A Modeling Toolbox and Its Applications to Statistical Process Modeling

Young-Don Ko and Helen Shang[1]
School of Engineering, Laurentian University, Sudbury, Ontario, Canada P3E 2C6
Email: {yko, hshang}@laurentian.ca

***Abstract*— Analyzing data that are measured or collected in processes requires statistical software. Most methods for analysis use a regression approach or its expanded forms such as polynomial and response surface models. However, the available commercial programs are inconvenient and costly for personal use, and require demanding pre-training in using them. A simple and user-friendly interface program adaptable to personal needs is therefore demanded. In this paper, a Matlab Toolbox, named the Regression Modeler (RM), is developed. The structure of the program and procedure of implementation are described. The program is verified using various datasets in process engineering. The toolbox provides users with solutions to regression models, polynomial models and response surface models with fine 2D and 3D plots.**

***Index Terms*— matlab toolbox, regression model, multiple regression model, polynomial model, response surface model.**

I. Introduction

With the development of system engineering in recent years, the volume of data containing valuable information has multiplied. The data can be used for diagnosis, decision-making and problem-solving. It is necessary to get reliable results from the data as they are directly related to productivity and cost in the industry. The analysis technique is therefore very important in examining the correlations of various factors and their effects on the systems. Specifically, the slope of a regression model obtained from a given dataset, whether positive or negative, accounts for the response direction of the system. For a response surface model, a prediction model including the prediction rate and a surface plot in three dimensions (3D) are needed to interpret the physical mechanisms or trends when process variables are controlled, changed, and shifted. In many applications, regression and quadratic models are recommended except when high-order nonlinear models and hybrid expert systems are needed.

There are many commercial programs or analysis packages available that use the methodologies described above. But most of them are expensive for limited analyses like solving regression problems or building quadratic models. Further, use of these programs requires training, coding, and skills such as knowledge of functions, parameter-setting, and code modification [1-2]. Available commercial packages, e.g., Minitab, Statistica, SAS and SPSS, can provide us with the prediction models along with various plots. Efforts have been made to develop Matlab toolboxes adapted to the needs of the users in some special disciplines. To analyze the sequence data for molecular biology and evolution, Cai *et al.* developed the MBE Toolbox [3]. A toolbox has been developed in the computer science for interactive development, validation, execution of distributed and parallel processing, data processing and visualization [4-5]. Ye *et al.* developed a GUI-based Matlab Toolbox to analyze hyperspectral images with a remote sensing methodology in the vegetation research [6]. For geological and geophysical data, Witten presented a three-dimensional viewer that was coded based on a GUI-Matlab environment [7]. A modeling toolbox based on a modified long-wavelength approximation was developed to describe the optical properties of single metal nanoparticles [8]. In addition, the toolboxes have been used for measuring the functional connectivity in brain image research [9], the biodata toolbox was applied to analyze spectral data including all relevant information regarding the spectra [10]. In the statistical analysis of time series with Granger causal connectivity [11], toolboxes have been developed for micromechanical analysis of composite materials with finite element analysis [12], for the simulation of metallic nanoparticle using a boundary element method [13], and plotting tool for the analysis of comparative genomic hybridization (CGH) microarray data [14]. As has been proved and addressed in many fields with various needs, the applications of Matlab Toolboxes are not limited to special fields and are widely utilized from data mining to medical analysis. They are applicable to various disciplines and fields, from social science, mathematics, computer science, and geology to nanotechnology and bioinformatics.

In this study, the Regression Modeler (RM) Matlab toolbox is developed to solve regression, multiple regression, response surface, and polynomial models for general engineering applications. Using the RM toolbox, the model results can be checked quickly online or offline. The developed toolbox can quickly generate a simple regression model, a multiple regression model, a response

[1] Corresponding author. Email: hshang@laurentian.ca

doi:10.4304/jsw.8.11.2706-2710

surface model (quadratic model), and a polynomial regression of second to fourth order. It also includes the response surface plots of two variables, model coefficients, and R-square values. However, a nonlinear model, including that of high-order, is not considered in the developed toolbox.

