Development of a Data Driven Dynamic Model for a Plasma Etching Reactor

Michael Nikolaou[a)] Haiyang Zhang and Ying Peng

Department of Chemical Engineering, University of Houston, Houston, TX 77204-4792

Submitted to Journal of Vacuum Science and Technology B (2000)

Abstract

We compare several methods for identification and validation of an empirical model for a helicon plasma reactor, on the basis of experimental data over an operating region. The model relates conveniently measurable process parameters to ellipsometry signals that capture etch depth information. Therefore, the complications of ellipsometry may be bypassed, and etch-depth may be inferred from conveniently measurable quantities in real time for potential use in real-time feedback control. The proposed identification approach shows clear improvement over a previously published study on the same experimental data.

I. INTRODUCTION

As a result of consistent demands on semiconductor manufacturers to produce circuits with increased density and complexity, tight process control has become an issue of growing importance. While run-to-run control has been widely used in the past, it is becoming evident that real-time control is needed in order to realize demands on high quality. To implement real-time control, advanced sensor, actuator, and controller technologies are needed.

Advanced process control techniques for real-time control rely on a model-based approach, i.e., an accurate model that characterizes process dynamics is explicitly used by the controller. For microelectronics manufacturing processes such as plasma etching, accurate models based on first principles may be developed. Models of plasma etching which relate the response of process outputs (such as etch rate or etch uniformity) to variations in input parameters (such as radio-frequency (rf) power, flow rate, dc bias or pressure) have been developed using first principles[2], empirical methodologies[10], or a combination of the two[11]. Plasma etching models based on first principles involve continuity, momentum balance, and energy balance equations inside a high-frequency, high-intensity electric field. Realistic simulation[9] for inductively coupled plasma reactors requires solution times on the order of several hours for a single operating point, thus being unacceptably slow for real-time feedback process control. Empirical models that are based on experimental data, on the other hand, while less insightful and limited within the range over which data are available, may be a lot more easily amenable to use in real-time computations.

Feedback control requires real-time measurements. Real-time sensing technologies for plasma etching may rely on standard techniques, such as full-wafer imaging ellipsometry[3], for etch rate measurement and end-point detection. The relationship between the ellipsometry signal and film thickness is well understood[1] on the basis of first principles. However, because of high complexity and cost, *in situ* monitoring of wafer attributes in real-time in a modern semiconductor manufacturing environment is essentially not feasible. Alternatively, the state of the wafer can be inferred in real time without costly ellipsometry measurements, if a reliable empirical model is available that relates *in situ* wafer-state signatures (ellipsometry signal) to process signatures (reflected rf power, dc bias, matching network tune, flow rate, optical emission spectra (OES)).

This work is a study on using experimental data from routine production tests to develop a model that can be used in real-time control of plasma etching. Specifically, the developed model shows (a) what is the effect of manipulated process inputs, such as rf power, pressure, dc bias, and gas flowrate, on the ellipsometry signal, and (b) how OES measurements, combined with knowledge of process inputs can be used to infer the ellipsometry signal (hence the etch rate and endpoint). In that context, we develop and validate an empirical finite-impulse-response (FIR) model based on 11 experimental data sets from a Helicon plasma reactor, by virtue of quadratic-spline (QS) wavelet–basis signal compression. The developed approach shows good improvement over previously attempted approaches with the same set of data. It should be stressed that the model developed in this work is valid only for the operating range covered by the available data and should not be extrapolated beyond that.

The rest of this paper is structured as follows. First, we present an overview of the prevailing schemes for the identification of FIR models, namely, partial least-squares (PLS), principal component regression (PCR), ridge regression (RR) (and its variant RRDlin) and continuum regression (CR). Next, by virtue of the discrete wavelet transform (DWT), the development of a parsimonious FIR model for multi-input single-output (MISO) system is shown. Finally, comparison is made of the Helicon plasma system identification results for various popular approaches.

II. BACKGROUND

A three-step gate etch was performed in a inductively coupled helicon plasma reactor, as described in [10]. Chlorine was used to provide a high etch rate and anisotropic etching profile while bromine was added to favor the polysilicon-oxide selectivity. Also, the amount of oxygen fed into the gas mixture was essential to obtain anisotropic etching profiles. Data were collected for 13 input variables: eight optical emission spectroscopy (OES) frequencies and five machine parameters (dc bias, rf power, matching network tune, bromine flow rate, oxygen flow rate). In addition, data were collected[10] for the Ψ and Δ ellipsometry values which can be used to calculate the etched film thickness. A neural-network-based model connecting the 13 input variables to the two ellipsometry signals was developed in [10]. The objective of this work is to develop an improved model for improved on-line control of plasma etching.

III. FIR MODEL

For a stable multi-input-multi-output (MIMO) system, the process output variable y for the l^{th} data-set at sampling time k can be expressed in terms of an FIR model as a linear combination of inputs consisting of lagged process input variable u:

$$y_l(k) = \sum_{i=1}^{m} \sum_{j=1}^{n(i)} h_{ij} u_{il}(k-j) + e(k) = \mathbf{h}^T \mathbf{u}_l(k) + e(k) , \tag{1}$$

where the quantities $\mathbf{h}$ and $\mathbf{u}_l$ are defined as,

$$\mathbf{h} = \begin{bmatrix} h_{11} \\ h_{12} \\ \vdots \\ h_{1n(1)} \\ \vdots \\ h_{i1} \\ \vdots \\ h_{in(i)} \\ \vdots \\ h_{m1} \\ h_{m2} \\ \vdots \\ h_{mn(m)} \end{bmatrix}, \qquad \mathbf{u}_l(k) = \begin{bmatrix} \mathbf{u}_{1l}(k-1) \\ \mathbf{u}_{1l}(k-2) \\ \vdots \\ \mathbf{u}_{1l}(k-n(1)) \\ \vdots \\ \mathbf{u}_{il}(k-1) \\ \vdots \\ \mathbf{u}_{il}(k-n(i)) \\ \vdots \\ \mathbf{u}_{ml}(k-1) \\ \mathbf{u}_{ml}(k-2) \\ \vdots \\ \mathbf{u}_{ml}(k-n(m)) \end{bmatrix}^{\mathbf{T}}, \tag{2}$$

