Selecting Transients Automatically for the Identification of Models for an Oil Well

Antônio H. Ribeiro & **Luis A. Aguirre**

*Department of Electronic Engineering, Universidade Federal de Minas Gerais (UFMG). Av. Antônio Carlos 6627, Belo Horizonte, MG, Brazil (*antonio-ribeiro@ufmg.com.br, aguirre@cpdee.ufmg.br*)*

Abstract: This paper proposes a procedure to automatically select transient windows for system identification from routine operation data. To this end two metrics are proposed. One quantifies the transient content in a given window and the other provides an overall measure of correlation between such transients and the chosen model input. The procedure is illustrated using data from an oil well that operates in deep waters.

Keywords: Automatic transient selection, system identification, soft sensors, intelligent oil fields

1. INTRODUCTION

The set of techniques for constructing models from observed data is known, in the control theory context, as *System Identification.*

To construct models from a data set it is necessary to have a set of data that contains relevant information about the system. Some times only "routine operation" data is available and, other times, it will be necessary to create tests for extract dynamic information about the system.

There are tests that excite a wide range of system frequencies (e.g. pseudorandom binary signal (PRBS)) and, therefore, are proper for linear system identification since the output data obtained will contain significant information about the system dynamics.

For nonlinear models there is a need to drive the system over a wider range of amplitudes and PRBS may not be the best choice as shown by Leontaritis and Billings (1987). Classical system identification textbooks offer some practical guidance in what concerns testing (Ljung, 1987).

Sometimes it will not be possible to perform experiments on the system and historical data is used. Since the data is recorded during "routine operation", the system will probably be in steady state most of the time and there will be few data windows containing relevant dynamical information.

Therefore in practical problems of system identification from routine operation data the choice of informative windows of data are both subjective and greatly time consuming. Hence this paper puts forward a criterion to aid in the choice of informative windows of data from a large data set. The method was developed for an oil well (Teixeira et al., 2014), but should also be useful in other applications.

In this paper, two metrics are proposed to classify windows from the oil well recorded data with the identification in view. Each metric addresses one of the following points:

(1) The transient is appropriate for identification only if it contains relevant information about the system dynamics;
(2) The output should be well correlated with the input, otherwise the transient is caused by an unmeasured disturbance and the window are not appropriate for identification of an input-output model.

Using such metrics it is possible to create automatic routines that choose good transients for identification.

The remainder of the paper is organized as follows. In Section 2 some mathematical concepts are quickly revisited. These concepts will be used on Sections 3 and 4 to define the metrics. In Section 5 the use of the metrics is illustrated considering a real numerical problem, and finally, some concluding remarks are provided in Section 6

2. MATHEMATICAL BACKGROUND

2.1 Singular Value Decomposition (SVD)

Theorem 1. Any $m \times n$ matrix $\mathbf{A}$ with rank r can be factored as (Strang, 1988):

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T,$$

where $\mathbf{U}$ $(m \times m)$ and $\mathbf{V}$ $(n \times n)$ are both ortogonal[1]. And Σ is a $m \times n$ matrix containing r elements on its main diagonal:

$$\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r).$$

The scalars $\{\sigma_1, \sigma_2, \dots, \sigma_r\}$ are the singular values of $\mathbf{A}$ and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$.

Proof. This is shown in (Strang, 1988, pp.450-451).

The key ideia behind the SVD factorization is to take into consideration the diagonalization of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$:

Corollary 2. The columns of $\mathbf{U}$ are the eigenvectors of the matrix $\mathbf{A}\mathbf{A}^T$ and the columns of $\mathbf{V}$ are the eigenvectors of

[1] The matrix U is ortogonal if and only if $UU^T = I$

the matrix $\mathbf{A}^T\mathbf{A}$. The singular values $\{\sigma_1, \sigma_2, \ldots, \sigma_r\}$ are the square roots of the nonzero eigenvalues of both $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$.