II. Toolbox Dscription

The hierarchical structure of the Regression Modeler Matlab toolbox is illustrated in Fig. 1. As shown, the toolbox consists of five categories. Among them, “Model” and “Results” have four and three submenus, respectively. The four specified regression models in “Model” include: “Linear regression,” “Multiple linear regression,” “Polynomial regression,” and “Response surface model.” “Results” is for displaying the model coefficients, the R-squared value, the 2D plot in linear and log scales, and the 3D surface plot. If a model is not applicable to either 2D or 3D plot representation, a notification “Not Application” is displayed above the inactive figure in red and unnecessary figures are not plotted. For example, there are no 2D plots for multiple regression or response surface cases and no 3D plots for the cases where simple linear regression and polynomial models are applied.

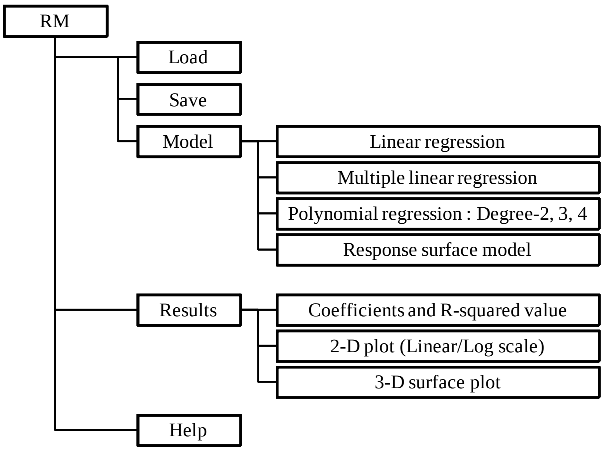

RM
Load
Save
Model
Linear regression
Multiple linear regression
Polynomial regression : Degree-2, 3, 4
Response surface model
Results
Coefficients and R-squared value
2-D plot (Linear/Log scale)
3-D surface plot
Help

Figure 1. Hierarchical structure of the RM Matlab toolbox.

The main active window of the RM toolbox is shown in Fig. 2 (a) and 2 (b). In Fig. 2 (a), the window has two boxes for “Command” and “Result.” In the “Command” window, we can specify the file location to load the data and save the results. The required model can be selected by the dropdown menu shown in Fig. 2 (b). Results containing the predicted and measured outputs can be saved as a text file in a user’s computer. Fig. 2 (c) shows the tutorials available in the “Help” menu in the toolbox. As described in the tutorial, four simple steps are necessary to run RM:

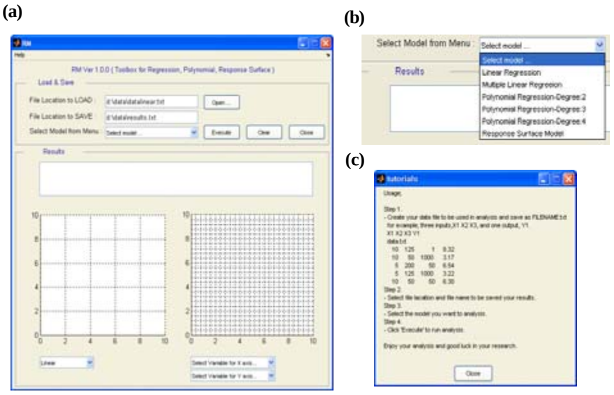
1) Create and load the data file to be used in RM in the format of .txt.

2) Select the file location to save the result.

3) Select the model you want to analyze: simple regression, multiple regression, polynomial model with second, third, and forth orders, response surface methodology.

4) Run RM.

A computer with an Intel Pentium Dual Core CPU 1.8 GHz, and with 2 GB RAM at 800 MHz FSB, was used as the development system, and Matlab Ver. 7.5.0.342 (R2007b) was used as the programming language. The RM toolbox was developed for both the Matlab-installed environment and for that without Matlab installed. For the software users without Matlab, the stand-alone executable files are available, along with the MCR Installer (Matlab Complier Runtime Installer).