$\mathbf{h}$ is the FIR model kernel to be estimated from the input-output data; $\mathbf{e}$(k) can be either white noise or colored noise; m is the number of inputs; l refers to the lth data-set; $n(i)$ is the memory length of the i^{th} input. The value of $n(i)$ must be greater than the settling time of the process. In practice, $n(i)$ is adjusted in an iterative fashion, so that all the trailing h_{ij} are negligible for j greater than $n(i)$. If the number of total observations is p with N_l observation in the l^{th} dataset, input-output data in equation (1) can then be rewritten in canonical form as

$$\mathbf{y} = \mathbf{\Phi h} + \mathbf{e} , \tag{3}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1(k_1 - N_1 + 1) \\ y_1(k_1 - N_1 + 2) \\ \vdots \\ y_1(k_1) \\ \vdots \\ y_l(k_l - N_l + 1) \\ \vdots \\ y_l(k_l) \\ \vdots \\ y_p(k_p - N_p + 1) \\ y_p(k_p - N_p + 2) \\ \vdots \\ y_p(k_p) \end{bmatrix}, \quad \Phi = \begin{bmatrix} \mathbf{u}_1(k_1 - N_1 + 1) \\ \mathbf{u}_1(k_1 - N_1 + 2) \\ \vdots \\ \mathbf{u}_1(k_1) \\ \vdots \\ \mathbf{u}_l(k_l - N_l + 1) \\ \vdots \\ \mathbf{u}_l(k_l) \\ \vdots \\ \mathbf{u}_p(k_p - N_p + 1) \\ \mathbf{u}_p(k_p - N_p + 2) \\ \vdots \\ \mathbf{u}_p(k_p) \end{bmatrix}. \tag{4}$$

IV. LINEAR PARAMETER ESTIMATION OVERVIEW

Ordinary Least Squares

Parameter estimation via the OLS method involves minimization of the objective function

$$\min_{\mathbf{h}} (\mathbf{y} - \Phi\mathbf{h})^T (\mathbf{y} - \Phi\mathbf{h}), \tag{5}$$

yielding the estimate

$$\hat{\mathbf{h}} = \left(\Phi^T \Phi\right)^{-1} \Phi^T \mathbf{y}, \tag{6}$$

of the FIR model kernel. OLS is currently the most straightforward method for the identification of FIR model. The main goal of OLS is to seek the best input-output correlation without regard to the input block structure. Thus OLS provides the best fit of the given data (in the least square sense) but may not the best predictor of new data. If, as usual for MIMO process data, the columns of Φ exhibit a high degree of multicollinearity, then $\mathbf{\Phi^T\Phi}$ is highly ill-conditioned and some of the eigenvalues of $\mathbf{\Phi^T\Phi}$ are nearly zero. Usually such eigenvalues will lead to poor prediction of outputs when using new data sets.

Ridge Regression and its Variant RRDlin

Ridge regression purports to deal with ill-conditioning of the matrix $\mathbf{\Phi^T\Phi}$ by performing the minimization

$$\min_{\mathbf{h}} (\mathbf{y} - \Phi\mathbf{h})^T (\mathbf{y} - \Phi\mathbf{h}) + K\mathbf{h}^T\mathbf{Q}\mathbf{h}\,, \tag{7}$$

where k is a nonnegative scalar and $\mathbf{Q}$ is the constraint or penalty matrix for the estimator $\mathbf{h}$. This objective function minimizes the sum of squares of the residuals subject to a constraint on the magnitude or length of the regression estimate of $\mathbf{h}$. Thus, a ("slightly") biased estimator $\hat{\mathbf{h}}$

$$\hat{\mathbf{h}} = (\mathbf{\Phi^T\Phi} + K\mathbf{Q})^{-1}\mathbf{\Phi^T y}\,, \tag{8}$$

is obtained, whose variance is ("much") lower than that of the OLS estimator.

Various versions of regularized least squares arise from different choices of $\mathbf{Q}$. For $\mathbf{Q}$ equal to the identity matrix, we obtain the well known Ridge Regression estimator[4]. In this estimator the length of the regression vector is constrained. The size of changes in adjacent values of the parameter estimates is penalized, thereby forcing successive values to be similar and smoothing the estimated impulse response coefficients. Furthermore, if one expects that the coefficients h_{ij} follow a "smooth" pattern for large values of j, then the following positive matrix $\mathbf{Q}$, suggested by Kozub (1994)[5] can be used in eqn. (7).

$$\mathbf{Q} = \mathbf{A}^T\mathbf{L}\mathbf{A}\,, \tag{9}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{bmatrix}, \tag{10}$$

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 2 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & n-1 & 0 \\ 0 & 0 & \cdots & 0 & n \end{bmatrix}. \tag{11}$$

Although the bias in the estimate of $\mathbf{h}$ increases with increasing K, the variance usually decreases much more rapidly. The value of the ridge parameter K can be selected by cross-validation. In general, too low a value of K yields estimates with low bias but large variances, while too large a value of K yields overly smoothed and biased estimates.

Principal Component Regression

Principal component regression (PCR) is another way to handle the problem of ill-conditioning of the information matrix $\mathbf{\Phi^T\Phi}$, as follows: The matrix $\mathbf{\Phi}$ can be represented as

$$\Phi = t_1 p_1^T + t_1 p_2^T + \cdots + t_1 p_r^T + E = \mathbf{T}\,P^T + \mathbf{E}\,, \tag{12}$$

where t_i is a set of orthogonal latent vectors (scores) calculated sequentially for each dimension (i=1,2,…r) and contain information on how the samples relate to each other. p_i are the loading vectors for $\mathbf{\Phi}$ **,** they contain information on how the variables relate to each other and express the contribution of each variable in $\mathbf{\Phi}$ toward defining the new latent vectors t_r. Instead of obtaining the regression coefficients from the original input-output data, the estimators are regressed on the principal component scores of the input information as follows.

$$\hat{\mathbf{h}} = p\ (\mathbf{T^T T})^{-1}\mathbf{T^T y}\ . \tag{13}$$

The number of principal components to retain in the model can be determined by means of cross-validation. If all the principal components are retained in the model, then the result is the same as that obtained by OLS. PCR runs the risk that useful (predictive) information will end up in discarded principal components and that some noise will remain in the components used for regression.

Projection to Latent Structures

Projection to latent structures (or partial least squares, PLS) occupying some middle ground between PCR and OLS, captures the covariance between input and output blocks. This can be thought of as an attempt to balance the two tasks of providing a reduced order description of the input data block and correlating input data to output data. PLS is a multivariable regression method ideally suited to studying the variation in large numbers of highly correlated process variables. It handles this by projecting the information in the data down into a low dimensional space defined by a small number of latent vectors (t_1, t_2,…t_r). These new latent vectors summarize the information contained in the original data set.

