Proceedings of International Conference on Heat Exchanger Fouling and Cleaning VIII - 2009 (Peer-reviewed)
June 14-19, 2009, Schladming, Austria
Editors: H. Müller-Steinhagen, M.R. Malayeri and A.P. Watkinson

Published online
www.heatexchanger-fouling.com

CONSTRUCTION OF HIGH PREDICTIVE FOULING MODELS USING STATISTICAL METHODS

Hiromasa Kaneko[1], Susumu Inasawa[1], Hirofumi Inokuchi[2], Kimito Funatsu[1]

[1] Department of Chemical System Engineering, Graduate School of Engineering,
The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-8656, Japan
E-mail (funatsu@chemsys.t.u-tokyo.ac.jp):
[2] Mitsubishi Chemical Corporation, 3-10, Ushiodori, Kurashiki, Okayama, 712-8054, Japan

ABSTRACT

Fouling is one of the important phenomena which should be understood in many fields. We constructed statistical models with which we were able to predict fouling phenomena. Input values are experimentally obtained parameters and some characteristic values such as the degree of supersaturation. Fouling parameters such as thermal resistance R_f are calculated by our models and the results are in good agreement with experimental results. Therefore, we concluded that our models predicted the fouling phenomena with high accuracy. Those models are constructed with linear and nonlinear regression methods. Furthermore, our statistical approaches are practical and successful in understanding fouling.

INTRODUCTION

Thermal resistance R_f and precipitation rate constant are very important parameters for studying fouling phenomena. However, in a real process, it is difficult to estimate those parameters with high accuracy because there is no predictive model that describes fouling phenomena at a satisfactory level. One of the many possible reasons is that we do not have well-established physical models to describe fouling phenomena. Although we have a lot of fouling data obtained in industrial plants and lab-scale experiments, lacking of physical models restricts us from analyzing these data scientifically, hence resulting in insufficient understanding of fouling. Since there is a lack of physical understanding, statistical approaches becomes an interesting and promising alternative. It is possible to extract some physically important parameters from experimental data and then, once we extract relationships between experimental conditions and resulting fouling phenomena, a statistically predictive model can be constructed. We use "chemoinformatics" (Gasteiger and Engel, 2003) in our works to study the fouling phenomena.

Chemoinformatics is a generic term used by many fields of chemistry. It is a field in which problems of chemistry are solved using informatics methods. There are many researches in the field, revolving around topics such as quantitative structural-property relationships, quantitative structure-activity relationships, reaction design and drug design.

In our study, we construct fouling models based on chemoinformatics methods that predict values of objective variables such as R_f from experimentally obtained parameters. We use a partial least squares (PLS) (Wold et al., 2001) method and a support vector regression (SVR) (Vapnik et al. 1995) method as linear and nonlinear regression methods. By using experimental data at various conditions, it was confirmed that PLS models with high accuracy were constructed and a SVR method improved predictive accuracy of our models.

In addition to experimentally obtained parameters, we improved predictive accuracy by considering characteristic values such as the degree of supersaturation. These values were extracted from experimental data using a set of heat transfer equations. Reconstructed models with these characteristic values predict fouling parameters with higher accuracy. By using our models, it is able to estimate fouling variables in a new experimental condition with high predictive accuracy.

PLS

PLS is a method for relating $\mathbf{X} \in R^d$ (where d is the number of variables) and $\mathbf{y} \in R$, by a linear multivariate model, but goes beyond traditional regression methods in that it models also the structures of $\mathbf{X}$ and $\mathbf{y}$. In PLS modeling, the covariance between score vector $\mathbf{t}_i \in R$ and $\mathbf{y}$ is maximized. Generally, PLS models have higher predictive power than those of multiple linear regression.

A PLS model consists of two equations as follows:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \quad (1)$$

$$\mathbf{y} = \mathbf{Tq} + \mathbf{f} \quad (2)$$

where $\mathbf{T} \in R^a$ (where a is the number of components) is the score matrix; $\mathbf{P} \in R^a$ is an $\mathbf{X}$-loading matrix; $\mathbf{q} \in R$ is a $\mathbf{y}$-loading vector; $\mathbf{E} \in R^d$ is a matrix of $\mathbf{X}$ residuals and $\mathbf{f} \in R$ is a vector of $\mathbf{y}$ residuals. The PLS-regression model is as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{const} \quad (3)$$

$$\mathbf{b} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{q} \quad (4)$$

where $\mathbf{W} \in R^a$ is an $\mathbf{X}$-weight matrix and $\mathbf{b} \in R$ is a vector of regression coefficients. The number of components must be appropriately decided to construct a highly predictive

model. r^2 and q^2 values are used as the measure and defined as follows:

$$r^2 = 1 - \frac{\sum (y_{obs} - y_{calc})^2}{\sum (y_{obs} - \bar{y})^2} \quad (5)$$

$$q^2 = 1 - \frac{\sum (y_{obs} - y_{pred})^2}{\sum (y_{obs} - \bar{y})^2} \quad (6)$$

where y_{obs} is the actual **y** value; y_{calc} is the calculated **y** value and y_{pred} is the predicted **y** value in the procedure of cross-validation such as leave-one-out. In this paper, the number of components is determined by maximum of q^2.