(a)
RM
Help
RM Ver 1.0.0 (Toolbox for Regression, Polynomial, Response Surface)
Load & Save
Open ...
Execute
Clear
Close
Results
10
8
6
4
2
0
0 2 4 6 8 10
Linear
Select Variable for X axis ...
Select Variable for Y axis ...
(b)
Results
Select model ...
Linear Regression
Multiple Linear Regresion
Polynomial Regression-Degree:2
Polynomial Regression-Degree:3
Polynomial Regression-Degree:4
Response Surface Model
(c)
tutorials
Close

File Location to LOAD : d:\data\datalinear.txt

File Location to SAVE : d:\data\results.txt

Select Model from Menu : Select model ...

```
Usage;

Step 1.
- Create your data file to be used in analysis and save as FILENAME.txt
 for example, three inputs,X1 X2 X3, and one output, Y1
 X1 X2 X3 Y1
 data.txt
   10   125      1   8.32
   10    50   1000   3.17
    5   200     50   6.54
    5   125   1000   3.22
   10    50     50   6.30
Step 2.
- Select file location and file name to be saved your results.
Step 3.
- Select the model you want to analysis.
Step 4.
- Click 'Execute' to run analysis.

Enjoy your analysis and good luck in your research.
```

Figure 2. The windows of RM Matlab toolbox: (a) the main window, (b) the model selection in menu, and (c) the tutorials.

III. Validation of the RM Toolbox in Engineering Applications

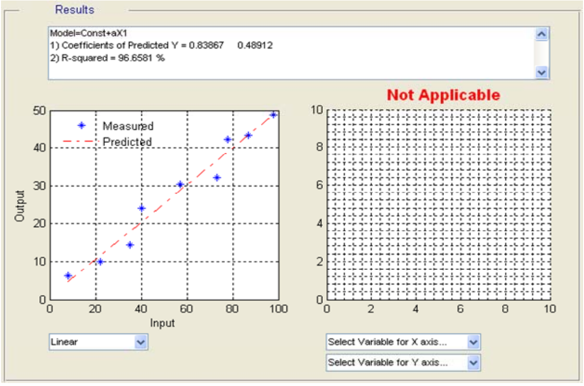
To evaluate the software’s performance, the Regression Modeler Toolbox was applied to a variety of datasets in engineering applications. Three datasets are used in this section for illustration and validation. Two are related to the characterization of etch rates and cell gaps in the thin-film manufacturing process in the semiconductor engineering [15-17]. The third dataset is about particle size distribution in the mineral engineering and the particle sizes of ore from mines were calculated using WipFlag© [18].

In the regression analysis, the semiconductor data for etch rates were employed to build a simple regression model. The data consists of input (marked *x*) and output (marked *y*). The input is the time and the output is the film thickness [15]. The relationship between the input and the output is explored, i.e., how the film thickness would be affected by the processing time. The etching process has been analyzed and modeled using different approaches along with its physical phenomenon [19-21]. The results of a simple regression analysis are shown in Fig. 3. A simple regression model is obtained as follows

$$y = 0.8367 + 0.48912 x_1 \tag{1}$$

This model has 96.66 percent of the R-squared value accounting for how well the model fits with the measurement data. The R-squared value can be calculated as a ratio of the regression sum of squares (SSR) and total sum of squares (SSTO), as redefined in [22], where SSTO = SSR + the error sum of squares (SSE). It represents the portion of variation in the response explained by the model. Thus, the R squared value can be expressed as:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{2}$$



Model=Const+aX1
1) Coefficients of Predicted Y = 0.83867 0.48912
2) R-squared = 96.6581 %

