Gaussian Process Regression for Virtual Metrology of Plasma Etch

Shane Lynn[†] , John Ringwood[†], and Niall MacGearailt[*]

[†]*Department of Electronic Engineering*
National University of Ireland, Maynooth.

School of Electronic Engineering
[*]*Dublin City University, Dublin*

E-mail: [†] shane.a.lynn@eeng.nuim.ie john.ringwood@eeng.nuim.ie,
[*]niall.macgearailt@eeng.dcu.ie

Abstract — **Plasma etch is a complex semiconductor manufacturing process in which material is removed from the surface of a silicon wafer using a gas in plasma form. As the process etch rate cannot be measured easily during or after processing, virtual metrology is employed to predict the etch rate instantly using ancillary process variables. Virtual metrology is the prediction of metrology variables using other easily accessible variables and mathematical models. This paper investigates the use of Gaussian process regression as a virtual metrology modelling technique for plasma etch data.**

Keywords — **Plasma etch, virtual metrology, Gaussian process regression**

I Introduction

Plasma etching is a semiconductor manufacturing process where material is removed in exact amounts from the surface of silicon wafers. Etchant gases in plasma form are directed towards the wafer surface using electric and magnetic fields, where they react and evaporate to remove matter. Plasma etching is preferred to wet etching methods as it is capable of producing a highly anisotropic etch, allowing deep and narrow trenches to be etched into the wafer surface [1].

It is important to control and monitor the etch rate as precisely as possible since the ultimate trench depth and profile has a dramatic impact on the performance of the devices being produced. Measurements of etch rate and etch depth are expensive, time consuming, and invasive, sometimes requiring destruction of the device. Etch depth measurements are not available to machine operators without considerable metrology delay and it is not practical for fabrication plants to measure every wafer processed. Control is difficult to implement under such conditions and machines running out of specification without detection can lead to days of scrapped product. As Moore's law drives component dimensions smaller [2], there is an urgent requirement for more accurate control of such fabrication processes.

Fortunately a great deal of information is available from the etch chambers as each wafer is processed, such as chamber temperatures, pressures, gas flows, and electrical signals that are easily monitored during processing. The aim of virtual metrology (VM) schemes is to find mathematical relationships between these easily measured variables and the variables of interest such as the etch rate or etch depth.

Fab wide VM schemes have been suggested that use feed forward and feedback of information between different processes [3]. However, the individual VM schemes for each process are not trivial. The plasma etching process is notoriously difficult to model and predict, with more challenges arising with emerging technologies and techniques [4]. While the etch process itself is quite complex, modelling of the process is further complicated by multi-step recipes with changing chemistries, chamber conditioning effects, unpredictable shifts in process characteristics due to preventative maintenance (PM) operations, and a limited amount of downstream metrology data on which to validate modelling attempts. A comprehensive review of virtual metrology in plasma etch can be found in [5].

Previous work has examined the disaggregation of data sets to combat the effect of PM events [6], the application of windowed modelling methods to maintain model accuracy [7], and the identification of key process variables from data sets [8]. This paper focuses on the use of Gaussian process regression (GPR), a non-parametric modelling technique, to plasma etch data, paying particular attention to covariance function choice. While partial least squares (PLS) regression [9] and artificial neural networks (ANNs) [10] are regularly employed in the semiconductor manufacturing literature, the use of Gaussian processes (GPs) for regression and classification is a relatively new concept.

The remainder of this paper is set out as follows: Section II describes the GPR technique in detail. Section III provides details on the etch data set used for modelling. Sections IV and V give the results and conclusions of the paper respectively.

II Gaussian Process Regression

A GP can be viewed as a collection of random variables $f(x_i)$ with joint multivariate Gaussian distribution $f(x_1), f(x_2), \cdots, f(x_n) \sim N(0, \Sigma)$, where Σ_{ij} gives the value of the covariance between $f(x_i)$ and $f(x_j)$, and is a function of the inputs x_i and x_j, $\Sigma_{ij} = k(x_i, x_j)$ [11]. For the purposes of this discussion, let us assume a one-dimensional input-output process.