Combining equations (3) and (12), we get

$$\mathbf{y} = \mathbf{T}\,p^T\mathbf{h} + \mathbf{F}\,. \tag{14}$$

Where $\hat{\mathbf{h}} = \widehat{\mathbf{W}}(p^{\mathbf{T}}\widehat{\mathbf{W}})^{-1}(\mathbf{T}^{\mathbf{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathbf{T}}\mathbf{y}$ and $\widehat{\mathbf{W}}$ are the weights used to maintain orthogonal scores. E and F are the residual matrices for $\mathbf{\Phi}$ and $\mathbf{y}$ blocks. Cross-validation is usually used to select the dimension r. If PLS is used on a data set of uncorrelated variables or if the number of transformed variables equals the original number of variables, then the results obtained are equivalent to OLS. One main drawback of PLS is that it performs poorly if a process has dead time.

Continuum Regression

As discussed above, each of the OLS, PLS and PCR methods strikes a particular balance between describing input block variance and achieving correlation with the output block. This balance between describing variance and capturing correlation can be changed continuously. Based on this insight, Stone and Brooks[12] unified these linear modeling methods as the CR approach by developing a common objective function that can specialize to OLS, PLS or PCR by selecting the appropriate value of an adjustable parameter. The CR adjustable parameter complements the effect of the number of basis functions on the degree of over-fitting or bias of the empirical model, and provides additional control over the quality of the empirical model. To achieve this, the CR method first does singular value decomposition on $\mathbf{\Phi}$

$$\mathbf{\Phi} = USV^T, \tag{15}$$

Where U and V are orthogonal matrices and S is a diagonal matrix of singular values. A modified $\mathbf{\Phi}$-matrix, defined as $\mathbf{\Phi}^{\mu}$, is then formed as follows:

$$\mathbf{\Phi}^{\mu} = US^{\mu}V^T. \tag{16}$$

Where the matrix S^{μ} contains the singular values raised to a specified power μ, $0 \le \mu < \infty$. We then apply the standard PLS algorithm to $\mathbf{\Phi}^{\mu}$ and $\mathbf{y}$ to form a set of regression models with a number of latent variables ranging from 1 to l. As μ decreases, $\mathbf{\Phi}^{\mu}$ becomes progressively less directional. This biases the PLS model toward the minor eigenvectors. Any rotation toward major eigenvectors results in the increase of the directionality of $\mathbf{\Phi}^{\mu}$. The rotation of the PLS latent vectors have no effect on the amount of captured $\mathbf{\Phi}$-block variance, so the CR algorithm focuses on the accuracy of the correlation.

Wavelet-compressed FIR model

The data-overfitting and high modeling uncertainty caused by the identification of FIR models from experimental data can be overcome by reparametrizing an FIR model in terms of considerably fewer coefficients. For typical FIR models, abrupt changes are usually captured by the first few components of the kernel vector **h**, while the tail-end of the kernel is usually smooth. This implies that the wavelet transform, $\tilde{\mathbf{h}}$, of **h** will have many components that are approximately equal to zero[7]. Consequently, **h** can be efficiently parametrized in terms of the remaining nonzero components of $\tilde{\mathbf{h}}$. Prior partial knowledge of **h** based on either a first-principles or an empirical model, may be used as the starting point to determine what coefficients of $\tilde{\mathbf{h}}$ are identically zero and subsequently identify **h** through regression. Computation of the wavelet transform and its inverse can be achieved efficiently using Mallat's multiresolution analysis[6] (MRA) algorithm. Nikolaou and Vuthandam (1998)[7] demonstrated that parsimony in FIR models could be obtained by compressing the underlying kernel with wavelets. The current work is an extension of their methodology to the case of a MIMO system.

The DWT of a vector is the result of a linear transformation that generates a new vector of dimension equal to that of the original vector. It is of the same dimension as the original vector **h**, and consists of the approximation at the lowest resolution level, the detail at that level, and all the details at higher levels. The characteristics of the particular wavelet used (e.g.smoothness) are also important. For given *low pass* filter C and *high pass* filter D, the filter matrices can be constructed in several ways to filter a finite-length signal, depending on the various methods of signal extension, such as periodic, symmetric or constant extension. In this work, we use the QS wavelet with constant extension. The main reasons for our choice are biorthogonality for the QS wavelet and its dual, and good interpolation properties. Filters C and D are constructed as

$$\mathbf{C}_{i,j}=\begin{bmatrix} c_0+c_1 & c_2 & c_3 & & & & \\ & c_0 & c_1 & c_2 & c_3 & & \\ & & & c_0 & c_1 & c_2 & c_3 \\ & & & & \ddots & \ddots & \ddots \\ & & & & & c_0 & c_1 & c_2+c_3 \end{bmatrix}, \tag{17}$$

and

$$\mathbf{D}_{i,j}=\begin{bmatrix} d_0+d_1 & d_2 & d_3 & & & & \\ & d_0 & d_1 & d_2 & d_3 & & \\ & & & d_0 & d_1 & d_2 & d_3 \\ & & & & \ddots & \ddots & \ddots \\ & & & & & d_0 & d_1 & d_2+d_3 \end{bmatrix} \tag{18}$$

The essentials of *low pass* and *high pass* filtering operations are to reveal the frequency content of segments of the linear kernel **h** in various frequency brands. The filter coefficients of QS wavelet in above two digital filters are defined as

$$\begin{pmatrix} c_0 & c_1 & c_2 & c_3 \end{pmatrix}=\frac{1}{4}\begin{pmatrix} -1 & 3 & 3 & -1 \end{pmatrix}, \tag{19}$$

$$\begin{pmatrix} d_0 & d_1 & d_2 & d_3 \end{pmatrix}=\frac{1}{4}\begin{pmatrix} -1 & 3 & -3 & 1 \end{pmatrix}, \tag{20}$$

