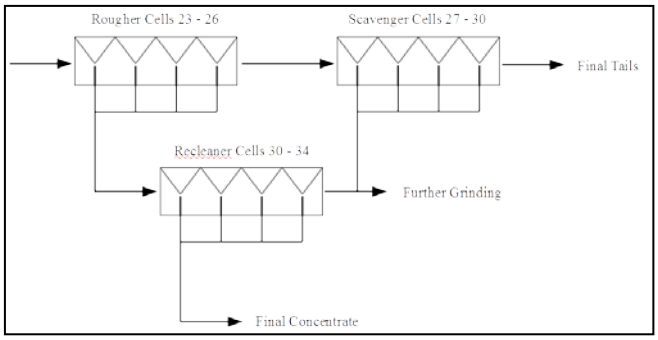

Figure 1: Zinc rougher, scavenger and recleaner circuit.

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 2, 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.

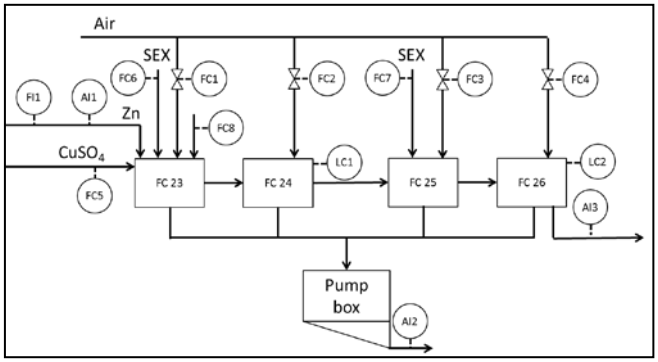

Figure 2: Rougher Bank Showing Control Loops and Analysers

4. TEST DATA

The data collected for this investigation consists of thirty-one days of plant operation. These were collected from the plant historian 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. There is a period of one day in the data where the feed falls away.

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.

Table 1: Test Variables

Tag	Attributes	Description
FI1	PV	Feed rate
AI1	Fe/Pb/Zn	Feed Compositions
FC5	SV/PV/MV	CuSO4 (reagent) to FC22
FC6	SV/PV/MV	EX (reagent) to FC23
FC7	SV/PV/MV	EX (reagent) to FC25
FC8	SV/PV/MV	NS (reagent) to FC3
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
LC1	SV/PV/MV	FC24 Level
LC2	SV/PV/MV	FC26 Level
AI2	Fe/Pb/Zn	Primary Rougher Concentrate Compositions
AI3	Fe/Pb/Zn	Primary Rougher Tailings Compositions

A design for a MPC on this unit has been derived. The manipulated variables (MVs) are the air flows, levels and the flows of the reagents. Feed-forward (FF) variables are expected to be the feed flow and feed composition or compositions. The outputs or controlled variables (CVs) are the zinc percentages in the concentrate and tailing streams.

5. RESULTS AND DISCUSION

Based on the analysis of the data set, five different cases will be considered:

1) **Case 1**: All data will be used for segmentation of the data set.
2) **Case 2**: Using three variables to segment the data set. The selected variables are LC1, LC2, and AI1Pb.
3) **Case 3**: Using three variables to segment the data set. The selected variables are FC1, FC2, and FC3.
4) **Case 4**: Using two variables to segment the data set. The selected variables are FC1 and FC3.
5) **Case 5**: Using expert knowledge to select the variables based on what variables should impact the model. The selected variables are FI1, FC5, FC6, FC7, FC8, LC1, and LC2.

For each case, the data set was segmented using the programme and a model using the "good data set" was developed using Aspentech® DMC Model to derive linear step response models. For the purposes of this study, only subspace methods were used. The variables were not conditioned before modelling. As well, a constant settling time of 90 minutes was selected for all the models. Furthermore, it can be noted that during data segmentation, whenever the output sensor failed, it was assumed that the mode had changed and that component was separated out of the model.

For Case 1, where all the available measurements were used, it was quickly determined that no useful information could be extracted, since some of the variables are correlated with each other, leading to strongly ill-conditioned matrices. This suggests that it is important to properly select the appropriate variables to consider.

For Case 2, the segmentation results are shown in Figure 3. It should be noted that a constant segment number represents a region where the data is assumed to belong to the same model. A segment value of −1 corresponds to those regions where the sensor failed. The segment number increases every time a data point fails to be good for identification. After every new segment, there will be a short transient region corresponding to the time it takes to have sufficient data for identification (40 data points are considered the minimum for identification).