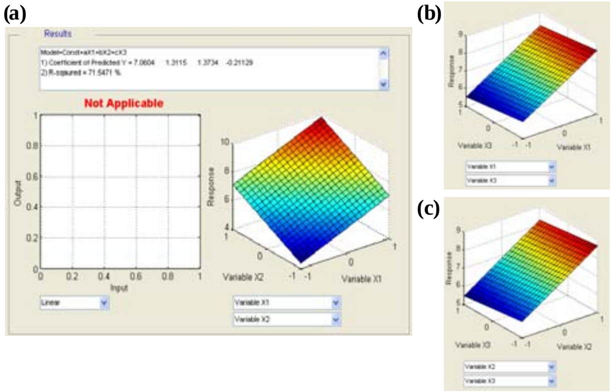
Figure 3. Results of a simple linear regression

The predicted and measured values are plotted in Fig. 3. As the model has a high accuracy with 96.66 percent, the predicted data points lie around the straight line. The RM toolbox can also provide a different view of the 2D plot in log scale, but it is unnecessary in this case for a simple regression. Note that the 2D plot in log scale is only applicable to both simple regression and polynomial models. Fig. 3 indicates that there is a strong linear relationship between the input and output, hence the simple regression can generate a well-fit model for the etching process. It is observed that the etch rate increases with the time linearly.

The data for cell gaps were also from the semiconductor research and are used to validate both the multiple regression model and the response surface model. The experimental setup and physical considerations can be found in [16]. The inputs were generated from the design of the experiments and are used for building both multiple regression and response surface models. In the design of the experiments, *D*-optimal design can be used to generate the experimental runs. This is a very effective way to minimize the variance of model parameters and select a flexible number of experiments, leading to the response surface models of second order. The inputs are coded values (-1, 0, and 1). Fig. 4 and Fig. 5 present the results of a multiple regression and a surface response model, respectively. It is noted that the response surface model shows better results than the multiple regression in predicting the cell-gap manufacturing process with a higher R-squared value. The R-squared value of the surface response model is 89.01 percent, implying that a curvature effect exists and it could affect the manufacturing process. This example shows that the response surface model may be a better choice for cases in which curvature effects are considered. The 3D surface plots enable us to examine the physical changes with varying process input, as illustrated in Fig. 4 (b)-(c) and Fig. 5 (b)-(c). The surface response model in this study is defined as follows:

$$y = b_0 + \sum_{i=1}^{n} b_i x_i + \sum_{j=i+1}^{n} \sum_{i=1}^{n} b_{ij} x_i x_j + \sum_{i=1}^{n} b_{ij} x_i^2 \tag{3}$$

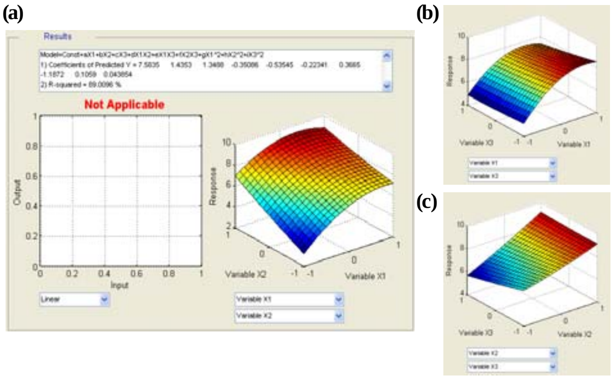
where y is the response variable, n is the number of independent process factor b's are model coefficients, and x_i 's are independent variables.



Model=Const+aX1+bX2+cX3
1) Coefficient of Predicted Y = 7.0604 1.3115 1.3734 -0.21129
2) R-squared = 71.5471 %

Figure 4. Results of a multiple linear regression: (a) the main result window, (b) X1 vs. X2 in 3D plot, and (c) X2 vs. X3 in 3D plot.