The DWT, $\widetilde{\mathbf{h}}$, of the FIR linear kernel **h** for a MIMO system has the same dimension as **h** and can be obtained as follows

$$\widetilde{\mathbf{h}}=\mathbf{W}\mathbf{h}, \tag{21}$$

where

$$\mathbf{W}=\begin{bmatrix} \mathbf{W}_1 & & & & \\ & \mathbf{W}_2 & & 0 & \\ & & \ddots & & \\ & & & \mathbf{W}_i & \\ & 0 & & & \ddots \\ & & & & & \mathbf{W}_m \end{bmatrix}, \tag{22}$$

$$\mathbf{W}_i=\mathbf{W}_{1i}\mathbf{W}_{2i}\cdots\mathbf{W}_{j,i}\cdots\mathbf{W}_{(nr-1),i}\mathbf{W}_{nr,i}, \tag{23}$$

nr is the resolution level; the matrices $\mathbf{W}_{j,i}$ are of the following form:

$$\mathbf{W}_{j,i}=\left[\begin{array}{c|c} \mathbf{C}_{i,j} & 0 \\ \mathbf{D}_{i,j} & 0 \\ \hline 0 & \mathbf{I}_{n(i)-n_i} \end{array}\right], \tag{24}$$

ni is the size of matrix $\begin{bmatrix} \mathbf{C}_{i,j}^T & \mathbf{D}_{i,j}^T \end{bmatrix}$. Because of the smooth nature of **h**, it is expected that the approximation coefficients and the detail coefficients at the first few lowest resolution levels are significantly larger than all other coefficients, and the magnitude of the detail coefficients decay rapidly as the resolution level increases. So at different resolution levels there are only few significant coefficients,

with the rest being negligible. Thus, by retaining only the significant DWT coefficients, and neglecting all the others (setting them to zero) we can achieve compression of the FIR kernels, which is achieved as

$$\tilde{\mathbf{h}}_c = \mathbf{P}^T \tilde{\mathbf{h}}, \tag{25}$$

Where $\tilde{\mathbf{h}}_c$ is a vector of length n_c, equal to the number of retained coefficients after compression; and **P** is a projection matrix of dimensions $n \times n_c$ to select the elements of **h** to be retained, and consists of ones and zeros arranged suitably. The compressed FIR model is then reconstructed using the IDWT as follows

$$\mathbf{h} = \mathbf{W}^{-1}\mathbf{P}\tilde{\mathbf{h}}_c, \tag{26}$$

Plugging equation (21) into equation (1), we get

$$y_l(k) = \mathbf{h}^T \mathbf{u}_l(k) + e(k) = (\mathbf{W}\mathbf{h})^T (\mathbf{W}^{-T}\mathbf{u}_l(k)) + e(k)$$

$$= \tilde{\mathbf{h}}^T \mathbf{W}^{-T}\mathbf{u}_l(k) + e(k) \tag{27}$$

Combining equations (25) and (27) leads to

$$y_l(k) \approx \tilde{\mathbf{h}}_c^T \mathbf{P}^T \mathbf{W}^{-T}\mathbf{u}_l(k) + e(k) = \tilde{\mathbf{h}}_c^T \breve{\mathbf{u}}_l(k) + e(k), \tag{28}$$

where $\breve{\mathbf{u}}_l(k) = \mathbf{P}^T \mathbf{W}^{-T}\mathbf{u}_l(k)$. The wavelet matrix **W** for the QS wavelet is orthonormal, i.e. $\mathbf{W}^{-T} = \mathbf{W}$. The retained coefficients $\tilde{\mathbf{h}}_c$ can be identified by means of OLS to get the estimate of the compressed DWT coefficients as

$$\hat{\tilde{\mathbf{h}}}_c = \left(\breve{\Phi}^T \breve{\Phi}\right)^{-1} \breve{\Phi}^T \mathbf{y}, \tag{29}$$

where

$$\breve{\Phi} = \Phi \mathbf{W}^{-1}\mathbf{P}, \tag{30}$$

The FIR model is then estimated by reconstruction using the IDWT as follows

$$\hat{\mathbf{h}}_c = \mathbf{W}^{-1}\mathbf{P}\hat{\tilde{\mathbf{h}}}_c. \tag{31}$$

V. MAIN RESULTS AND DISCUSSION

Eleven data sets were collected from an inductively coupled helicon plasma reactor, as discussed in section II.

The ellipsometry data in the 11 data sets can be divided into three groups that have one, two, and three peaks, respectively (Figure 1). For the identified model to be as realistic as possible, we selected representative sets exhibiting one, two, and three peaks, both in identification and validation of the model.

Because the identified model is MIMO, the FIR coefficients for each input-output pair are expected to decay at different rate. Therefore, different memory length is needed for each FIR kernel corresponding to each input-output pair. The length of each FIR kernel is selected by trial and error, where prior knowledge about the identified system is used in the first trial. Generally, too short memory lengths (severely under-parametrized model) will not capture input-output relationships accurately enough, while too long memory lengths (severely over-parametrized model) will generate inaccurate parameter estimates by fitting data to noise.

The linear parameter estimation methods discussed in section IV were applied on eight of the 11 data sets, and models were developed. All models were validated on the remaining three data sets. Validation results for the Ψ ellipsometry signal are shown in Figures 1 through 5. Similar results were obtained for the Δ signal. All computations were performed in Matlab. Use of the PLS toolbox was made.

Figure 1 shows validation results for OLS, RR, and RRDlin. Here the solid line is the experimental data, and ●, ▲, and ■ represent predictions by models developed through OLS, RR, and RRDlin, respectively. It is apparent that the RRDlin performance is better than that of OLS and RR, especially during the steady-state period. This is because the matrix $\mathbf{\Phi}^T\mathbf{\Phi}$, which must be inverted to obtain the OLS estimates (eqn. (6)), is ill-conditioned, so the OLS estimator has large variance. This behavior is typical of OLS in FIR modeling, which involves the estimation of many highly correlated parameters. The optimal degree of constraints for RR and its variant RRDlin are determined from the relationship between the regression parameter K and total predictive residual sum-of-squares (PRESS) for both outputs (Ψ and Δ) on the basis of the three validation data sets. Usually, the value of PRESS as a function of K decreases rather rapidly at first to reach a minimum for some K and then it increases again, to approach quasi steady state (slow growth).

Figure 2 shows identification and validation results for CR. When employing the CR algorithm, data have to be auto-scaled before use. Otherwise, numerically larger variables would influence the results more than numerically smaller variables, regardless of whether they were really important in describing the

data set. The "leave-one-out" cross-validation method was used to determine the optimum number of latent variables. The cross-validation procedure in CR involves a two-dimensional search for the optimal number of latent variables (LVs) and the "SVD power" μ that minimizes the prediction error, as discussed in section IV. Figure 2 shows how the value of PRESS varies as a function of the number of LVs and μ. Between PCR and OLS μ varies in logarithmically spaced intervals from μ=10 (near PCR) to μ=0.01 (near OLS). It is clear that PCR does not depend strongly on the number of LVs, while the larger LV is, the lower the PRESS is for OLS. For the same LV, the value of PRESS will decrease with increasing μ. Several typical features of the CR validation error surface can be found in the second diagram of Figure 2. The flat "OLS plain" indicates that the number of LVs will have little effect on the models in the region with low μ. As μ increases, the regression gives more importance to input block variance. Thus models change from the OLS through the PLS to the PCR as the number of LVs increases. At the other extreme, the "PCR mountain" with few latent variables represented in the PCR, has poor prediction accuracy. Between the extremes lies a region where the optimal PRESS is located. Figure 2 indicates an optimum for number of LVs approximately equal to 3, and $\mu \approx 1.5$.