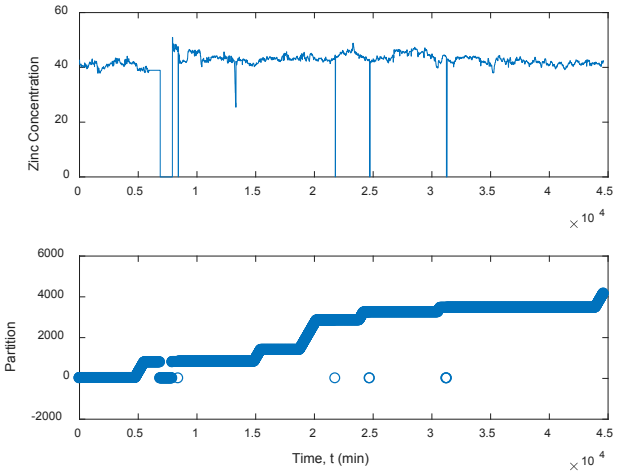

Figure 3: Segmentation Results for Case 2

For Case 3, the segmentation results are shown in Figure 4. The same definitions have been used as for Case 2. It can be seen that the number of segments is quite different even though 3 variables have been used.

For Case 4, the segmentation results are shown in Figure 5. The same definitions have been used as for Case 2. Here it can be seen that decreasing the number of variables has lead to an increase in the regions that are not sufficiently good for identification. However, this could easily be a function of the variables selected. Nevertheless, selecting an appropriate subset of 2 variables could be difficult as it would involve a large search.

Finally, for Case 5, the segmentation results are shown in Figure 6. Here it can be seen that there are large areas of

constant value located between the sensor faults. It would be possible to determine if the adjacent segments are actually similar and warrant being combined. Doing this would provide additional data for model building.

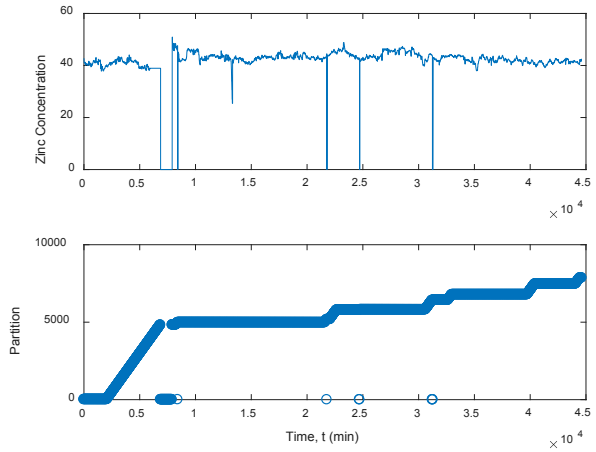

Figure 4: Segmentation Results for Case 3

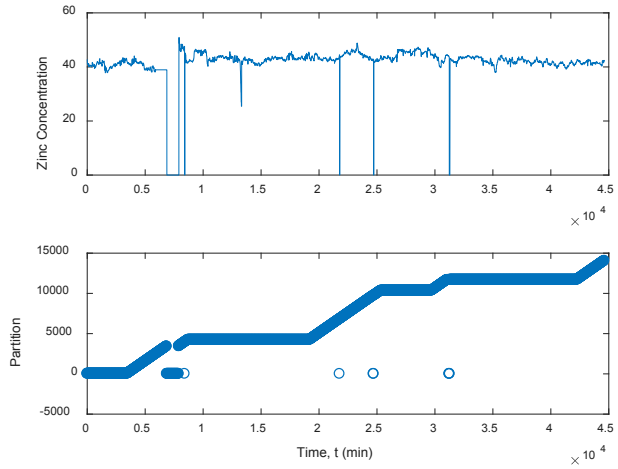

Figure 5: Segmentation Results for Case 4

The resulting models for all the cases are shown in Figure 7. The step responses for each of the variables of interest and the resulting models have been provided for both inputs. It can be noted that in practice the air flow rates are combined into a single variable. The same is done for the EX reagent. In general, it can be seen that the quality of the resulting model strongly depends on the segmentation results. It can be seen that Cases 2, 4, and 5 present similar results, while Case 3 (denoted by the black line) often gives models that deviate strongly from the consensus. Noting that the purpose of this modelling exercise is to develop "seed model" for use as the initial values for the MPC model creation software, it should be noted that the overall accuracy of the model is not all that important, except that it provide the correct overall picture.

Table 2 shows the root mean square error and R^2 for the fit of the zinc concentration models for the first output. It can be seen that in general the fit for all the cases is relatively low. However, of the considered cases, Case 4 has the best fit. This suggestions that the segmentation method can accurately determine which regions should be used for modelling and which ones should not. Furthermore, since the data was extracted from a data historian without any prior data conditioning, there is no guarantee that the data set itself can provide decent models.