This can be easily understood considering:

$$\mathbf{A}\mathbf{A}^T = (\mathbf{U}\Sigma\mathbf{V}^T)(\mathbf{V}\Sigma^T\mathbf{U}^T) = \mathbf{U}\Sigma\Sigma^T\mathbf{U}^T,$$

and interpreting $\mathbf{U}\Sigma\Sigma^T\mathbf{U}^T$ as the diagonalization of the symmetric matrix $\mathbf{A}\mathbf{A}^T$. The diagonal matrix containing the eigenvalues of $\mathbf{A}\mathbf{A}^T$ is:

$$\Sigma\Sigma^T = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2),$$

and the columns of $\mathbf{U}$ are its eigenvectors.

Analogously, the columns of $\mathbf{V}$ are eigenvectors of $\mathbf{A}^T\mathbf{A}$ and its eigenvalues are the squared singular values.

2.2 Cross Correlation Function

Given two stationary signals $y(k)$ and $u(k)$, the cross correlation function measures the similarity between u and copies of y shifted (lagged) by τ. It is defined as the expected value ($\mathrm{E}\{\cdot\}$) of u times a shifted copy of y:

$$r_{uy}(\tau) = \mathrm{E}\{u(t)y(t+\tau)\}, \tag{1}$$

The cross correlation function can be estimated for a finite time series with m samples by:

$$\hat{r}_{uy}(\tau) = \frac{1}{m}\sum_{k=1}^{m} u(k)y(k+\tau). \tag{2}$$

An important practical consideration is how large the cross correlation should be so it indicate statistically significant correlation between two signals. And, in fact, for two uncorrelated signals $u(t)$ and $y(t)$ (where the expected cross correlation is zero) there is a 95% probability that the *estimated* normalized cross correlation falls within the *confidence interval*:

$$-\frac{1.96}{\sqrt{m}} \leq \rho_{uy}(\tau) \leq \frac{1.96}{\sqrt{m}}, \tag{3}$$

where $\rho_{uy} = \hat{r}_{uy}/\sigma_u\sigma_y$ is the normalized cross correlation (σ stands for standard deviation).

Hence if the cross correlation is outside the confidence interval at some lag (τ) it is fair to say that the two signals have a high probability to be correlated at lag τ.

The correlation between a stationary signal $y(t)$ and a shifted version of itself is known as autocorrelation and is denoted by:

$$r_{yy}(\tau) = \mathrm{E}\{y(t)y(t+\tau)\}. \tag{4}$$

All previous considerations are equally valid for the autocorrelation function.

3. DYNAMIC BASED METRICS

More informative signals are better suited for identification and will yield better parameter estimation. The next section will discuss how to adjust an autoregressive model to the signal y and a clear way to evaluate this signal activity will arise as consequence.

3.1 Autoregressive (AR) Models and Regressor Matrix

A linear *autoregressive* (AR) model is defined as

$$y(k) = a_1y(k-1)+a_2y(k-2)+\cdots+a_ny(k-n)+e(k), \tag{5}$$

where $e(k)$ is white noise and the scalar parameters $\{a_1, a_2, \ldots, a_n\}$ may be estimated from recorded data. If the signal y is known from the instant $k = 1$ to the instant $k = m$, then

$$\begin{aligned}
y\,(1) &= a_1y(0) + a_2y(-1) + \cdots + a_ny(-n+1) + e(1)\\
y\,(2) &= a_1y(1) + a_2y(0) + \cdots + a_ny(-n+2) + e(2)\\
\vdots\quad & \quad\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots\\
y\,(m) &= a_1y(m-1) + a_2y(m-2) + \cdots + a_ny(m-n) + e(m),
\end{aligned} \tag{6}$$

which can be rewritten in the matrix form as:

$$\mathbf{A}\mathbf{x} + \mathbf{e} = \mathbf{y}, \tag{7}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ is called the *AR regressor matrix.* Such matrix and vectors have the following structure:

$$\mathbf{A} = \begin{bmatrix} y(0) & y(-1) & \ldots & y(-n+1)\\ y(1) & y(0) & \ldots & y(-n+2)\\ \vdots & \vdots & \ddots & \vdots\\ y(m-1) & y(m-2) & \ldots & y(m-n)\end{bmatrix};$$

$$\mathbf{y} = \begin{bmatrix} y(1)\\ y(2)\\ \vdots\\ y(m)\end{bmatrix}; \quad \mathbf{e} = \begin{bmatrix} e(1)\\ e(2)\\ \vdots\\ e(m)\end{bmatrix}; \quad \mathbf{x} = \begin{bmatrix} a_1\\ a_2\\ \vdots\\ a_n\end{bmatrix}.$$

The AR regressor matrix may be written as:

$$\mathbf{A} = [\mathbf{A}_1 \quad \mathbf{A}_2 \quad \ldots \quad \mathbf{A}_n].$$

where $\mathbf{A}_i \in \mathbb{R}^m$ is the i-th column of matrix $\mathbf{A}$. Equation 7 may be rewritten as:

$$[\mathbf{A}_1 \quad \mathbf{A}_2 \quad \ldots \quad \mathbf{A}_n]\begin{bmatrix} a_1\\ a_2\\ \vdots\\ a_n\end{bmatrix} + \mathbf{e} = \mathbf{y},$$

or, equivalently as $a_1\mathbf{A}_1 + a_2\mathbf{A}_2 + \cdots + a_n\mathbf{A}_n + \mathbf{e} = \mathbf{y}$.

The *range* of matrix a $\mathbf{A}$ is defined as the vector space that contais all possible results of $\mathbf{A}\mathbf{x}$ or, in other words, all vectors that may be written as linear combinations of the columns of $\mathbf{A}$: $a_1\mathbf{A}_1 + a_2\mathbf{A}_2 + \cdots + a_n\mathbf{A}_n$.

Because of the noise vector $\mathbf{e}$, the output vector $\mathbf{y}$ does not belong to the *range* space of the matrix $\mathbf{A}$ and there is no general exact solution to (7). Its is generally possible to find a solution $\hat{\mathbf{x}}$ in the *least square* sense, which corresponds to the orthogonal projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto the *range* space of $\mathbf{A}$. In this case the solution $\hat{\mathbf{x}}$ is such that:

$$\mathbf{A}\hat{\mathbf{x}} = \hat{\mathbf{y}},$$

or, equivalently $\hat{a}_1\mathbf{A}_1 + \hat{a}_2\mathbf{A}_2 + \cdots + \hat{a}_n\mathbf{A}_n = \hat{\mathbf{y}}$.

The next theorem states the importance of the rank r of the matrix $\mathbf{A}$ on the parameter estimates.

Theorem 3. Given the equation

$$\mathbf{A}\hat{\mathbf{x}} = \hat{\mathbf{y}}, \tag{8}$$

with rank($\mathbf{A}$) $= r$, $\hat{\mathbf{x}} \in \mathbb{R}^n$ and $\hat{\mathbf{y}} \in \mathbb{R}^m$ belongs to the *range* of $\mathbf{A}$. The solution $\hat{\mathbf{x}}$ belongs to a subspace of dimension $n - r$.

Proof. Matrix $\mathbf{A}$ has r linear independent rows, hence Eq. 8 imposes r linearly independent constraints to the solution $\hat{\mathbf{x}}$. The vector $\hat{\mathbf{x}}$ belongs to the $\mathbb{R}^n$ of dimension n. Each constraint narrows by one the solution subspace, so the solution $\hat{\mathbf{x}}$ belongs to a subspace of dimension $n-r$.