The polynomial analysis was validated using the data of particle-size distribution. As particle sizes are typically distributed in the polynomial pattern, they are examined using polynomial fitting in this section. As shown in Fig. 6-8, the polynomials from the second to fourth order are



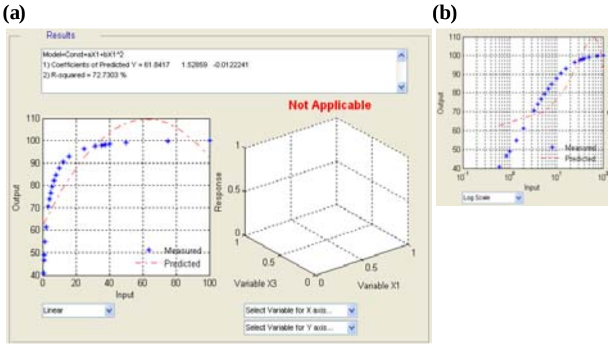
Model=Const+aX1+bX2+cX3+dX1X2+eX1X3+fX2X3+gX1^2+hX2^2+iX3^2
1) Coefficients of Predicted Y = 7.5835 1.4353 1.3488 -0.35086 -0.53545 -0.22341 0.3685 -1.1872 0.1059 0.043854
2) R-squared = 89.0096 %

Figure 5. Results of a response surface model: (a) the main result window, (b) X1 vs. X2 in 3D plot, and (c) X2 vs. X3 in 3D plot.

considered in building the models. From the results from the RM toolbox, it is noted that the R-square values of the second and third order are 72.73 percent and 84.98 percent, and that of the fourth order is 94.08 percent.

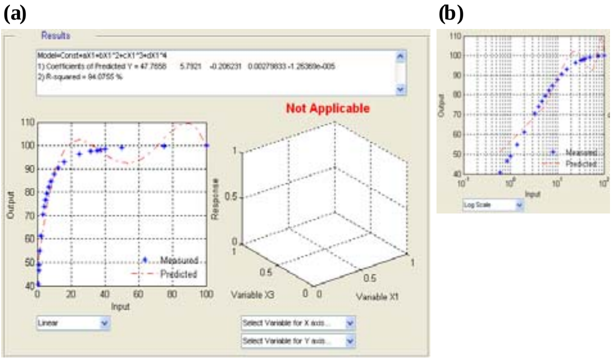
It can be said that the fourth-order polynomial model provides a better predictive model and its model is defined as

$$y = 47.7658 + 5.792x_1 - 0.206231x_1^2 + 0.00279833x_1^3 - 1.26369 \times 10^{-5} x_1^4 \quad (4)$$



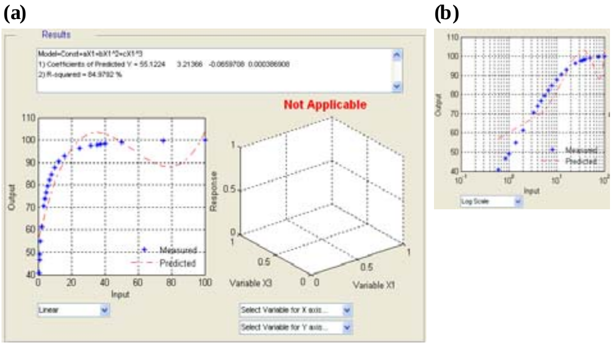
Model=Const+aX1+bX1^2
1) Coefficients of Predicted Y = 61.8417 1.52859 -0.0122241
2) R-squared = 72.7303 %

Figure 6. Results of a second order polynomial regression: (a) the main result window and (b) 2D plot in log scale.



Model=Const+aX1+bX1^2+cX1^3+dX1^4
1) Coefficients of Predicted Y = 47.7658 5.7921 -0.206231 0.00279833 -1.26369e-005
2) R-squared = 94.0755 %

Figure 8. Results of a fourth order polynomial regression: (a) the main result window and (b) 2D plot in log scale.



Model=Const+aX1+bX1^2+cX1^3
1) Coefficients of Predicted Y = 55.1224 3.21366 -0.0659708 0.000386908
2) R-squared = 84.9792 %

Figure 7. Results of a third order polynomial regression: (a) the main result window and (b) 2D plot in log scale

The RM toolbox can be applied to a variety of data and generates regression, quadratic and polynomial models. It provides a simple and user-friendly interface including 2D or 3D plotting for easy analysis. The toolbox allows the researchers and industrial practitioners to get the needed results quickly before proceeding with advanced nonlinear computational problems, and it therefore can save time and reduce cost.