The covariance function $k(x_i, x_j)$ can be any function, provided that it generates a positive definite covariance matrix Σ. One of the most commonly used covariance functions is the *squared exponential* (SE) covariance function, which has the form:

$$k(x_i, x_j) = \nu^2 exp(-\frac{(x_i - x_j)^2}{2l^2}) \quad (1)$$

where ν and l are *hyperparameters* that vary the properties of the covariance function to best suit the training data set. The SE covariance function assumes that input points that are close together in the input space correspond to outputs that are more correlated than outputs corresponding to input points which are further apart. The parameter ν controls the scale of the variations between points x_i and x_j in the output space, while l, known as the length scale, determines the degree of variation in the input dimension. Hence, variations in l and ν control the smoothness of the covariance function. Examples of the effects of different length scales for a single-input single-output GP are shown in Figure 1. It can be shown that the use of a GP with a squared exponential covariance function is equivalent to modelling with a linear combination of an infinite number of Guassian shaped basis functions in the input space [12].

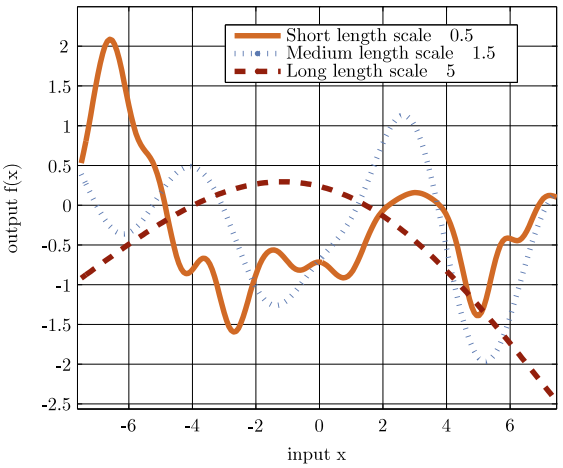

Fig. 1: Three possible outputs from GP models with differing length scales. The GP models with shorter length scales have more "flexibility" in the output space.

Gaussian process models fit naturally into the Bayesian modelling framework where, instead of parameterising the model function $f(x)$, we place a Gaussian prior on the range of possible functions that could represent the mapping of inputs x to outputs y. The Gaussian prior incorporates the analyst's knowledge about the underlying function in the data, and is specified using the GP covariance function.

We let the underlying function of our data be $y = f(x) + \epsilon$, where ϵ is a Gaussian white noise term with variance σ_n^2 such that $\epsilon \sim N(0, \sigma_n^2)$. A Gaussian process prior is put on the range of possible underlying functions $f(x)$ with covariance function as exemplified in (1) with unknown hyperparameters.

Hence we have

$$y_1, y_2, \cdots, y_n \sim N(0, K) \quad (2)$$

$$K = \Sigma + \sigma_n^2 I \quad (3)$$

where $\sigma_n^2 I$ represents the covariance between outputs due to white noise, where σ_n^2 is the noise variance and I is the $n \times n$ identity matrix. Our aim now is to use the set of training data points $\{x_i, y_i\}_{i=1}^n$ to find the posterior distribution of y_*, given input x_*, that is $p(y_*|x_*, \mathbf{x}_{tr}, \mathbf{y}_{tr})$, where $\{x_*, y_*\}$ denotes an unseen test data point and $\mathbf{x}_{tr}$ and $\mathbf{y}_{tr}$ denote the input and output training data. Before we find the posterior distribution of y_*, the unknown hyperparameters of the covariance function (1), l, ν, and the noise variance σ_n^2, must be optimised. This can be performed via a Monte Carlo method or, more typically, via maximisation of the log marginal likelihood

$$log(p(y|X)) = -\frac{1}{2}\mathbf{y}_{tr}^T K^{-1} \mathbf{y}_{tr} - \frac{1}{2} log(|K|) - \frac{N}{2} log(2\pi).$$

(4)

Equation (4) is made up of a combination of what can be termed a *data fit* term, $\frac{1}{2}\mathbf{y}_{tr}^T K^{-1}\mathbf{y}_{tr}$, that determines the success of the model in fitting the output data, along with a model complexity penalty $\frac{1}{2}log(|K|)$. By maximising the log marginal likelihood we find the model with the least complexity that fits the input-output data set most accurately. This optimisation problem is non-convex and requires the computation of the derivative of (4) with respect to each of the hyperparameters in the covariance function (1). Since typical gradient descent optimisation routines are sensitive to the initial choice of hyperparameters, during modelling, the initial values of the hyperparameters are initialised randomly up to five times to try to find the global minimum and the model with the lowest log likelihood is chosen for use on the test data.