The key point to obtain a good model in the wavelet-based parsimonious identification scheme is the selection of the nonzero coefficients in $\tilde{\mathbf{h}}$. Coefficients that must set to zero have to be selected iteratively[7]. An advantage of the method is that it avoids a combinatorially large number of trials in order to arrive at the best $\tilde{\mathbf{h}}_c$. Indeed, a large number of coefficients corresponding to the tail ends of the kernels are virtually equal to zero. At the starting point of the iterations, one can use information from either first principles modeling or form empirical modeling at a different operating point. As an example, Figure 3 shows the DWT of the OLS-identified linear kernel between the Ψ signal and the OES288 input signal at three resolution levels. Obviously, in the tail part and/or high-resolution level, most of the coefficients are negligible, and it appears that setting them to zero will not affect the actual kernel significantly. For this study, no prior information about the identified process is available, so only the generic nature of FIR can be made use of to compress the linear kernel $\tilde{\mathbf{h}}$.

Figures 4 and 5 show validation results for wavelet-compressed FIR models for the Ψ signal and Δ signal, respectively. The degree of compression was 73% for Ψ and 79% for Δ. In each figure, the thick

solid line represents the experimental data, while the solid points are generated by the model. Clearly, the amplitude and frequency of both Ψ and Δ are captured reasonably well. The autocorrelation of residuals for the Δ signal are shown in Figure 6. The residuals are not white noise, indicating that the model developed has fitted some noise. This is due to the fact that the original data were collected under conditions that were selected to be close to normal operating conditions and not conditions optimal for model development (i.e., the identified process was not excited enough to reveal its actual behavior).

Finally, all validation results by the linear approaches mentioned above are summarized in Tables I and II. For comparison purposes, Tables I and II also include identification results obtained by using the PO-MOESP subspace identification method[8]. The reason for considering the PO-MOESP method is that it is effective at finding accurate models without preliminary estimation of various structural indexes, such as model order, and can reduce the effects of noise on input and output simultaneously. Variance-accounted-for (vaf), an indication of closeness between the original signal and its model estimate, is applied to evaluate the identification approaches studied above. The best value for vaf is 100 and the worst -1000. Comparison indicates that the wavelet-compressed FIR model produces excellent results in terms of both higher vaf value and good shape predictability (trends and peaks).

VI. CONCLUSIONS

In this paper, we have compared several estimation methods to identify an empirical model for a helicon plasma reactor from limited data. Identification and validation comparisons have been made on the basis of closeness of fit to the true process. Of the proposed model structures, the parsimonious FIR structure generated the best results. It also exhibited improved performance over a neural network-based approach[10] for the same experimental data. The CR algorithm can also provide good identification results. The optimal models generated by CR fall in the middle of the continuum region between OLS and PLS in this case, the exact position depending on the selection of the CR parameters μ and the number of latent variables.

The results of this work indicate that while an empirical model can be built for a plasma etching reactor on the basis of data alone, the conditions under which the data are generated are important. Future

studies should be undertaken to reveal how data can be generated without upsetting the reactor, and how production data could be used to refine an existing model.

Acknowledgement

The first three authors would like to thank Dr. Ed Rietman of Lucent Technologies for allowing access to the experimental data on which this work is based, as well as for providing valuable suggestions on how to improve the manuscript.

---

1. Azzam, R.N.M., and N.M. Bashara, *Ellipsometry and Polarized Light*, North-Holland, Amsterdam, (1987).
2. Economou, D. J., T. Panagopoulos, and M. Meyyappan, *MICRO magazine*, 101 (1998).
3. Haas et al, *J.Vac.Sci. Technol. A*, 16 (3), 1117 (1998).
4. Hoeral, A.E., and Kennard, R.W., *Technometrics*, 12, 55 (1970).
5. Kozub, D., *Personal Communication*, 1994.
6. Mallat, S. G., *Trans.Amer.Math.Soc.*,315, 69 (1989).
7. Nikolaou, M. and P. Vuthandam, , *AICHE J.*, 44, 141 (1998).
8. Nikolaou, M., and H. Zhang, *Proceedings of the Electrochem. Soc. Annual Meeting* (1999).
9. Panagopoulos,T., *Three-Dimensional Simulation of Inductively Coupled Plasma Reactors*, Ph.D. Thesis, University of Houston, (1999).
10. Rietman, E.A., D.E. Lbbotson, and J.T.Lee, *J.Vac.Sci.Technol. B* 16, 131 (1998).
11. Sharfaty, M., C. Baum, M.Harper, N.Hershkowitz, and J.L.Shohet, *Jpn. J. Appl. Phy.s*, 37, 2381 (1998).
12. Stone, M., and R. J. Brooks, *J. R. Stat. Soc. B*, 52, 237-269 (1990).
13. Wise, B. M., and N. L. Ricker, *Journal of Chemometrics*, 7, 1 (1993).