As demonstrated, the developed tool box provides the models that are explained by various types of plots and can generate a wide range of modeling forms from simple regression to poly-nominal and quadratic forms. This toolbox does not support the multi-dimensional models for complex systems and future work is, therefore, needed. The future work will deal with nonlinear models in which the multimodal variables exist and many fluctuations happens in process system. If necessary, the neural networks, time-dependent models, multivariate models, and special types of modeling forms depending on the data patterns, will be used in the future work. In multivariate models, parameter selection methods, cross validation, re-sampling techniques will be addressed and added to toolbox. The hybrid modeling approach will also be investigated.

Matlab toolboxes have proven to be an important tool for evaluating and developing solutions to the problems of prediction models. This will enable us to generate the results quickly from the industrial measurement and to help come up with solutions to satisfy the industrial needs.

IV. CONCLUTIONS

In this paper, a Matlab toolbox, named Regression Modeler, is developed to solve regression, response surface, and polynomial modeling problems. The toolbox was implemented and validated using the data from a variety of engineering problems. The results indicate that the RM toolbox provides a convenient approach to obtaining a quick result needed for analysis, including for the cases where problems are not amenable to regression, quadratic, and polynomial solutions. The toolbox allows users to have a simple and fast analysis on the spot, saving both time and the cost of using the existing commercial software.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by Centre for Excellence in Mining Innovation (CEMI) and the Ministry of Research and Innovation (MRI) Postdoctoral Fellowship (PDF) Program.

REFERENCES

[1] I. M. Baskir and A.V. Drozd, “New Matlab software for wavelength selection,” *Chemometr. Intell. Lab*, vol. 66, no. 1, pp. 89-91, 2003.

[2] M. Daszykowski, S. Serneels, K. Kaczmarek, P. V. Espen, C. Croux, and B. Walcazk, “TOMCAT: A MATLAB toolbox for multivariate calibration techniques,” *Chemometr. Intell. Lab*, vol. 85, no. 2, pp. 269-277, 2007.

[3] J. J. Cai, D. K. Smith, X. Xia, and K.-y. Yuen, “MBEToolbox: a Matlab toolbox for sequence data analysis in molecular biology and evolution,” *BMC Bioinformatics*, 6, 2005.

[4] S. Pawletta, W. Drewelow, P. Duenow, T. Pawletta, and M. Suesse, “A MATLAB Toolbox for Distributed and Parallel Processing,” *Matlab Conference 95*, Cambridge, MA., 1995.

[5] H. Teufelsbauer, “Linking laser scanning to snowpack modeling: Data processing and visualization,” *Comput. Geosci.*, vol. 35, no. 7, pp. 1481-1490, 2009.

[6] X. Ye, K. Sakai, H. Okamoto, and L. O. Garciano, “A groung-based hyperspectral imaging system for characterizing vegetation spectral features,” *Comput. Electron. Agr.*, vol. 63, no. 1, pp. 13-21, 2008.

[7] A. Witten, “A MATLAB-based three-dimensional viewer,” *Comput. Geosci.*, vol. 30, no. 7, pp. 693-703, 2004.

[8] K. D. Ko and K. C. Toussaint Jr., “A simple GUI for modeling the optical properties of single metal nanoparticles,” *J. Quant. Spectrosc. Ra.*, vol. 110, no. 12, pp. 1037-1043, 2009.

[9] D. Zhou, W. K. Thompson, and G. Siegle, "MATLAB toolbox for functional connectivity," *NeuroImage*, vol. 47, no. 1, pp. 1590-1607, 2009.

[10] K. D. Gussen, J. D. Gelder, P. Vandenabeele, and L. Moens, "The Biodata toolbox for MATLAB," *Chemometr. Intell. Lab*, vol. 95, no. 1, pp. 49-52, 2009.