Theorem 3 makes it easy to see that if $n > r$ (there are redundant regressors in the model), then infinitely many solutions are possible to Eq. 8. Hence in order to have a unique solution, the number of parameters should be equal to the rank of the matrix $\mathbf{A}$. Hence *the rank of a given AR regressor matrix is an upper bound to the number of parameters that can be estimated for such regressors.*

Thus the rank of the AR regressor matrix $\mathbf{A}$ is deeply related with how much information one can extract from a signal and, therefore, is a good metric to evaluate if a window is suitable for identification.

3.2 Covariance Matrix

The covariance matrix of $\mathbf{y} \in \mathbb{R}^m$ is defined as:

$$C_y = \mathrm{E}\{(\mathbf{y} - \mu_y)(\mathbf{y} - \mu_y)^T\} \in \mathbb{R}^{m \times m}, \tag{9}$$

where μ_y is the mean value of $\mathbf{y}$, and for the purpose of linear system identification it is zero. Under this assumption C_y can be estimated as

$$\hat{C}_y = \lim_{m \to \infty} \frac{1}{m} \mathbf{A}^T \mathbf{A}. \tag{10}$$

Theorem 4. The AR regressor matrix $\mathbf{A}$ (rank r) and the estimated covariance matrix $\hat{C}_y$ have the same rank.

Proof. From Corollary 2 the eigenvalues of $\hat{C}_y$ equals the singular values of $\mathbf{A}$ squared. From Theorem 1, the matrix $\mathbf{A}$ have r non-zero singular values. Thus, $\hat{C}_y$ have r non-zero eigenvalues. Since the number of non-zero eigenvalues is equal the rank of a matrix, so rank($\hat{C}_y$) $=$ rank($\mathbf{A}$) $= r$.

In this paper the rank of the *AR regressor matrix* will be used to evaluate the suitability of a window for system identification. However, accordingly to Theorem 4, the rank of the *covariance matrix* could also be used.

For a stationary *input* signal, the rank of $\hat{C}_u$ is known as *persistence of excitation* (Ljung, 1987). There are two main diferences between the concept of persistence of excitation and the concept of AR regressor matrix rank (proposed in this paper):

- *Persistence of excitation* is usually applied to the *input* of a system. It quantifies which tests will result in a good set of data for system identification. On the other hand, the rank of the *AR regressor matrix* is applied to the system *output*, usually obtained from routine operation.
- *Persistence of excitation* is defined for *stationary* signals. The purpose of the rank of the *AR regressor matrix* is the complete opposite of it, evaluate *transients* from recorded data.

3.3 Effective rank

Two ways to estimate the effective rank will be presented:

(1) From Theorem 1 the number of nonzero singular values is the rank r of the corresponding matrix. Hence, one measure of the *effective rank* r_1^{ef} of a matrix is the number of singular values that have normalized values σ_i/σ_1, greater than a minimum value l_1:

$$\frac{\sigma_i}{\sigma_1} \geq l_1.$$

(2) Another measure of *effective rank* r_2^{ef} can be obtained counting for how many singular values $\sigma_i - \sigma_{i-1}$ is greater them a minimum value l_2:

$$r_2^{\mathrm{ef}} = \sum_{i=2}^{n} \mathrm{H}[(\sigma_{i-1} - \sigma_i) - l_2], \tag{11}$$

where H is the Heaviside (step) function which returns 1 if the argument is non-negative and returns 0 if it is negative.

The first alternative can always be applied and is a common form to calculate the effective matrix rank. However, in the context of evaluate the signal activity in a transient, the second approach have produced more coherent results in numerical experiments and will be used from now on.

4. CORRELATION-BASED METRICS

The analysis of the output signal activity is very important to see if a window is suitable for identification. However the output activity may be caused by an unmeasured disturbance, and any attempt to use such a window of data to identify a model that explain the output as a function of the input will be ill-fated, since the output is not caused by the measured inputs.