List of Tables

List of Figures

TABLE I. Comparison of the identification (ID) and validation results for the Ψ ellipsometry signal

Method	ID (vaf)	Validation (vaf)		
		1^{st} data-set	5^{th} data-set	7^{th} data-set
OLS	72.9	-92.9	54.9	46.3
PCR	15.3	30.5	10.1	2.7
PLS	15.1	33.1	9.1	4.7
PO-MOESP	73.2	-148.3	59.6	35.4
CR	64.5	-12.4	48.4	46.0
RR	66.1	6.2	48.4	43.9
RRDlin	67.6	0.8	54.5	44.9
QS-Wavelet	63.1	56.9	69.0	60.0

TABLE II. Comparison of the identification (ID) and validation results for the Δ ellipsometry signal

Method	ID (vaf)	Validation (vaf)		
		1^{st} data-set	5^{th} data-set	7^{th} data-set
OLS	67.0	-1.9	-274.3	-36.4
PCR	10.8	15.0	-1.5	21.5
PLS	11.6	15.9	-1.7	21.4
PO-MOESP	67.1	20.8	14.6	6.9
CR	40.0	32.9	25.6	36.0
RR	50.4	34.7	12.5	42.9
RRDlin	47.9	33.2	10.7	29.3
QS-Wavelet	53.6	36.3	66.9	51.0

FIG 1 M Nikolaou

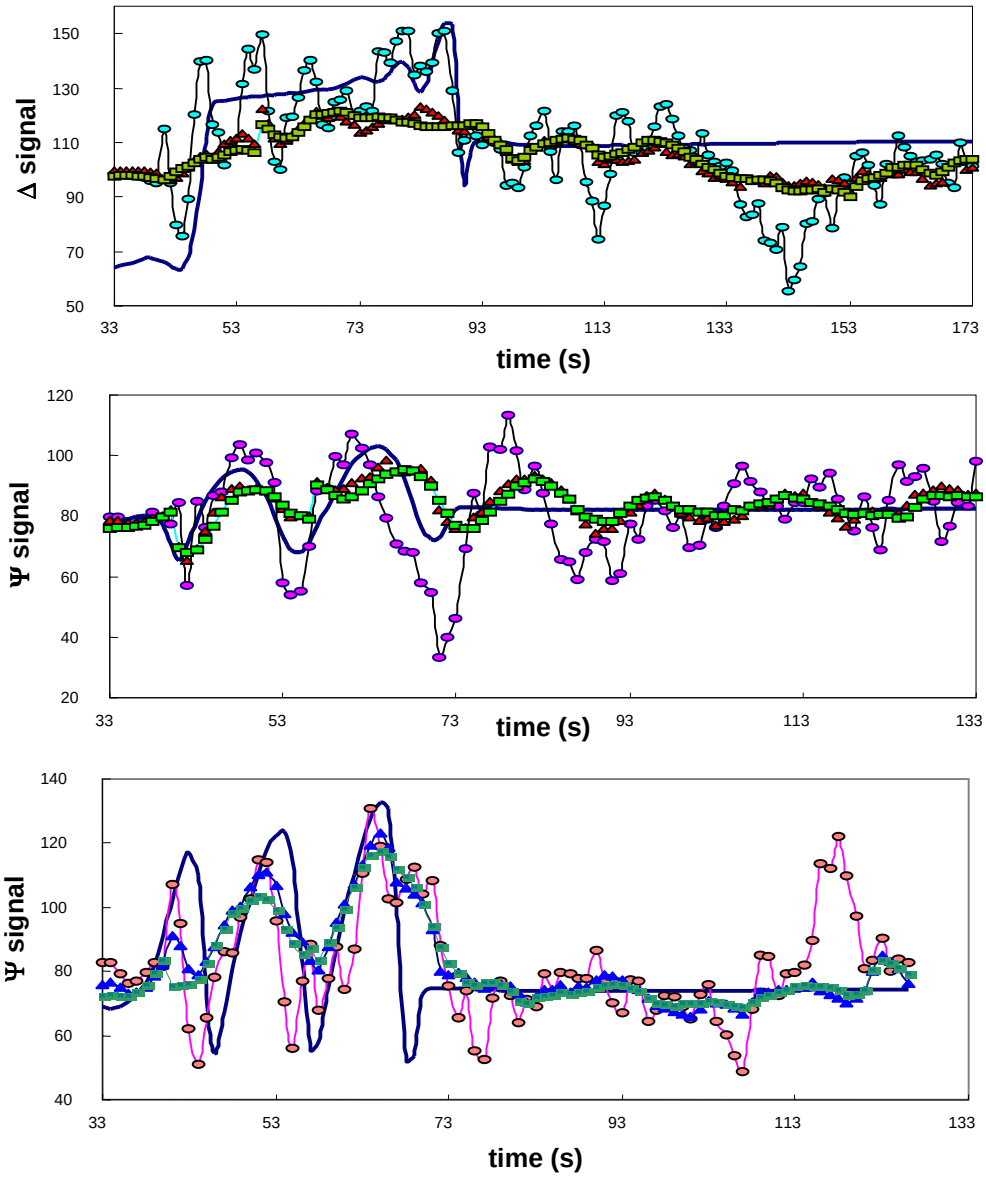
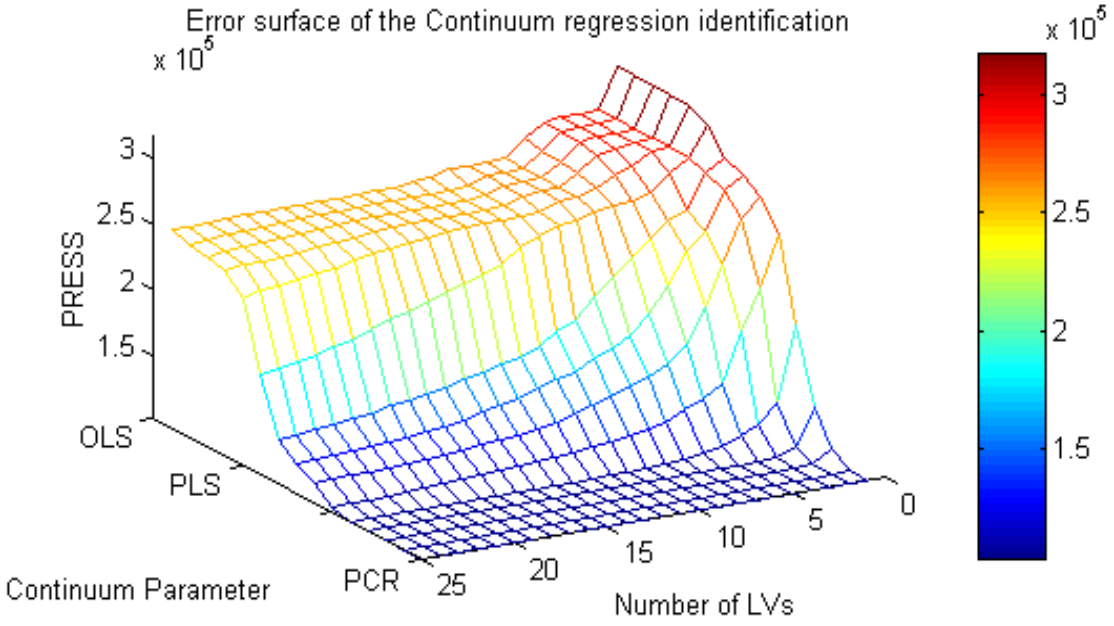


