Virtual Metrology for Plasma Etch using Tool Variables

Shane Lynn*, John Ringwood*, Emanuele Ragnoli*, Sean McLoone* and Niall MacGearailt†
*Department of Electronic Engineering
National University of Ireland, Maynooth, Co. Kildare, Ireland
Email: shane.a.lynn@eeng.nuim.ie
†Dublin City University, Dublin, Co. Dublin, Ireland
Email: niall.macgearailt@dcu.ie

***Abstract*—This paper presents work carried out with data from an industrial plasma etch process. Etch tool parameters, available during wafer processing time, are used to predict wafer etch rate. These parameters include variables such as power, pressure, temperature, and RF measurement. A number of variable selection techniques are examined, and a novel piecewise modelling effort is discussed. The achievable accuracy and complexity trade-offs of plasma etch modelling are discussed in detail.**

I. Introduction

Plasma etching is a complex dynamic process used in semiconductor manufacture during which material is removed from the surface of product wafers using gases in plasma form. Plasma etch is preferred to older wet etch methods as anisotropic etch profiles can be achieved with excellent across-wafer uniformity. However, due to the complexity of the process, plasma etching is notoriously difficult to model and hence troublesome to control. Measurements of etch depth and etch profile are not usually available to machine operators for several days after the completion of the etch process. Hence, control is difficult to implement with this inherent measurement delay. A machine running out of specification without being detected can lead to days of scrapped materials and/or damaged equipment.

Whereas downstream data is recorded intermittently and delayed, a great deal of inline data is often recorded from the processing tool during the etch process. Variables such as pressure, temperature, RF power and phase, plasma impedance, and gas flow rates are available in real time during wafer processing. "Virtual metrology" is a relatively new technology that is gaining a following in this industry whereby readily available measurements are used to estimate the actual wafer state and etch results [1]. With a full virtual metrology system in place, "virtual" measurements of etch depth would be available for each wafer directly after processing, reducing wafer scraps and enhancing the etch process dramatically.

The determination of a model for virtual metrology schemes is challenging due to the low supply of actual metrology values for training, the natural drifting behavior of the tools, and the effect of periodic maintenance cycles. This paper explores some methods of variable selection for modelling of etch rate, examines a piecewise modelling effort to counteract the effect of maintenance cycles, and discusses the obtainable accuracy and complexity compromises of modelling the plasma etch process.

II. Variable Selection Techniques

During plasma etch processing, a vast amount of information is recorded from etching machines and various other diagnostics. This situation leads to a surplus of available data for each wafer processed. Deciding which variables are most useful to explain variations in the etch output is a challenging task. Modelling from first principals is a complicated option, and leads to computer models that take hours or days to compute seconds of a plasma etch simulation [2]. Relating the bulk plasma and etch tool parameters to etch parameters on a nanometre scale is a daunting task, and engineers in the area often turn to statistical methods to model variations in the etch process [3] [4]. This paper examines three different statistically-based methods for variable selection.

A. Principal Component Analysis

Principal Component Analysis (PCA) is a method used to transform a set of correlated variables into new uncorrelated variables, known as *principal components* (PCs). Each PC is a linear combination of the original variables. They are arranged in order of the variance that each one explains in the original dataset [5]. It is often used with plasma etch to compress Optical Emission Spectroscopy (OES) measurements [6].

Before PCA is performed on a set of data, $\mathbf{X}$, made up of n samples (rows) and m variables (columns), it is usual to offset each variable to have zero mean, and sometimes to scale each variable to unit variance. This is useful if the original data has variables with many different magnitude scales. Scaling to unit variance gives all variables equal importance for the analysis. PCA performs an eigen-decomposition of the covariance or correlation matrix of the data matrix $\mathbf{X}$, which decomposes $\mathbf{X}$ as the sum of the outer product of vectors $\mathbf{t}_i$ and $\mathbf{p}_i$ plus a residual matrix $\mathbf{E}$ [6].