[11] A. K. Seth, "A MATLAB toolbox for Granger causal connectivity analysis," *J. Neurosci. Methods*, vol. 196, no. 2, pp. 262-273, 2010.

[12] U. Hohenester, A. Trügler, "MNPBEM-A Matlab toolbox for the simulatin of plasmonic nanoparticles," Comput Phys Commun, vol. 183, no. 2, pp. 370-381.

[13] C. T. McCarthy and T. J. Vaughan, "A MATLAB toolbox for micromechanical analysis of composite materials," J Compos Mater, vol. 46, no. 14, pp. 1715-1729, 2011.

[14] R. Autio, S. Hautaniemi, P. Kauraniemi, O. Yli-Harga, J. Astola, M. Wolf, and A. Kallioniemi, "CGH-Plotter: MATLAB toolbox for CGH-data anaysis," Bioinformatics, vol. 19, no. 13, pp. 1714-1715, 2003.

[15] G. S. May and C. J. Spanos, *Fundamentals of Semiconductor Manufacturing and Process Control*, Wiley & Sons: New Jersey, USA, 2006.

[16] Y.-D. Ko, J.-Y. Hwang, D.-S. Seo and I. Yun, “Investigation of Cell Gap on the Polymer Substrates using Statistical Modelling for Flexible Liquid Crystal Display Applications,” *Int. J. Nanomanuf.,* vol. 2, no. 4, pp. 361-374, 2008.

[17] G. S. May and S. M. Sze, *Fundamentals of Semiconductor Fabrication*, Wiley & Sons: New Jersey, USA, 2004.

[18] *WipFrag User’s Manual*, WipWare Inc.

[19] Y.-D. Ko, Y. Jeong, M.-K. Jeong, A. Garcia-Diaz, and B. Kim, ”Functional Kernel-based Modeling of Wavelet Compressed Optical Emission Spectral Data: Prediction of Plasma Etch Process,” *IEEE Sens. J.*, vol. 10, no. 3, pp. 746-754, 2010.

[20] B. Kim, J. Kim, and S. Choi, “Use of neural network to model X-ray photoelectron spectroscopy data for diagnosis of plasma etch equipment,” *Expert. Syst. Appl.*, vol. 36, no. 8, pp. 11347-11351, 2009.

[21] B. Kim and M. Kwon, “Prediction of plasma etch process by using actinometry-based optical emission spectroscopy data and neural network,” *J. Mater. Process. Tech.*, vol. 209, no. 5, pp. 2620-2626, 2009.

[22] Y.-D. Ko, P. Moon, C. E. Kim, M.-H. Ham, J.-M. Myoung and I. Yun, “Modeling and Optimization of the Growth Rate for ZnO Thin Films using Neural Networks and Genetic Algorithms,” *Expert. Syst. Appl.*, vol. 36, no. 2, pp. 4061-4066, 2009.

Young-Don Ko received his Ph.D. in Electrical and Electronic Engineering, Yonsei University, Seoul, Korea in 2007. He is currently a Postdoctoral Fellow, Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital, and Institute of Biomaterials and Biomedical Engineering, University of Toronto, Canada, and was a Postdoctoral Fellow at, University of Alberta, Laurentian University, and the University of Tennessee, Knoxville. His research interests include modeling, diagnosis, statistical analysis, and optimization of semiconductor engineering, mineral processing, and biomedical engineering. He received Postdoctoral Fellowship, National Research Foundation, Korea and Ministry of Research and Innovation, ON, Canada in 2008 and 2010.

Helen Shang studied Chemical Engineering in Tsinghua University, China where she received her BASc. and MASc.. For her PhD, in the Department of Chemical Engineering, University of Alberta, Canada, she carried out research on advanced control methods for complicated systems. Since 2002, she has been teaching process control and other engineering courses at Laurentian University, where she has also conducted significant research in both theoretical and applied process modelling and control. She is also a past chair of the System and Control Division of the Canadian Society of Chemical Engineering (CSChE).