Therefore a way to evaluate if the output and the input have relation between them is needed. The cross-correlation function (CCF), explained in Section 2, arise as a good alternative to solve this problem and will be the main topic of this section.

4.1 A scalar metric based on the CCF

The cross-correlation function (CCF) of a signal is a sequence of values. In fact, it is a function of the lag. This gives specific information as to how correlated are two signals at each lag. However, if one desires an overall picture of the correlation of the signals, the use of CCF is hard to automatize.

The test proposed by Ljung and Box (1978) summarizes the auto-correlation function in only one number. A similar attempt will be made here for the CCF. In order to get an overall assessment of the level of correlation of two signals, the following scalar metric is proposed:

$$s = \sum_{\tau=-\tau_{\max}}^{\tau_{\max}} g(\rho(\tau),\, \tau,\, p), \tag{12}$$

where $\tau_{\max}$ is the maximum lag of interest and $g(\rho(\tau), \tau, p)$ is defined as:

$$g(\rho(\tau), \tau, l) = \begin{cases} 0, & \text{if} \quad |\rho(\tau)| \leq p, \\ \dfrac{|\rho(\tau)| - p}{|\tau|}, & \text{if} \quad |\rho(\tau)| > p \quad \text{and} \quad \tau \neq 0, \\ |\rho(\tau)| - p, & \text{if} \quad |\rho(\tau)| > p \quad \text{and} \quad \tau = 0, \end{cases}$$

$\rho(\tau)$ is the normalized CCF and $p = 1.96/\sqrt{m}$, m is the window length in samples. The 95% confidence interval is given by $\pm p$.

When a single number is used instead of the CCF a lot of information will be lost. However, using (12) it is easy to establish "overall correlation" between the variables. Of course this is an oversimplified approach, however it is an objetive way to evaluate correlation between input and output and a computer can do it. This approach will be used in this paper.

5. NUMERICAL RESULTS

In this section, data recorded from an oil well will be analyzed in search of suitable windows for identification. The set of data recorded was sampled at one sample per minute. Data from several oil wells recorded during the last years are available.

Using only routine operation data, the final goal of the global project is to develop soft sensors for the downhole pressure. This is achieved by estimating models using system identification and Kalman filtering methods, that infer the desired pressure based on seabed and platform measurements. This article deals with a subproblem that arises from it: how to choose which are the windows of data that are most relevant for system identification. The procedure must be simple yet effective in order to be implemented on computer and automatically analyze large amounts of recorded data.

The metrics r_2^{ef} (11) and s (12) will be used to implement this task on a computer. Before, some specificities of the oil well problem should be considered:

- The pressure measure from the permanent downhole gauge (PDG) sensor is the variable of interest (model output), and seabed and platform measures will be used as input for the model;
- In the case of severe slugging, before computing the aforementioned metrics, the signals will be low-pass filtered.
- During some valve maneuvers, the system dynamics do not correspond to normal operation dynamics, for which the soft sensors are being developed. The corresponding windows of data should be discarded by the algorithm.

A more complete review of the process and the variables involved can be found in (Teixeira et al., 2014)

Using computer routines a large set of data can be scanned looking for transients with highest r_2^{ef} and that satisfy $s > 3$. The procedure is based on a sliding window and is described on Algorithm 1. It is important to highlight that r_2^{ef} will be preferred because the signal is typically oversampled.