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + ... + \mathbf{t}_l\mathbf{p}_l^T + \mathbf{E} \quad (1)$$

$$= \mathbf{TP^T} + \mathbf{E} \quad (2)$$

where,

$$\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ...\mathbf{t}_l] \quad (3)$$

$$\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ...\mathbf{p}_l] \quad (4)$$

and l is the number of PCs. The vectors $\mathbf{t}_i$ are known as the *scores* and $\mathbf{T} \in \mathbb{R}^{n\times 1}$ the score matrix; the $\mathbf{p}_i$ vectors are the *loadings* and $P \in \mathbb{R}^{m\times 1}$ the loadings matrix. For PCA, the decomposition of $\mathbf{X}$ is such that the loading matrix $\mathbf{P}$ is orthonormal and the score matrix $\mathbf{T}$ is orthogonal. The first PC is the linear combination of the m original variables that explains the greatest amount of variability ($\mathbf{t}_1 = \mathbf{X}\mathbf{p}_1$). In the m-dimensional variable space, the loading vector $\mathbf{p}_1$ defines the direction of the greatest variance [7]. Overall, loadings represent how the original variables are combined to make the PCs, scores represent original data projected onto the new uncorrelated variables, and finally, $\mathbf{E}$, the residual, represents the data that is left unrepresented by the model. For a matrix $\mathbf{X}$ of rank r, r PCs can be calculated. However, the first k ($k < r$) of these may be sufficient to explain the bulk of the variance in the data. If $k = dim(\mathbf{X})$, then $\mathbf{E} = 0$, and the representation of the data is exact for the new variables (PCs).

PCA can be used as a variable selection technique by examining the loading vectors for the first few principal components. The variables that contribute the most variance to these components will have the highest loading values. These variables can then be selected as inputs to etch rate models. Using the principal components themselves as inputs to regression based models is known as Principal Component Regression (PCR), and has been applied to plasma processing in [8].

B. Correlation Methodology

An arguably simpler method to select important variables is to analyse the linear correlations between each etch chamber variable and the etch rate recorded. The correlation between two variables is defined as

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{E((x-\mu_x)(y-\mu_y))}{\sigma_x \sigma_y}, \quad (5)$$

where x and y are two variables, with mean values μ_x and μ_y and standard deviations σ_x and σ_y. E is the expected value operator, and cov denotes covariance. The correlation coefficient $\rho_{x,y}$ cannot exceed 1 in absolute value, and is a measure of the degree of linear relationship between two random variables. The closer the correlation coefficient is to -1 or +1 the more closely the two variables are related.

As correlations measures only the degree of the linear relationship between two variables, it is useful to pass the input variables through non-linear transforms before correlation tests, as a test for some non-linear relationships. For this variable selection technique, each input is raised to a number of powers before correlation testing (e.g. x^1, x^2...x^n). It was found that this increased the correlation between input and output vectors dramatically for some variables.

After all of the variables have been correlated with the output, they are ranked in order of correlation coefficient, and then the most correlated variables are used as inputs to regression or neural-network based models.

C. Stepwise Regression

Stepwise regression is a technique in which a linear regression model is produced to model an output variable, but the regression variables are chosen automatically using a number of criteria. Stepwise Regression is a combination of two separate stages, forward selection, and backward elimination.

The forward selection algorithm begins with a model with no predictor variables. Variables are entered into the model one at a time in an order determined by the strength of their correlation with the output. At each step, the p-value of an F-statistic is computed to test models with and without a new potential variable. If a variable is not currently in the model, the null hypothesis that the term would have a zero coefficient if added to the model is tested. If there is sufficient evidence to reject the null hypothesis, the term is added to the model. The F-distribution at each stage can be expressed as [9]

$$F = \frac{RSS_0 - RSS_1}{RSS_1/(n-p+1)} \quad (6)$$

where RSS_0 is the residual sum of squares of the original model without the additional variable, RSS_1 is the residual sum of squares with the new variable included, p is the number of variables present in the larger model, and n is the number of samples.

A backward selection algorithm starts with all of the available regression variables in the model. These are then removed in order of the weakest predictors first. Removal continues until only useful predictor variables remain in the model.

A stepwise algorithm is a combination of the above methods. Variables are added in sequence to a model, and their value assessed. If the variable adds value, it is retained, but all other variables in the model are then retested to examine if they are still contributing to the success of the model. They are removed if they do not contribute significantly [10]. The p-value limits that are used to judge whether variables are kept or added to the model are set by the user. For the purpose of this study, p-values of 0.05 and 0.10 were used for addition and removal of regression variables respectively.