With the hyperparameters optimised, we can now use the GP model to predict the distribution of y_* at the input x_*. The predictive distribution of y_*, $p(y_*|x_*, \mathbf{x}_{tr}, \mathbf{y}_{tr})$, can be shown to be Gaussian [12], with mean and variance

$$\mu(x_*) = \mathbf{k}_* K^{-1}\mathbf{y}_{tr} \tag{5}$$

$$\sigma^2(x_*) = k_{**} - \mathbf{k}_* K^{-1}\mathbf{k}_*^T + \sigma_n^2 \tag{6}$$

respectively, where $\mathbf{k}_* = [k(x_*, x_1), k(x_*, x_2), \cdots k(x_*, x_n)]$ is a vector of covariances between the test and training data points and $k_{**} = k(x_*, x_*)$ is the autocovariance of the test input.

The vector $\mathbf{k}_* K^{-1}$ can be seen as a vector of weights that form a linear combination of the observed outputs $\mathbf{y}_{tr}$ to form the prediction at x_*. The variance on the predicted values, $\sigma^2(x_*)$ is given by the prior variance k_{**}, which is a positive term, minus the posterior variance $\mathbf{k}_* K^{-1}\mathbf{k}_*^T$ which is also positive. The posterior variance will be inversely proportional to the distance between the test point and the training points in the input space, as it depends on $\mathbf{k}_*$, resulting in large variances for test points that are far from training points, as shown in Fig. 2.

Other covariance functions can be employed to fit the training set, depending on the prior knowledge of the analyst about the data. Useful covariance functions include:

1. The *squared exponential* (SE) covariance function (as above) detailed in Equation (1) is one of the most commonly used covariance functions in GP modelling applications and has the general form

$$k_{SE}(r) = exp(-\frac{r^2}{2l^2}) \tag{7}$$

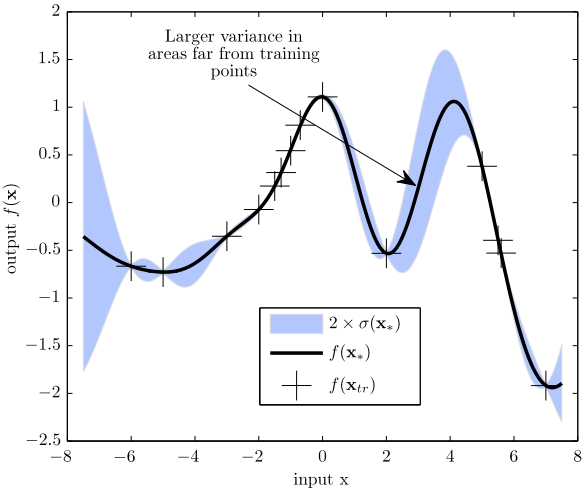

Fig. 2: Example prediction and 95% confidence intervals (2 standard deviation distance) for one dimensional GP. Note how the variance on the prediction grows with the distance from observed training points.

where $r = |\mathbf{x_i} - \mathbf{x_j}|$ and l defines the characteristic length scale. When a different length scale is specified for each dimension, input dimensions that do not contribute to the model output are automatically assigned long length scales during optimisation. This is known as automatic relevance determination (ARD). The SE covariance function is infinitely differentiable and, therefore, is very smooth.

2. The *linear* covariance function has the general form

$$k_{lin}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T P^{-1}\mathbf{x}_j \tag{8}$$

where $P \in \mathbb{R}^{d\times d}$ is a diagonal matrix of ARD parameters $p_{11}, p_{22}, \cdots p_{dd}$, where d is the dimension of the input space.

3. The *Matérn* class of covariance functions is given by

$$k_{Matern}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{l}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}r}{l}\right) \tag{9}$$

with positive parameters ν and l, where K_ν is a modified Bessel function. The Matérn covariance functions are $\nu - 1$ times differentiable. Hence the parameter ν can be used to allow very jagged outputs. As $\nu \to \infty$, $k_{Matern}(r) \to k_{SE}(r)$.

4. The *rational quadratic* (RQ) covariance function has the form

$$k_{RQ}(r) = \left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha} \tag{10}$$

with α, $l > 0$ and can be visualised as an infinite sum of squared exponential covariance functions of differing length scales. The RQ covariance function hence allows the GP to vary the length scale over the range of each input dimension. The limit of the RQ covariance for $\alpha \to \infty$ is the SE covariance function.