FIG 2 M.Nikolaou JVSTB



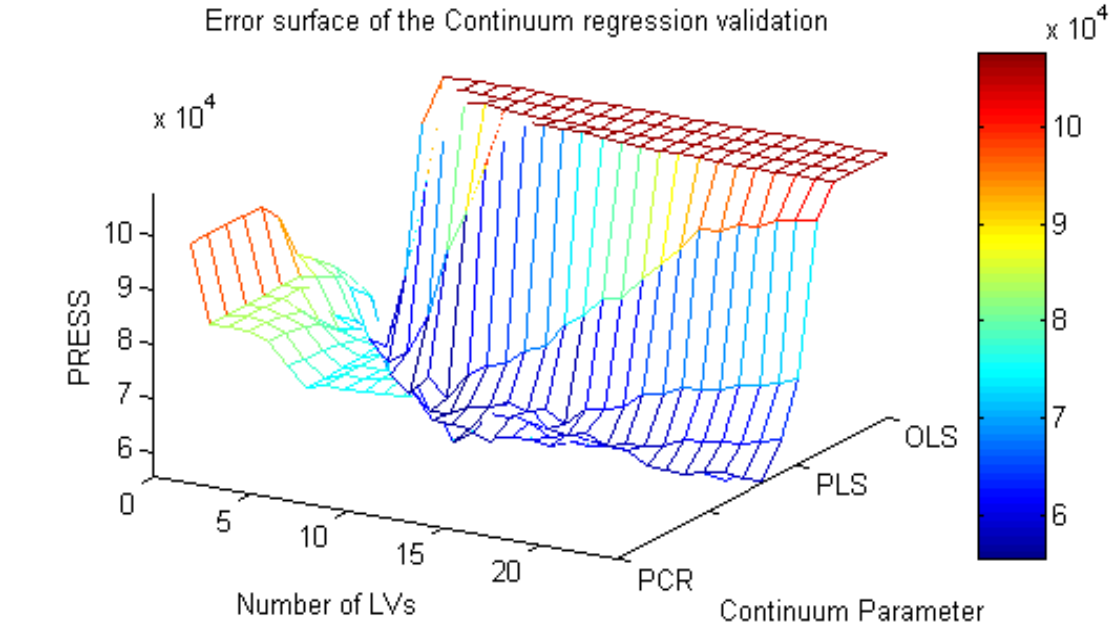


FIG 3 M.Nikolaou JVSTB

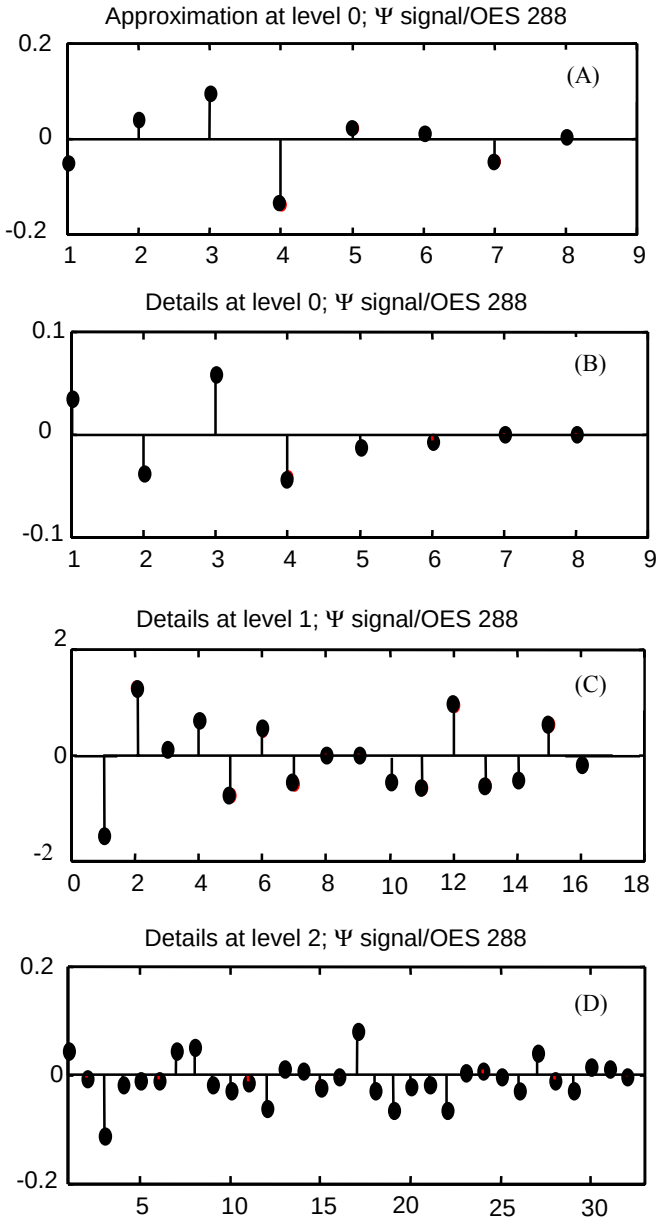


Figure 3

FIG 4 M.Nikolaou JVSTB

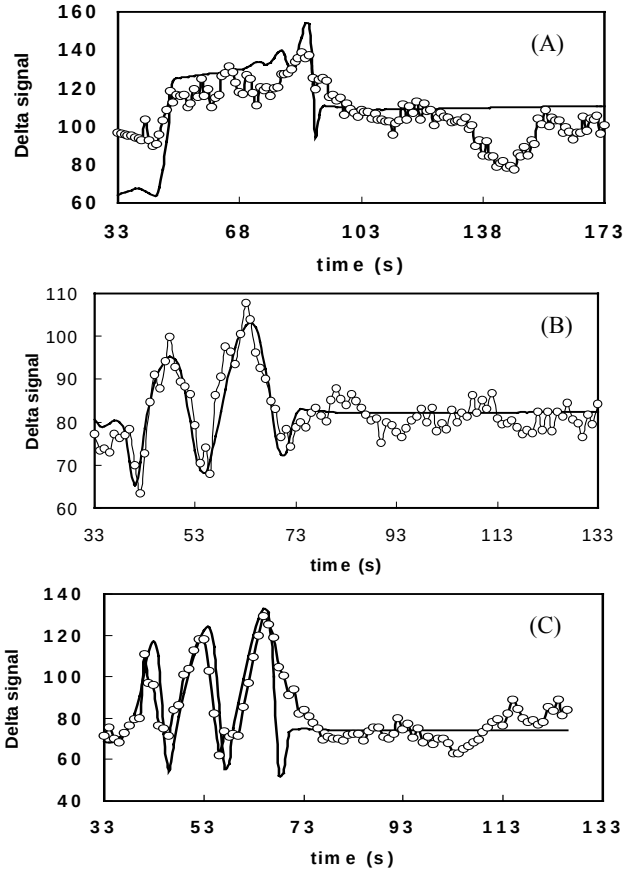


FIG. 4. Validation results of Δ signal for three test data-sets by QS wavelet compression