III. Description of Dataset

For this paper, data was collected from an industrial silicon etch process over a period of six months. The analysis is carried out on production data for a well-controlled, capacitively coupled trench etch process. All of the data is obtained from the same etch machine, and the dataset spans approximately 18 maintenance cycles that are carried out approximately once every 1000 wafers. The process in question has a total of five etch steps, each entailing different etch gases and materials. Chamber setting such as power and gas flow are adjusted for each step.

Time series data for the chamber parameters are measured during etch time. Due to the large amount of data present, these are further compressed into bulk statistics (mean and standard

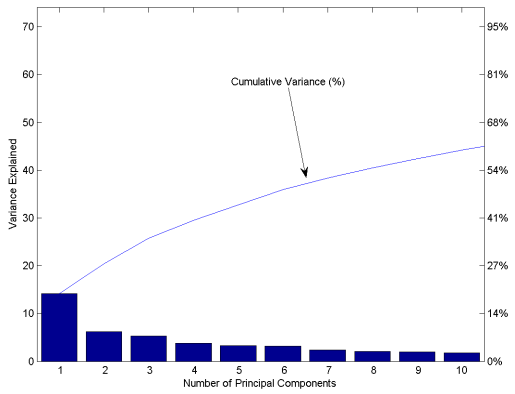

Fig. 1. Variance Explained as a function of Principal Components for input data.

deviation) for the purpose of our investigations. For each step in the process, an average of 30 variables are collected, leading to over 150 variables in total. As the main etch step is later in the process, only variables from steps 3-5 are included in this analysis. This reduces the variable set to 85. The variables measured from each step include:

- Mean and standard deviations of gas flow rates into the plasma etch chamber taken from flow rate meters.
- Readings from the RF system such as phase, power, voltage and a calculation of plasma impedance.
- Measurements from the etch chamber's impedance matching unit, such as capacitor and coil positions.
- Chamber measurements, such as pressure and electrode temperature.
- Endpoint traces from monochronometers used for end-point detection of some of the process steps.
- Wafer context data, such as processing time, lot number, wafer number, maintenance counter and product type.

Although etch tool variables are measured for every wafer processed, actual measurements of etch depth are carried out at a much lower frequency. In total, there are approximately eight hundred useable measurements of etch rate contained in the dataset. The wafers are processed in lots of 25, and etch depth measurements are taken from slots 13 and 25 approximately once in every 2 lots, leading to measurements of approximately 4% of wafers. There is a measurement delay for etch rate of up to three days for each measured lot.

A PCA of the data, after mean and standard deviation normalisation demonstrates that the individual variables have very little correlation between them. This is seen as the overall variance for the dataset cannot be explained using a small number of principal components (see Figure 1).

IV. Modelling Techniques

After variable selection has been carried out using one of the techniques outlined in section II, the next step is to build a model around the chosen variables in order to estimate etch rate for future wafers. Two modelling techniques are used in this work.

A. Multiple Linear Regression

Multiple Linear Regression (MLR), also known as Ordinary Least Squares Regression (OLSR) is a linear method that attempts to model the relationship between two or more regression variables and an output variable by fitting a linear equation to the observed data. This leads to a model of the form

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1} \tag{7}$$

being fit to each data point. Here, y is a measured output, and $x_1 \ldots x_{p-1}$ are system inputs that can be used as regression variables. We denote the data points to be

$$y_i, x_{i,1}, x_{i,2}, \ldots, x_{i,p-1}, i = 1, ...n$$

The output observations y_i will be represented by the vector $\mathbf{y}$, the unknown model parameters, $\beta_0, \beta_1, \ldots, \beta_{p-1}$ by the vector $\boldsymbol{\beta}$, and the data matrix $\mathbf{X}_{n\times p}$ takes the form:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,p-1} \end{bmatrix} \tag{8}$$

Hence for a given $\boldsymbol{\beta}$, the vector of predicted values $\hat{\mathbf{y}}$ is given by

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} \tag{9}$$

The ordinary least squares solution for $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{10}$$

provided that $\mathbf{X}^T\mathbf{X}$ is nonsingular [11]. Models created using MLR are computationally quick to train as they require only simple matrix operations.