Algorithm 1 *Scanning a data set looking for windows for system identification*

```
 1: $\mathbf{y} = [y(1), y(2), \ldots, y(N)]^T$ % recorded data: $\boldsymbol{y} \in \mathbb{R}^N$.
 2: Define the window increment $c$
 3: $N_c = \lfloor N/c \rfloor$ % Number of windows
 4: for $i = 1 : N_c$ do
 5:     $\mathbf{y}\{i\} = \mathbf{y}(ic : ic + m)$ % ith window
 6:     Build the AR regressor matrix $\mathbf{A}_i \in \mathbb{R}^{m \times n}$
 7:     Calculate $r_{2i}^{\text{ef}}$ (for an appropriate $l_2$) for each $\mathbf{A}_i$
 8: end for
 9: Sort the effective ranks $r_{2i}^{\text{ef}}$ and obtain a list $q$ with the
    index of windows in descending order of effective rank;
10: % Note that if $m > c$, consecutive windows $\boldsymbol{y}_i$ and $\boldsymbol{y}_{i+1}$
    will have elements in common. When this happens we
    will say two windows overlap
11: for $i = 1 : N_c$ do
12:     if $\mathbf{y}\{q(i)\}$ overlaps with $\mathbf{y}\{q(1 : i - 1)\}$ then
13:         Remove $q(i)$ from the list
14:     end if
15: end for
16: for $i = 1 : \text{size}(q)$ do
17:     Calculate the index $s_i$ for $\mathbf{y}\{q(i)\}$
18:     if $s_i < 3$ then
19:         Remove $q(i)$ from the list $q$
20:     end if
21: end for
22: Pick the first 3 windows from list $q$ and choose the one
    that suits you better for identification
```

Figure 1 shows a month of recorded data from the downhole pressure ($N = 43201$). Running the first fifteen lines from Algorithm 1 a list of non-overlapping window sort by the effective rank is obtained. The Figure 1 is colored accordingly to this list: Red indicates transients with high r_2^{ef} (higher positions on the list) and colors close to yellow correspond to lower values of that metric (lower positions on the list).The gaps between the windows have been plotted using black lines and the transients have been numbered from 1 to 11 to future reference.

The parameters used for this case is $c = 50$, $m = 3000$, $l_2 = 0.1$ and $n = 100$. The choice of these parameters are not so critical, however the user should be attempt to the following guidelines when choosing it:

- The number of columns of the regressor matriz n is related with the time of computation. High value of n will lead to more time of computation. We recommend to use $n = m/30$;
- The maximum effective rank is equals to $min(n, m)$. Since $n < m$ the maximum effective rank is n. The effective rank is not useful to discriminate windows with full effective rank. The situation when lots of windows have the full effective rank can be avoided setting the parameter l_2 with a higher value.

From previous studies it is known that a good variable to be used as an input to the model is a pressure measure from the platform. The metric s (12) that is an overall measure of the correlation between this variable and the PDG downhole pressure is shown in Table 2 for each of the numbered transients.