SVR

SVR is a method applying support vector machine (SVM) to a regression analysis and can construct non-linear models by applying the kernel trick as well as SVM. Primal form of SVR can be shown to be a following optimization problem:

Minimize

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i |y_i - f(\mathbf{x}_i)|_e \quad (7)$$

subject to

$$|y_i - f(\mathbf{x}_i)|_e = \max(0, |y_i - f(\mathbf{x}_i)| - e) \quad (8)$$

where y_i and $\mathbf{x}_i$, $\mathbf{w} \in R$ are training data; $\mathbf{w} \in R$ is a weight vector; e is a threshold and C is a penalizing factor which controls a trade-off between a training error and a margin. By minimization of eq. (7), we can construct a regression model which has a well balance between adaptive ability to the training data and generalization capability. A kernel function in our application is a radial basis function:

$$K(x, x') = e^{-g*|x-x'|^2} \quad (9)$$

where g is a tuning parameter controlling width of the kernel function. In this paper, LIBSVM (Chang et al., 2001) is used as a machine learning software.

DATA

In this study, we analyzed fouling data measured by a concentric cylinder viscometer. Explanatory variables **X** are experimentally obtained parameters and characteristic values, which are extracted from experimental data. The formers are shear rate [s^{-1}], slurry concentration at the end of cooling [wt%], cooling time constant [s] and degree of supersaturation at the onset of precipitation [-], and the latters are precipitation time constant [s], maximum degree of supersaturation [-]. With these variables, values of objective variables, R_f [m^2 K W^{-1}] and viscosity [mPa s], are predicted. In addition, logarithmic transformation was used to viscosity. All variables were transformed to zero mean and unit variance as a preprocessing. The number of samples is 38. The details of experimental setups and obtained data were introduced in the talk by Inasawa *et al.* in the Eurotherm conference on Heat Exchanger Fouling and Cleaning (2009).

RESULTS AND DISCUSSION

Table 1 shows modeling results of R_f and viscosity. RMSE (root mean square error) is defined as follows:

$$RMSE = \sqrt{\frac{\sum (y_{obs} - y_{calc,pred})^2}{n}} \quad (10)$$

where $y_{calc,pred}$ is the calculated or predicted y value and n is the number of samples.

In PLS modeling, predictive accuracy increased by using characteristic values in addition to experimentally obtained parameters. A piece of information extracted from experimental data would closely relate to R_f. Fig. 1 shows standard regression coefficients of R_f models in PLS modeling. From Fig. 1(a), shear rates contributed negatively, and slurry concentrations at the end of cooling, cooling time constants and degree of supersaturation at the onset of precipitation contributed positively to R_f. From Fig. 1(b), experimentally obtained parameters had the same trend as in Fig. 1(a), and precipitation time constants and degree of supersaturation contributed positively to R_f. Because these contributions were not conflicted, we can say that we constructed appropriate models predicting values of R_f. Then, in SVR modeling, predictive accuracy decreased by using characteristic values in addition to experimentally obtained parameters. Observation errors and errors of calculation of characteristic values could affect the accuracy in the SVR model. Fig. 2 shows the relationship between measured and predicted values of R_f. In Fig. 2, predicted values were obtained from the SVR model which was constructed from only experimentally obtained parameters. The plot shows an almost linear trend along the diagonal, reflecting the high prediction accuracy in R_f.

We also constructed prediction models of viscosity. The results are shown in Table 1. In PLS modeling, predictive accuracy increased when both characteristic values and experimentally obtained parameters are used. A piece of information extracted from experimental data would also closely relate to viscosity. Fig. 3 shows standard regression coefficients of viscosity models in PLS modeling. From Fig. 3, all explanatory variables had the same trend as in Fig. 1. Because these contributions were not conflicted, we could construct appropriate models which predict values of viscosity. Then, in SVR modeling, predictive accuracy increased when both characteristic values and experimentally obtained parameters. High predictive model of viscosity was also constructed by using only experimentally obtained parameters.

If RMSE is smaller than experimental errors, regression models could be constructed, but in that case, the models fit even the experimental errors. By comparing RMSE with these errors, an appropriate model should be selected.

CONCLUSIONS

A practical statistic approach has been proposed in this paper. First the key parameters such as thermal resistance values R_f and viscosity were obtained via regression using experimental results. Then PLS and SVR are used as linear and nonlinear regression methodologies to obtain the necessary parameters for the fouling models. This model has

been tested and validated at our laboratory and successful results have been obtained.