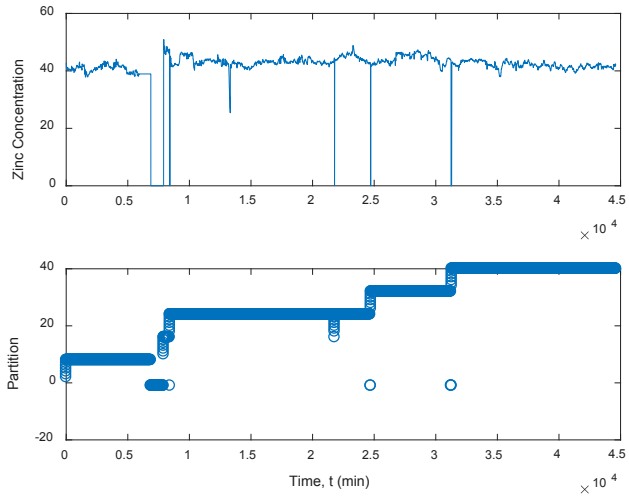

Figure 6: Segmentation Results for Case 5

Table 2: Summary Statistics for AI2.ZN

	Case 2	Case 3	Case 4	Case 5
RMSE	1.53	1.95	1.54	2.07
R^2	0.18	0.13	0.38	0.13

6. CONCLUSIONS

This paper examined the application of a data segmentation algorithm to the zinc flotation cell. In this case, the Laguerre approach to data segmentation was used, since it did not require knowledge of the time delays. Furthermore, since multiple inputs were available, different sets of variables were tested in order to determine which if the variables could be used for quickly segmenting the data set. The larger the number of variables, the longer it takes to properly segment the data set. As well, variables which do not have an influence on the model should be removed when data segmentation is performed.

The above observations were validated using data extracted from a historian for a zinc flotation cell. The best segmentation, both in terms of the number of segments and their accuracy, was using all the relevant variables. Furthermore, the resulting models were sufficiently accurate to be used for the initial seed for model predictive controllers.

Therefore, when dealing with multiple inputs, it is important in selecting the appropriate set of variables to consider for segmentation purposes. Too large and too small of a number can have an impact on the final quality of the models.

Future work will focus on determining if a subset of variables can be used to obtain better segmentation results.

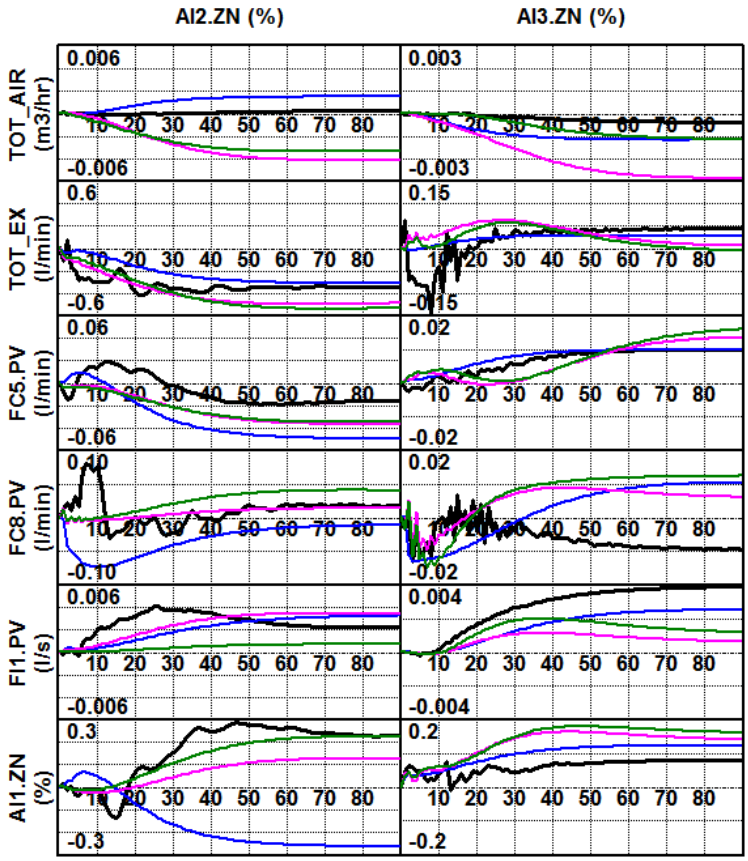

Figure 7: Unit Step Response Models (Case 2: black, Case 3: blue, Case 4: pink, and Case 5: green)

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.

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.

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.

Darby, M. L. & Nikolaou, M., 2014. Identification for multivariable model-based control: An industrial perspective. *Control Engineering Practice,* 22(1), pp. 165-180.

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

Kalafatis, A. et al., 2006. Multivariate step testing for MPC projects reduce crude unit testing time. *Hydrocarbon Processing,* pp. 93-400.

Olivier, L. E. & Craig, I. K., 2017. Should I Shut down my Processing Plant? − An Analysis in the Presence of Faults. *Journal of Process Control,* Volume 56, pp. 35-47.

Patel, A., 2016. *Data Mining of Process Data in Mutlivariable Systems,* Stockholm, Sweden: Royal Institute of Technology.

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.

Qin, S. J. & Badgwell, T. A., 2003. A survey of industrial model predictive control technology. *Control Engineering Practice,* 11(7), pp. 733-764.

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.

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