B. Artificial Neural Networks

Artificial Neural Networks (ANNs) are networks of interconnected artificial neurons that can be used for information processing. Each artificial neuron, or node of the network, receives inputs from other nodes, and produces an output based on its internal activation function (which can be non-linear). The nodes are usually arranged in layers, with all neurons in a layer receiving weighted outputs from all neurons in the preceding layer, and in turn, passing their outputs through weights to the next layer. This is known as a feedforward neural network. It has been shown that a feedforward ANN network with one hidden layer can approximate almost any continuous function [12]. Neural networks have been applied often in the literature to plasma processes due to their ability to approximate nonlinear functions, and have been shown to outperform statistical techniques such as PCR, MLR, and Partial Least Squares Regression (PLSR) on some datasets [13][8].

This paper makes use of Multi-Layer Perceptron (MLP) neural networks consisting of three layers: an input layer,

an output layer, and one hidden layer. The output neurons used linear activation functions, while the other layers used nonlinear sigmoid activation functions.

MLPs are trained by starting the network with random weights (multiplicative values applied to signals between neurons). The dataset is split into two parts, a training set and a validation set. For the training set, the output of the network is calculated. The sum squared error E is defined as [14]

$$E = \sum_{j=1}^{N} (y_j - \hat{y}_j)^2 \tag{11}$$

where N is the number of samples in the training set, y_j is the desired output for sample j, and $\hat{y}_j$ is the calculated output at sample j. When training the MLP this error, E, is minimised via a *gradient descent* algorithm, where the weights are adjusted each iteration in the direction of decreasing E. A general rule for MLP weight optimisation is expressed as:

$$\mathbf{W}(m+1) = \mathbf{W}(m) + \eta \Delta \mathbf{W}(m) \tag{12}$$

where $\mathbf{W}$ is a matrix of the network weights and $\Delta\mathbf{W}$ is the calculated change in weight to minimize the error, E.

$$\Delta \mathbf{W} = -\frac{\delta \mathbf{E}}{\delta \mathbf{W}} \tag{13}$$

The other parameters, m and η are the training iteration number (the gradient descent is iteratively completed on the dataset until an error minimum is found) and the *learning rate* respectively [15]. Test datasets can be used in parallel to the training to assess when networks are becoming overtrained on the training set. In practical circumstances, more sophisticated and computationally efficient error minimisation techniques are used, such as the 2nd order gradient descent BroydenFletcherGoldfarbShannon (BFGS) method [16].

To combat the possible effects of random weight initialisation, several networks with the same structure are trained, initialising the weights randomly each time. Optimisation of the number of hidden neurons and the gradient of activation functions can also assist with accuracy [17]. In general, neural networks can be computationally demanding to train due to the iterations required for the gradient descent method to converge. They are also quite data-hungry, often requiring a large number of samples to develop a useful model [18].

V. Data Disaggregation

In an attempt to more accurately model the variations in etch rate over the dataset, a disaggregation of the input data is explored. For the wafers in the dataset, a *count* variable is provided that indicates the position of that wafer in the current maintenance cycle. Data is disaggregated into three separate datasets. Each dataset contains wafers with different ranges the count variable. The first set contains all input and output information for wafers numbered 1–300, the second 301–600, and finally the third 601–1000. Each different dataset is then modelled completely separately from the others as shown in Figure 2.

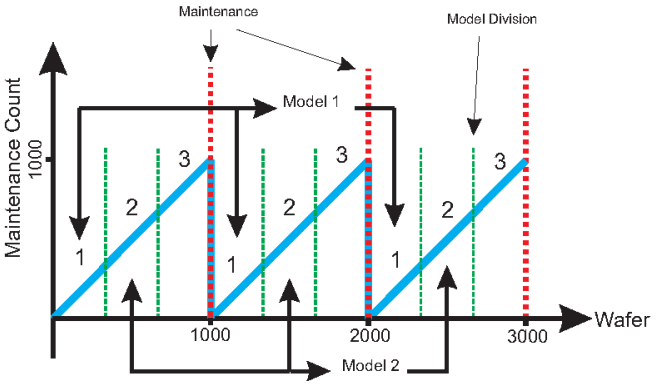

Fig. 2. Schematic of data disaggregation scheme.