5. The *neural network* (NN) covariance function has the form

$$k_{NN}(\mathbf{x}_i,\mathbf{x}_j) = \sigma \sin^{-1} \dots \left(\frac{\tilde{\mathbf{x}}_i^T P^{-1} \tilde{\mathbf{x}}_j}{\sqrt{(1+\tilde{\mathbf{x}}_i^T P^{-1} \tilde{\mathbf{x}}_j)(1-\tilde{\mathbf{x}}_i^T P^{-1} \tilde{\mathbf{x}}_j)}} \right) \tag{11}$$

where $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$ vectors are augmented vectors such that $\tilde{\mathbf{x}} = (1, x_1, x_2, \cdots x_d)^T$. $P = cI$ where c is constant, and σ controls the signal variance.

Depending on the the form of the training data, covariance functions can be summed or multiplied to form suitable GP models. Different models can then be compared using the marginal likelihood of each. The choice of this covariance structure originated from prior knowledge of the target signal. A more complete treatment of the covariance functions mentioned, along with some others, can be found in [12].

GP models have several advantages over other modelling techniques. Using GPR, useful models can be created from training data sets with a relatively small number of training points and the analyst's prior beliefs about the data can be encapsulated in the GP covariance function. Because the model form is not specified explicitly, both linear and non-linear functions can be approximated. Finally, confidence intervals on predictions can be evaluated easily as each prediction is given in the form of a distribution. However, during the training procedure, GP models require the inversion of large covariance matrices, the size of which are determined by the number of training data points. This computational demand can be prohibitive in applications with large training data sets.

III Description of data set

The models examined in this report are constructed and tested on a data set collected from a multi-step trench etch process in an industrial semiconductor fabrication plant over a period of six months. The data consists of measurements collected from three different sources.

Etch process (EP) data consist of 131 variables such as temperatures, pressures, and gas flow rates for each process step collected directly from the processing tool. These EP data are reduced to a set of 28 variables by discarding variables unrelated to the main etch step and variables with little or no variance.

Number of Wafers	12133
Measured Wafers	529
PM Cycles	12
Measurement Frequency	4.4 %

Table 1: Data sets contents.

A plasma impedance monitor (PIM) sensor records an additional 159 variables for every wafer, comprising 53 harmonics each of electrode current, voltage, and phase. From these variables, it is possible to calculate the reactance (X) and resistance (R) of the chamber at each of the 52 harmonics (denoted XR data). Calculations of power and impedance are also possible. An extra collection of variables, denoted EP+ data, is formed with a combination of EP data along with power and impedance values calculated from the XR data.

Optical measurements of the etch depth, taken downstream from the etch process, are available for a limited number of wafers.

Summary statistics such as mean and standard deviation are derived from the time series traces for each variable. Wafers recorded with erroneous data are detected using a T^2 statistic and removed (see [13]). After removal of wafers with incorrectly recorded information, a total of 12133 wafers with correctly recorded EP and PIM data are available for analysis. Table 1 summarises the contents of the data set.

The data from each of the etch tools are manipulated in two separate ways. In order to test the prediction accuracy of models, first the data set is kept in chronological order, where the first 7/10 of wafers is used to train the model, and the remaining 3/10 is designated as test data to test model performance.

To test the prediction accuracy in a second way, an interleaved data set is introduced. For the interleaved data sets, the training and training sets are interleaved throughout the full set of wafers. Again the wafers are split using the ratios 7/10 and 3/10 for training and test data respectively.

IV Modelling results

Nine different input variable selections are investigated as candidate modelling inputs. Stepwise selection [14] and principal component analysis (PCA) [15] are investigated as variable selection and data reduction techniques for the input variables. The five covariance function forms described