REFERENCES

C. C. Chang, C. J. Lin, 2001, LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

J. Gasteiger and T. Engel, 2003, *Chemoinformatics-A Textbook,* Wiley-VCH, Weinheim, Germany.

S. Wold, M. Sjöström, L. Eriksson, 2001, PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, vol. 58, pp.109-130.

V. N. Vapnik, 1995, The Nature of Statistical Learning Theory; Springer, Berlin, Germany.

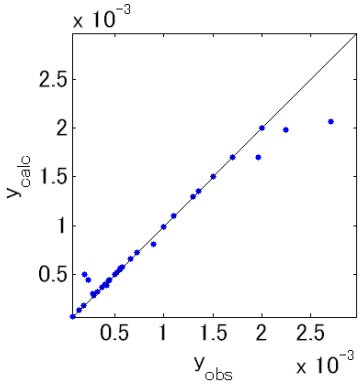
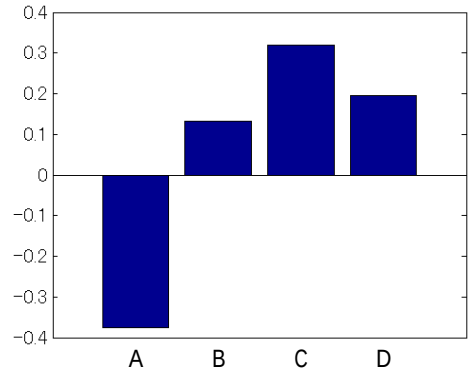
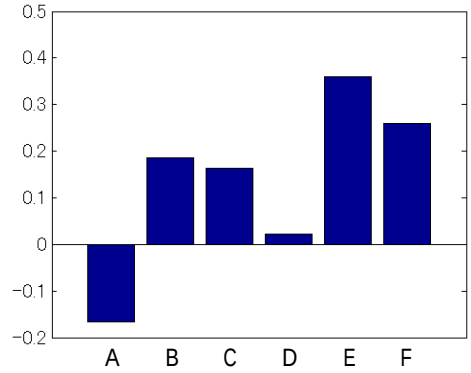


Fig. 2 The relationship between measured and predicted R_f in SVR



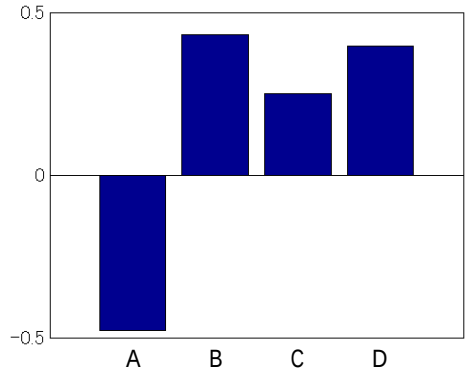
(a) only experimentally obtained parameters



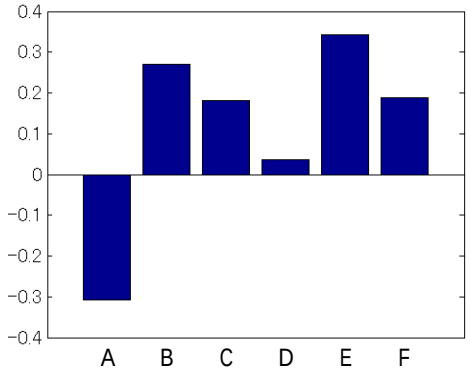
(b) all explanatory variables

Fig. 1 Standard regression coefficients of R_f models in PLS modeling.

A: shear rate, B: slurry concentration at the end of cooling, C: cooling time constant, D: degree of supersaturation on precipitation, E: precipitation time constant, F: maximum degree of supersaturation



(a) only experimentally obtained parameters



(b) all explanatory variables

Fig. 3 Standard regression coefficients of viscosity models in PLS modeling.

Meanings of A-F are the same as those of Fig. 1.

Table 1. Modeling Results of R_f and Viscosity

method	explanatory variables	R_f r^2	R_f RMSE ($\times 10^{-4}$)	R_f q^2	R_f RMSE ($\times 10^{-4}$)	viscosity r^2	viscosity RMSE	viscosity q^2	viscosity RMSE
PLS	a	0.453	4.66	0.342	5.11	0.693	1.05	0.591	1.21
	b	0.656	3.69	0.550	4.22	0.775	0.916	0.690	1.06
SVR	a	0.954	1.35	0.694	3.48	0.913	0.561	0.736	0.977
	b	0.875	2.22	0.667	3.68	0.983	0.247	0.899	0.603

a:experimentally obtained parameters, b:experimentally obtained parameters and characteristic values, which are extracted from experimental data