The aim of this scheme is to exploit any similarities that may exist between different stages of maintenance cycles. It is conjectured that the beginning sections, middle sections and end sections of individual maintenance cycles may be more similar to the corresponding sections in other cycles than to the different sections of the same cycle.

VI. Results

This section outlines the performance of the variable selection and modelling techniques investigated. The performance ratings used for this study are Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE), defined by

$$\text{APE}(\%) = \frac{\sum_{i=1}^{N} \frac{|(\hat{y}_i - y_i)|}{y_i}}{N} \times 100 \tag{14}$$

and

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}}, \tag{15}$$

where $\hat{y}$ is the predicted value, y is the actual value and N denotes the number of samples available.

A. Variable Selection

A number of models are generated to model the data from a global perspective, with the intent of investigating the effects of different variable selection techniques. The dataset is split into two sections; one for the training of the models, and the second to act as completely "unseen" data for the predictions. For the 18 maintenance cycles available, 16 are used to train the prediction models, and 2 to judge generalisation performance. In terms of metrology vectors, this leads to a training set with 703 etch rate measurements, and a validation set with 103 measurements.

The results for the models, based on multiple linear regression, are shown in Table I. All models are based on a subset of approximately 16 predictor variables, as this was the number of variables chosen by the stepwise regression technique.

It can be immediately noticed that the PCA based selection method performs most poorly out of the three methods examined. Increasing the number of regression variables for this model to 21 (7 variables from each of the first 3 principal

TABLE I
MULTIPLE LINEAR REGRESSION MODEL PERFORMANCE WITH DIFFERING VARIABLE SELECTION TECHNIQUES

Method	MAPE (%)	RMSE
Correlation	1.45	1.14
PCA	2.04	1.53
Stepwise	1.38	1.13

components) leads to a model with much better error figures of 1.3% and 1.076 for MAPE and RMSE respectively.

For the models in Table I, the main disadvantage of the correlation and PCA based variable selection method is that there is a high probability of the algorithms choosing predictor variables that are correlated with one another. These extra variables are added to the prediction model, but do not add much extra information or value to the prediction accuracy. This phenomenon arises in the correlation selection algorithm from signals such as power and pressure from different process steps. For example, power from step 3 and step 4 are selected as candidate variables by this technique, whereas the correlation between these signals is 0.9981. Adding both to a linear model is of very little value.

In the case of PCA, all of the variables from the same principal component are likely to be similar as they are used to describe the same component of the dataset variance. Hence, selecting five variables from the same principal component may actually add very little new information to the model. Another complication to this selection technique is that the PCs are calculated without any reference to the output. The variables selected by the principal component model may best explain the largest variance in the input data, but may be poor predictors of the system output. A PCR model investigated for this dataset yielded very poor prediction results.

The stepwise regression method of variable selection has the advantage that the selected predictor variables are unlikely to be highly correlated. As per section IV, each variable is added to the model only if it contributes to the accuracy of the prediction. Adding a variable that is highly correlated to an existing variable in the model will not contribute significantly, and so there is a low probability of many correlated variables existing in the final model structure. As variables are assessed during the algorithm and removed if they no longer contribute, stepwise regression should lead to the most parsimonious model. Figures 3 and 4 show the correlation structure between the variables selected by the different algorithms. It is clear that the variables chosen by stepwise regression are less correlated. There are only four variables chosen in common between the two methods whereas the model's accuracy is similar. This suggests that there are very few key drivers of the variability in the dataset.

B. Data disaggregation

Table II describes the effect of the data disaggregation scheme on the model accuracies. The data is split for training and test in the same proportions as described in the previous section.

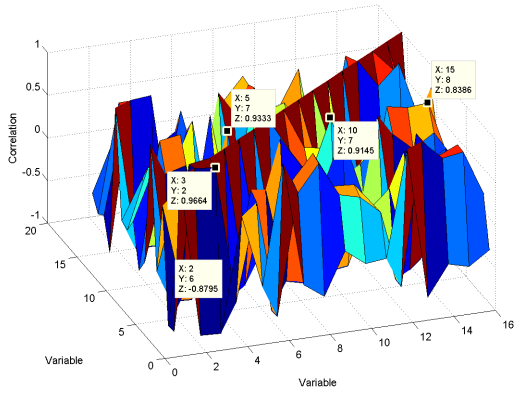

Fig. 3. Correlation structure for variables chosen with correlation method.