FIG 5 M.Nikolaou JVSTB

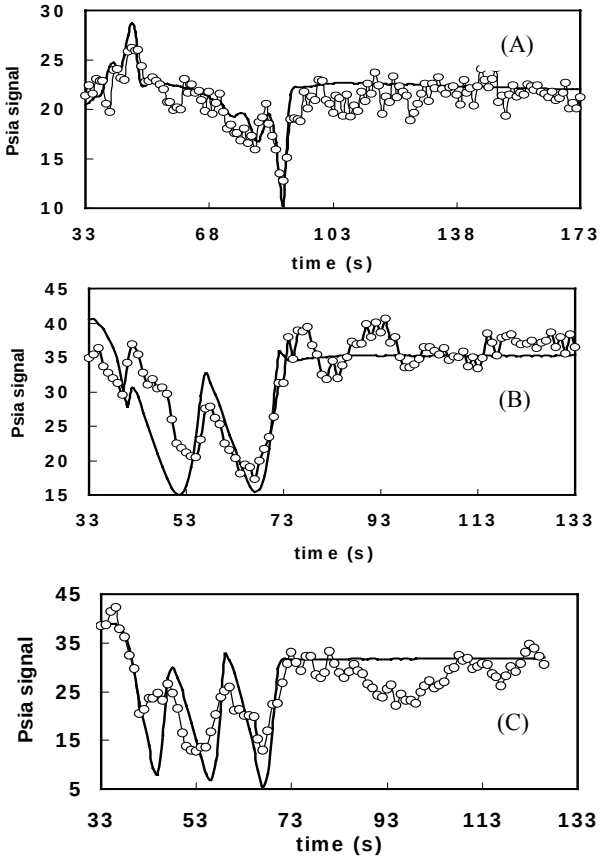


Figure 5

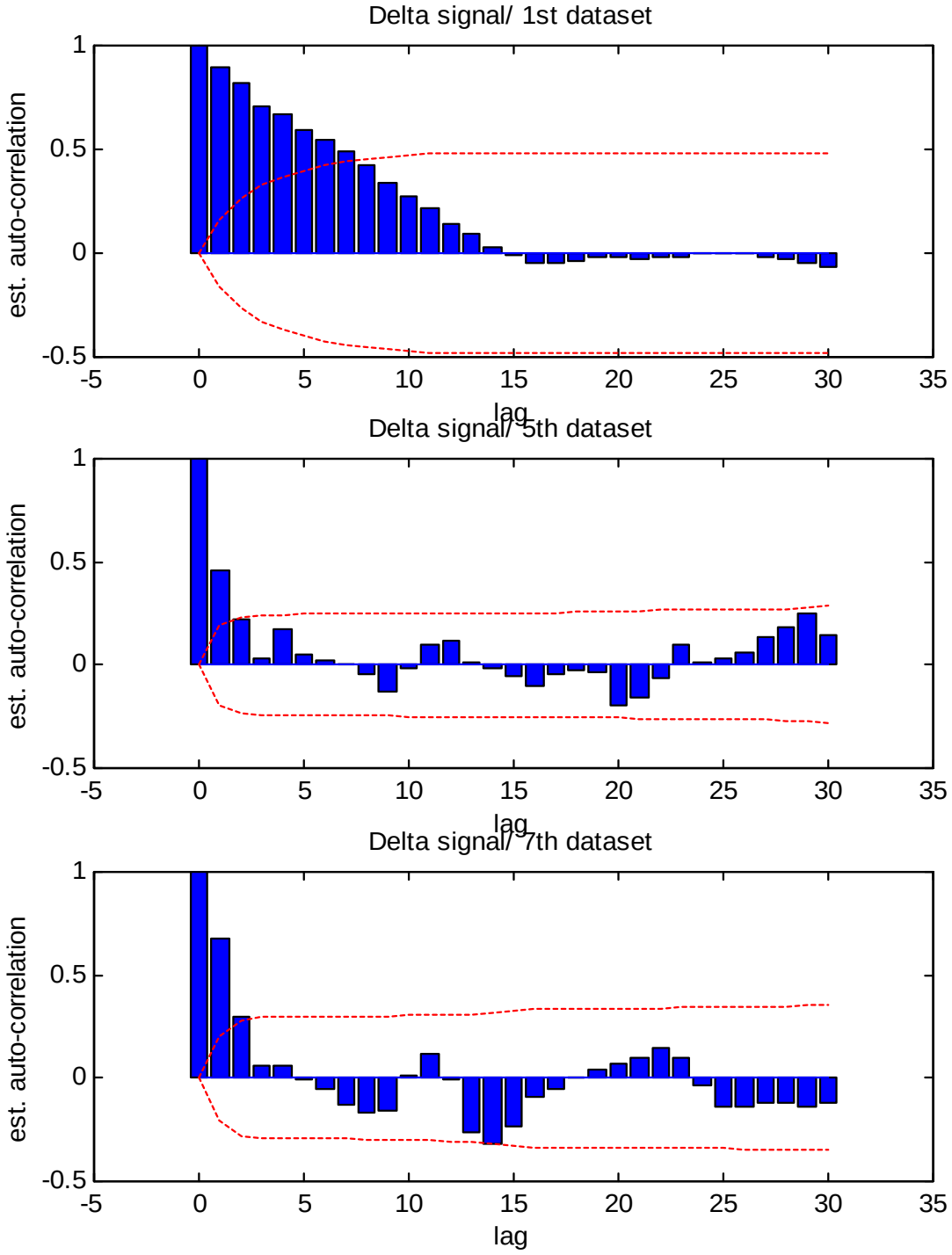

Delta signal/ 1st dataset
est. auto-correlation
lag
Delta signal/ 5th dataset
est. auto-correlation
lag
Delta signal/ 7th dataset
est. auto-correlation
lag

Figure 6