A New Process Variable and Dynamics Selection Method Based on a Genetic Algorithm-Based Wavelength Selection Method

Hiromasa Kaneko and Kimito Funatsu
Dept. of Chemical System Engineering, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan

DOI 10.1002/aic.13814
Published online April 23, 2012 in Wiley Online Library (wileyonlinelibrary.com).

Soft sensors have been used in industrial plants to estimate process variables that are difficult to measure online. Soft sensor models predicting an objective variable should be constructed with only important explanatory variables in terms of predictive ability, better interpretation of models and lower measurement costs. Besides, some process variables can affect an objective variable with time-delays. Therefore, we have proposed the methods for selecting important process variables and optimal time-delays of each variable simultaneously, by modifying the genetic algorithm-based wavelength selection method that is one of the wavelength selection methods in spectrum analysis. The proposed methods can select time-regions of process variables as a unit by using process data that includes process variables that are delayed in the range from zero to a set/given maximum value. The case study with simulation data and real industrial data confirmed that predictive, easy-to-interpret, and appropriate models were constructed using the proposed methods. © 2012 American Institute of Chemical Engineers *AIChE J,* 58: 1829–1840, 2012

Keywords: soft sensor, variable selection, process dynamics, maintenance, process control

Introduction

Plant operators have to monitor the operating conditions of industrial plants and control process variables, such as temperature, pressure, liquid level, and concentration of products. These variables, therefore, must be measured online. However, it is not easy to measure all variables online due to technological limitations, large measurement delays, and high investment costs. Thus, soft sensors have been widely used to estimate process variables that are difficult to measure online.[1,2] An inferential model is constructed between those variables that are easy to measure online and those that are not, and an objective variable **y** is then estimated using that model. Through the use of soft sensors, the values of objective variables can be estimated with a high degree of accuracy in real-time. In addition, soft sensors can give useful information in terms of fault detection by working with hardware sensors in parallel.[3,4]

It is often the case that the number of explanatory variables (**X**-variables) is very large and there is high correlation among **X**-variables when process data is analyzed for the construction of soft sensor models. For instance, modeling a relationship between **X** and **y** is done by using multiple linear regression (MLR), which works well as long as **X**-variables are few and uncorrelated. It is impossible for MLR to construct a regression model when the number of **X**-variables is more than the number of samples and a model is likely to be unsteady using highly correlated **X**-variables. Thus, the principal component regression method[5] and the partial least-squares (PLS) method[6] are widely used under such situations. Since the score matrix **T** translated from **X** is dimensionally reduced and mutually orthogonal, a stable model would be constructed even when the number of **X**-variables is larger than the number of samples.

In addition, Chong and Jun[7] remarked that process engineers are often interested in finding the few vital **X**-variables that would be most influential on a **y**-variable.[7] In PLS modeling, standard regression coefficients[7] and the variable importance in the projection (VIP) scores[6] are used as importance measures of **X**-variables for a **y**-variable. However, it is dangerous for regression coefficients to be simply considered as important for each variable because there is high correlation among **X**-variables[8] and VIP can find variables that are important not only for prediction but also for describing **X**-variables.[9] When soft sensor models are reconstructed with such process data, dramatic changes in these measures make it difficult for process engineers to find important **X**-variables and obtain process knowledge from the models.

Variable selection, therefore, has been widely focused to obtain improvement of the predictive ability, a better interpretation, or lower measurement costs.[9] In fact, models with higher accuracy and smaller number of **X**-variables are examined by using variable selection methods such as the least absolute shrinkage and selection operator (Lasso) method[10] and stepwise selection methods.[11] Some of the methods do not consider predictive accuracy but only fitting accuracy in variable selection while the stepwise selection method using the root-mean-squared error of cross validation

Correspondence concerning this article should be addressed to K. Funatsu at funatsu@chemsys.t.u-tokyo.ac.jp.

($\mathrm{RMSE}_{\mathrm{CV}}$) shown in Appendix as a criterion can take predictive accuracy into account, for example. In addition, it is probably difficult for stepwise selection methods to choose important combinations of **X**-variables because each variable is selected one by one. Meanwhile, the genetic algorithm-based PLS (GAPLS) method,[12] a variable selection method applying GA, aims to select important combinations of variables. A set of variables that is able to construct the PLS model having the optimum q^2 value, as explained in Appendix, can be obtained by using the GAPLS method. However, if the number of **X**-variables is large, the same variables cannot be selected after every GAPLS calculation, since the results are inconsistent, the number of selected variables can be still large, and it is very hard for process engineers to interpret GAPLS models and extract useful information from them.

Moreover, some process variables can affect a **y**-variable with a time-delay, that is, a time-delayed process variable having a stronger relationship with a **y**-variable than the nondelayed process variable. Therefore, the predictive accuracy of soft sensor models changes through the setting of time-delays.[13] Though it is expected that the selection of optimum time-delays of each process variable will improve predictive accuracy and interpretability of soft sensor models, there is very high correlation among time-delayed process variables and this makes it difficult to select the optimum values. In some studies, soft sensor models are constructed using process dynamics[13] and in others, the selection of process variables is used to increase the predictive accuracy.[14,15] However, no one has yet realized the optimization of both considerations in process dynamics and process variable selection.

In view of this, we have attempted to select process variables and process dynamics in plants, that is, optimal time-delays of each process variable, in a simultaneous fashion. In the field of spectrum analysis, data are analyzed in which there is high correlation among wavelengths, and much attention has been given to wavelength selection, by using wavelength regions as a unit of measurement. For example, stacked PLS (SPLS),[16] searching combination moving window PLS (SCMWPLS),[16] and genetic algorithm-based wavelength selection (GAWLS)[17] were proposed as region selection methods. Arakawa et al.,[17] carried out multiple case studies and concluded that GAWLS works better than SPLS or SCMWPLS in terms of predictive accuracy. GAWLS is one of the variable selection methods that are used for spectrum data. By using a genetic algorithm, it can select wavelengths with important information using regions as a unit of measurement.