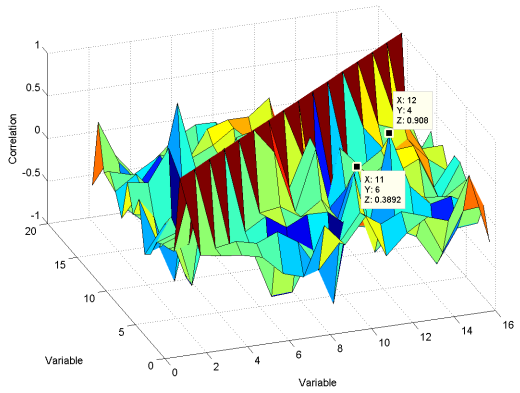

Fig. 4. Correlation structure for variables chosen with stepise regression.

TABLE II
EFFECT OF DATA DISAGGREGATION ON MODEL ACCURACY. THE FIRST COLUMN DENOTES THE METHOD OF VARIABLE SELECTION AND THE MODEL TYPE.

Method/Model	Full set		Disaggregated	
	MAPE (%)	RMSE	MAPE (%)	RMSE
Corr / NN	1.23	0.99	1.48	1.20
Corr / MLR	1.45	1.14	1.59	1.23
Step / MLR	1.38	1.13	1.67	1.30
Step / NN	1.36	1.12	1.49	1.18

The best performing model over all of the techniques investigated is a neural network based model using correlation selected variables (including higher orders of the variables). We can see from the results that the disaggregation of the data does not provide any increase in accuracy for the models, with the models based on the global dataset outperforming the others for all input combinations. The global nonlinear NN models appear to be capable of modelling the output across multiple maintenance cycles. It can be seen in Figure 5 that general trends are followed successfully, but high frequency etch rate deviations are sometimes missed. However, these are

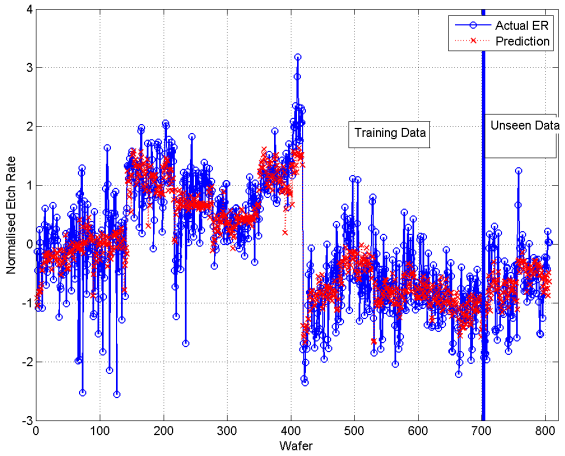

Fig. 5. Etch rate prediction from globally based neural network model.

well within tolerances.

VII. Conclusions

This paper has detailed an examination of various variable selection and modelling techniques in an effort to relate etch rate measurements to plasma processing tool parameters.

The achievable accuracy with such parameters for the trench process investigated is approximately 1.2% MAPE. It should be noted that this accuracy has been achieved with only etch tool measurements. Further accuracy may be possible with more sophisticated sensors, such as OES and PIM. However, a substantial cost can be associated with the installation of such systems. The models developed in this study are capable of following mean trends in the etch rate data, but the high frequency changes are not modelled accurately. To successfully model the more rapid changes in etch depth may require more advanced metrology. These changes may however be attributable to another step in the manufacturing process.

There is little difference in accuracy between models developed using different variable selection routines. This was despite the fact that different algorithms chose different inputs for their models. We can conclude that there are few variables that can act as "key" indicators of the process. A PCA analysis of the machine parameters confirmed that the dataset variance is spread across many variables. This is further compounded by the fact that completely different variables are chosen when the same techniques are applied to a different etch chamber. Both the correlation based and PCA based methods have the disadvantage that highly correlated variables can be chosen as model inputs. This does not apply to the same extent for the stepwise algorithm.