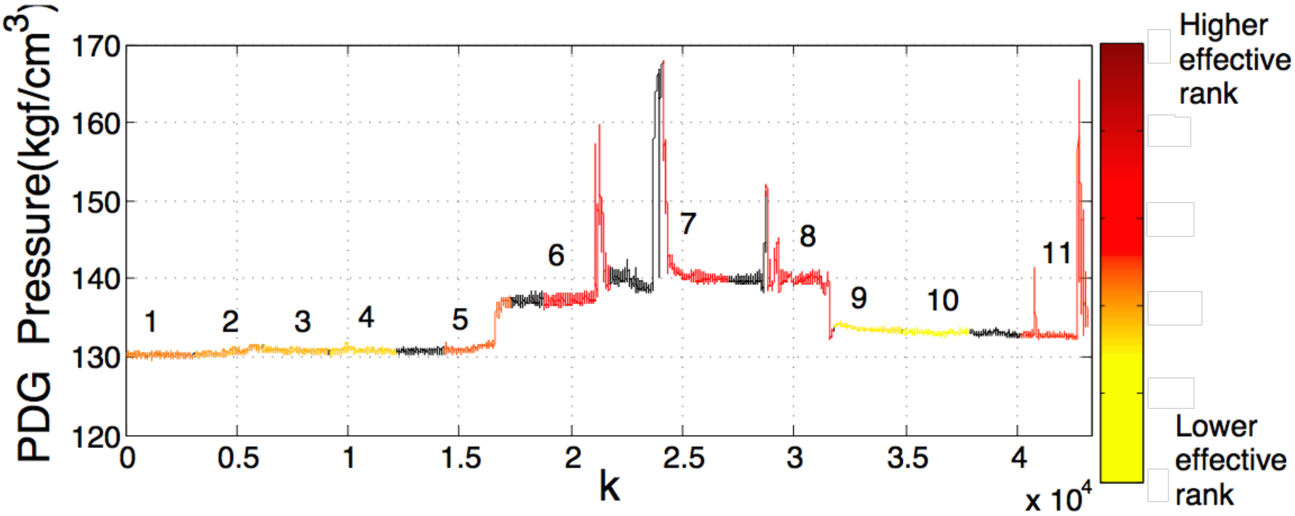

Fig. 1. PDG pressure colored according to r_2^{ef}.

Table 2: Correlation index s for each of the numbered transients

transients	s	transients	s
1	1.26	6	1.90
2	**3.47**	7	0.09
3	1.27	8	**4.95**
4	1.36	9	1.18
5	**7.59**	10	0.57
		11	**6.40**

Accordingly to Algorithm 1 transients that have the index s greater than 3 will be considered possible windows for identification, the rest of transients will be discarded from the list. So only transients 2, 5, 8 and 11 (indicated in boldface in Table 2) are possible choices for identification. The three best classified windows are: 5, 8 and 11. The Algorithm selection of windows match with the authors experience.

It is interesting to notice that although transients 6 and 7 have high values of r_2^{ef}, they are not correlated to the potential input and, therefore, should not be used. This is automatically detected by the procedure.

The computation of this example required 3.6 seconds on a MacBook Air with processor Intel Core i5, CPU 1.3 GHz.

6. CONCLUDING REMARKS

The main steps in system identification are: (i) experiment design and data collection; (ii) choice of the mathematical representation; (iii) choice of model structure; (iv) parameter estimation; and (v) model validation.

This paper aims at providing tools that will facilitate the first step of the system identification procedure (described above). In order to do so, a procedure has been devised to test if a data window is suitable for parameter estimation. One advantage of the proposed procedure is the possibility to create automatic routines capable of finding, within a possibly very long data set, transients that are adequate for system identification.

Although the final models will be ARX or NARX (polynomial or neural), the paper proposed to build an information matrix composed of AR regressors only. It is argued in this paper that the rank of such a matrix can be interpreted as an indicator of "signal activity", which is considered to be one of two important features that a data window must have to be used in identification. The second characteristic is that input and output must be correlated withing the data window.

It is important to notice that the fact that only AR regressors are used to detect signal activity in no way imposes restrictions on the model class used at a later step in the identification. As a matter of fact polynomial NARX models have been estimated from data windows selected by the procedure put forward in this paper.

The user may desire to perform adjustments to the windows automatically selected by the proposed method, but still the total time required to choose windows for system identification is greatly reduced. The metrics proposed in this article can be readily generalized to the multivariable case.

ACKNOWLEDGEMENTS

This work was partially supported by Petrobras and CNPq.

REFERENCES

Leontaritis, I.J. and Billings, S.A. (1987). Experimental design and identifiability for non-linear systems. *International Journal of Systems Science*, 18(1), 189–202.

Ljung, G.M. and Box, G.E.P. (1978). On a measure of a lack of fit in time series models. *Biometrika*, 65(2), 297–303.

Ljung, L. (1987). *System Identification: Theory for the User.* Prentice Hall, first edition.

Strang, G. (1988). *Linear Algebra and Its Applications.* Harcourt Brace Jovanovich.

Teixeira, B. O. S., Castro, W. S., Teixeira, A. F., and Aguirre, L. A. (2014). Data-driven soft sensor of downhole pressure for a gas-lift oil well. *Control Engineering Practice*, 22, 34 – 43.