Data Source	Covariance Function									
	Linear		**SE**		**RQ**		**Matern**		**NN**	
	R2	MAPE	R2	MAPE	R2	MAPE	R2	MAPE	R2	MAPE
EP	0.31	1.18	0.18	1.23	0.17	1.24	0.25	1.31	0.27	1.29
EP Step	0.31	1.18	**0.26**	**1.17**	0.26	1.17	0.20	1.61	0.22	1.51
PIM	0.12	1.36	0.00	3.22	0.12	1.33	0.18	1.28	0.20	1.23
PIM PCA	0.17	1.45	0.01	2.42	0.06	1.33	0.16	1.29	0.16	1.26
PIM Step	0.14	1.46	0.17	1.47	0.16	1.48	0.13	1.58	0.13	1.55
XR	0.07	1.40	0.01	2.07	0.13	1.35	0.11	1.34	0.17	1.28
XR PCA	0.16	1.32	0.08	1.39	0.07	1.31	0.06	1.38	0.10	1.34
XR Step	0.16	1.43	0.18	1.33	0.18	1.33	0.17	1.38	0.18	1.42
EP+	0.16	1.43	0.18	1.39	0.18	1.39	0.17	1.38	0.17	1.42

Table 2: GPR modelling results for chronologically ordered data.

Data Source	Covariance Function									
	Linear		**SE**		**RQ**		**Matern**		**NN**	
	R2	MAPE	R2	MAPE	R2	MAPE	R2	MAPE	R2	MAPE
EP	0.67	1.23	0.56	1.44	0.69	1.17	0.71	1.12	0.70	1.14
EP Step	0.69	1.17	0.69	1.12	0.70	1.12	0.66	1.17	0.66	1.18
PIM	0.69	1.16	0.00	2.23	0.72	1.12	0.73	1.10	0.72	1.11
PIM PCA	0.72	1.13	0.72	1.11	0.71	1.15	**0.73**	**1.10**	0.73	1.11
PIM Step	0.72	1.13	0.71	1.12	0.71	1.14	0.73	1.11	0.73	1.11
XR	0.70	1.16	0.01	2.23	0.72	1.13	0.72	1.12	0.71	1.14
XR PCA	0.68	1.19	0.67	1.24	0.65	1.26	0.71	1.14	0.70	1.16
XR Step	0.68	1.20	0.69	1.19	0.69	1.19	0.69	1.20	0.68	1.21
EP+	0.67	1.20	0.67	1.20	0.67	1.20	0.67	1.18	0.66	1.20

Table 3: GPR modelling results for interleaved data.

in II are investigated as potential covariance functions where the multidimensional ARD version of each covariance function is used, along with constant and noise terms. The results of the analysis are presented in Tables 2 and 3 for chronological and global modelling respectively. Data sources labelled "PCA" are first translated to a principal component space before modelling. Stepwise regression is applied to choose the most relevant input variables for data sources marked "Step". Models performance is reported in terms of their mean absolute percentage error (MAPE) and coefficient of determination (R^2) on unseen data.

The best model performance for the chronological data set is achieved using stepwise selected EP data and the SE covariance function. The best model performance for the interleaved data is achieved using a Matérn covariance function with the PCA reduced PIM data as input variables.

In general, the performance for the interleaved data is more accurate than for the chronological data. For the interleaved data set, the training data points contain information from the same PM cycles as the test data points. For chronological data investigations, models are expected to predict etch rate in completely unseen PM cycles, where the system behaviour may have changed.

The results suggest that model performance for the etch rate data is relatively insensitive to covariance function choice, with the exception of the SE covariance function that fails with input sets with a relatively large number of variables (PIM and XR inputs). Further tests with combinations of different covariance functions used together do not yield any substantial further improvement in prediction accuracy. Best results from modelling tests with neural networks, a commonly applied modelling technique for such tasks, fail to better the results of Tables 2 and 3, yielding MAPEs of 1.37% and 1.21% for chronological and interleaved data respectively (see [16] for neural network implementation details).

The simple evaluation of confidence intervals (CIs) on the etch rate predictions is useful in plasma etch monitoring as it can provide a measure of model accuracy, and can be used by engineers to maintain etch rate within specification. While models cannot follow the high frequency fluctuations of the real etch rate value, these variations are captured within the CIs. These variations may arise due to external disturbances in the manufacturing process not reflected in the collected data set. Fig. 3 shows the etch rate predictions and CIs from an GPR model using interleaved PIM data as input variables.

V Conclusions

Gaussian process regression models have been applied to semiconductor etch data and shown to accurately predict real etch rate using process data.