Using the data disaggregation technique suggested leads to no improvement in model accuracy for linear or nonlinear models. Global models perform better for this particular dataset, capturing the movement of etch rate across multiple maintenances. We can conclude that there is little consistent behaviour across different data segments across maintenance cycles. The inaccuracy of the disaggregated models may also be somewhat affected by the smaller training sets that arise from splitting up an already sparse dataset.

Acknowledgment

This project was funded by the Irish Research Council for Science Engineering and Technology (IRCSET).

References

[1] Y.-J. Chang, Y. Kang, C.-L. Hsu, C.-T. Chang, and T. Y. Chan, "Virtual metrology technique for semiconductor manufacturing," in *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, 2006, pp. 5289–5293.

[2] H. Kim, F. Iza, S. Yang, M. Radmilovic-Radjenovic, and J. Lee, "Particle and fluid simulations of low-temperature plasma discharges: benchmarks and kinetic effects," *Journal of Physics D: Applied Physics*, vol. 38, pp. R283–R301, Sept 2005.

[3] D. White, B. Goodlin, A. Gower, D. Boning, H. Chen, H. Sawin, and T. Dalton, "Low open-area endpoint detection using a pca-based t^2 statistic and q statistic on optical emission spectroscopy measurements," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 13, no. 2, pp. 193–207, May 2000.

[4] B. M. Wise, N. B. Gallagher, D. D. Butler, S. W.and White, and G. G. Barna, "A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process," *Journal of Chemometrics*, vol. 13, no. 3-4, pp. 379–396, 1999.

[5] A. Afifi, V. A. Clarke, and S. May, *Computer-Aided Multivariate Analysis*, 4th ed., C. Chatfield, M. Tanner, and J. Zidek, Eds. Chapman & Hall/CRC, 2004.

[6] H. H.Yue, S. J. Qin, J. Wiseman, and A. Toprac, "Plasma etching endpoint detection using multiple wavelengths for small open-area wafers," *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, vol. 19, no. 1, pp. 66–75, Jan/Feb 2001.

[7] J. Zhang, E. Martin, and A. Morris, "Fault-detection and diagnosis using multivariate statistical techniques," *Chemical Engineering Research & Design*, vol. 74, no. 1, pp. 89–96, January 1996.

[8] S. Lee and C. Spanos, "Prediction of wafer state after plasma processing using real-time tool data," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 8, no. 3, pp. 252–261, August 1995.

[9] D. Tsunami, J. McNames, B. Whitefield, P.Rudolph, and J. Zola, "Oxide etch rate estimation using plasma impedance monitoring," in *Proceedings of SPIE*, I. Emami, Ed., vol. 5755, no. 1. SPIE, 2005, pp. 59–68.

[10] N. Brace, R. Kemp, and R. Snelgar, *SPSS for Psychologists*. Palgrave Macmillon, 2006.

[11] J. A. Rice, *Mathematical Statistics and Data Analysis*, third edition ed., J. A. Rice, Ed. Thomson Brooks Cole, 2007.

[12] K. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Networks*, vol. 2, no. 3, pp. 183–192, 1989.

[13] C. Himmel and G. May, "Advantages of plasma etch modeling using neural networks over statistical techniques," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 6, no. 2, pp. 103–111, 1993.

[14] B. Kim and S. Kim, "Partial diagnostic data to plasma etch modeling using neural network," *Microelectron. Eng.*, vol. 75, no. 4, pp. 397–404, July 2004.

[15] B. Kim, K.-H. Kwon, S.-K. Kwon, J.-M. Park, S. Yoo, K.-S. Park, and B.-W. Kim, "Modeling oxide etching in a magnetically enhanced reactive ion plasma using neural networks," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 20, no. 5, pp. 2113–2119, 2002.

[16] R. Battiti and F. Masulli, "Bfgs optimization for faster and automated supervised learning," in *International Neural Network Conference*, vol. 2, 1990, pp. 757–760.

[17] S. Hong and G. May, "Neural-network-based sensor fusion of optical emission and mass spectroscopy data for real-time fault detection in reactive ion etching," *Industrial Electronics, IEEE Transactions on*, vol. 52, no. 4, pp. 1063–1072, Aug 2005.

[18] P. A. Cerny, "Data mining and neural networks from a commercial perspective," in *ORSNZ Conference Twenty Naught One*, 2001.