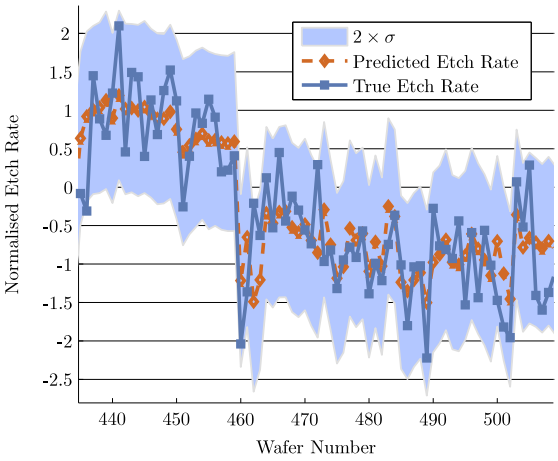

Fig. 3: Normalised etch rate prediction with confidence intervals.

The modelling results have shown that the prediction accuracy is relatively insensitive to covariance function choice for this data set. For analysts, this insensitivity simplifies model setup and allows a quick modelling time, without needing extensive investigation into different covariance functions.

Predictions are given in the form of distributions, allowing simple calculations of confidence intervals on predicted values. These confidence intervals are useful to engineers, allowing an estimation of the degree of variation possible on each predicted value.

The prediction performance and ease of CI calculation means that GPR models have the capability to gain increased popularity in industry in the near future, competing with ANNs and other statistical methods as a virtual metrology technique.

References

[1] F. F. Chen and J. P. Chang, *Lecture Notes on Principles of Plasma Processing*. Kluwer Academic/Plenum Publishers, 2003.

[2] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, 1965.

[3] A. Khan, D. Tilbury, and J. Moyne, "Fab-wide virtual metrology and feedback control," in *Asian AEC/APC Symposium*. NSF Engineering Research Center for reconfigurable Manufacturing Systems, University of Michigan., 2006.

[4] Y. Yang, M. Wang, and M. J. Kushner, "Progress, opportunities and challenges in modeling of plasma etching," in *Int. Interconnect Tech. Conf. (IITC)*, Jun. 2008, pp. 90–92.

[5] J. V. Ringwood, S. Lynn, G. Bacelli, B. Ma, E. Ragnoli, and S. McLoone, "Estimation and control in semiconductor etch: Practice and possibilities," *IEEE Trans. Semicond. Manuf.*, vol. 23, no. 1, pp. 87–98, Feb. 2010.

[6] S. Lynn, J. Ringwood, E. Ragnoli, S. McLoone, and N. MacGearailt, "Virtual metrology for plasma etch using tool variables," in *Proc. Adv. Semicond. Manuf. Conf. (ASMC)*, May. 2009, pp. 143–148.

[7] S. Lynn, J. V. Ringwood, and N. MacGearailt, "Weighted windowed PLS models for virtual metrology of an industrial plasma etch process," in *IEEE Int. Conf. Indust. Tech.*, Valpariaso, Chile, Mar. 2010, pp. 271–276.

[8] E. Ragnoli, S. McLoone, S. Lynn, J. Ringwood, and N. Macgearailt, "Identifying key process characteristics and predicting etch rate from high-dimension datasets," in *Proc. Adv. Semicond. Manuf. Conf. (ASMC)*, May. 2009, pp. 106–111.

[9] A. Khan, J. Moyne, and D. Tilbury, "An approach for factory-wide control utilizing virtual metrology," *IEEE Trans. Semicond. Manuf.*, vol. 20, no. 4, pp. 364–375, Nov. 2007.

[10] T.-H. Lin, F.-T. Cheng, W.-M. Wu, C.-A. Kao, A.-J. Ye, and F.-C. Chang, "NN-based key-variable selection method for enhancing virtual metrology accuracy," *IEEE Trans. Semicond. Manuf.*, vol. 22, no. 1, pp. 204–211, Feb. 2009.

[11] J. Kocijan, "Gaussian process models for systems identification," in *Proc. 9th Int. PhD Workshop on Sys. and Cont.*, 2008.

[12] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambride, MA.: MIT Press, 2006.

[13] S. Lynn, "Local modelling of a plasma etch data set," Elec. Eng. Dept., NUI Maynooth, Ireland, Ireland, Tech. Rep. EE/JVR/1/2010, February 2010.

[14] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 4th ed. New York: John Wiley & Sons Inc., 2001.

[15] J. E. Jackson, *A User's guide to principal components*. Wiley Interscience, 1991.

[16] S. Lynn, "Global modlelling of a plasma-etch data set," Elec. Eng. Dept., NUI Maynooth, Ireland, Tech. Rep. EE./JVR/3/2009, 2009.