In this study, the GAWLS method was applied to process data to incorporate information on process variables and process dynamics into soft sensor models, simultaneously. We have named this method as genetic algorithm-based process variables and dynamics selection (GAVDS) method because it involves selecting objects that are not wavelengths but process variables and dynamics, though the concept of GAVDS is identical with that of GAWLS. The GAVDS method treats process data represented in Figure 1 and involves all **X**-variables that are delayed in the range from zero to a set/given maximum value, and we are then able to select a number of time regions for each important variable.

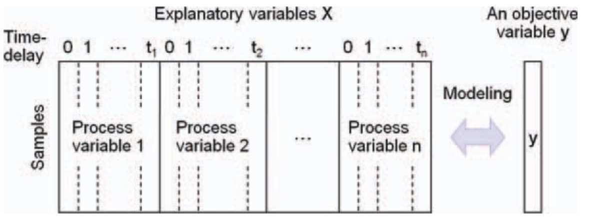

Figure 1. A data representation, considering process variables and dynamics.

The total number of X-variables is (t_1 + 1) + (t_2 + 1) + … + (t_n + 1). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

We have also modified the GAVDS method to use averages of selected variables, that is, selected time regions as input variables of a regression model. This method is called average GAVDS (aGAVDS). By using averages of selected variables, the number of input variables is the same as the number of time regions and thereby a reduction in the number of selected variables can be achieved. Traditional methods for process variable selection such as the Lasso method, stepwise selection method, and the GAPLS method did not consider adjacent variables that are similarly time-delayed process variables and it is difficult to select process variables with consideration of the process dynamics. Models constructed with a small number of **X**-variables by using GAVDS or aGAVDS will make interpretation easier and will lead to further consideration of final explanatory variables for the construction of soft sensors. Additionally, we are able to decrease the amount of uncertainty in parameters and effort put into the maintenance of soft sensor models during model reconstruction by using the proposed methods.

To verify the usefulness of the proposed methods, we applied them to simulation data and real industrial data. We analyzed the simulation data where **X**-variables have high correlation and time-delayed variables have coefficients of the contribution to **y**. The data obtained from the operation of the distillation column at the Mizushima works, Mitsubishi Chemical Corporation. Figure 1 shows the data format used in this study. By using this data format, we compared the standard regression coefficients and the VIP values of the PLS models, the Lasso method, the stepwise selection methods, the GAPLS method, the GAVDS method, and the aGAVDS method.

Method

PLS[6]

PLS is a method for relating explanatory variables, **X**, and an objective variable, **y**, using a linear multivariate model. The approach goes beyond traditional regression methods in that it also models the structures of **X** and **y**. In PLS modeling, the covariance between **y** and the score vector $\mathbf{t}_i$ is maximized. A PLS model has higher predictive power than ordinary least-squares models.[18]

A PLS model consists of the following two equations

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} \tag{1}$$

$$\mathbf{y} = \mathbf{T}\mathbf{q} + \mathbf{f} \tag{2}$$

where **T** is a score matrix, **P** is an **X**-loading matrix, **q** is a **y**-loading vector, **E** is a matrix of **X** residuals, and **f** is the vector of **y** residuals. The PLS regression model is as follows

$$\mathbf{y} = \mathbf{Xb} + \mathbf{const} \tag{3}$$

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

where $\mathbf{W}$ is an $\mathbf{X}$-weight matrix and $\mathbf{b}$ is a vector of regression coefficients. To construct a highly predictive model, the number of components in the PLS models must be chosen appropriately. In this study, the q^2 values were used as the measure and defined in Appendix. The optimum number of components was determined by the first local maximum of q^2.

Standard regression coefficients[7] and the VIP score[6] are used as the measure of the importance of $\mathbf{X}$-variables. The VIP score for the ith variable is defined as follows

$$\mathrm{VIP}_i = \sqrt{p\sum_{j=1}^{A}\left\{SS(q_j\mathbf{t}_j)\left(w_{ij}\big/\|w_j\|^2\right)\right\}\bigg/\sum_{j=1}^{A}SS(q_j\mathbf{t}_j)} \tag{5}$$

$$SS(q_j\mathbf{t}_j) = q_j^2\mathbf{t}_j^t\mathbf{t}_j \tag{6}$$

where p is the number of $\mathbf{X}$-variables and A is the number of latent variables.

Lasso[10]

Lasso is one of regularized least-squares methods and for calculation of regression coefficients $\mathbf{b}$ in Eq. 3 minimizes the regularized error function as follows

$$\frac{1}{2}\|\mathbf{y} - \mathbf{Xb}\|^2 + \frac{\lambda}{2}\sum_{i=1}^{p}|b_i| \tag{7}$$

where λ is the regularization coefficient that controls the relative importance of the sum-of-squares error and sum of the absolute value of b_i. The Lasso method has the property that some of the values of b_i are likely to be zero, which means variable selection, if λ is sufficiently large because of the constraint of the regularization term in Eq. 7. Minimizing Eq. 7 is equal to the following optimization problem:

Minimize

$$\frac{1}{2}\|\mathbf{y} - \mathbf{Xb}\|^2 \tag{8}$$

subject to

$$\sum_{i=1}^{p}|b_i| \le \eta \tag{9}$$

Where η is a constant. The optimum η value was determined by the best q^2 value when it changes from 0 to 10 in increments of 0.5 in this study.

Stepwise selection[11]

The stepwise selection method is one of the variable selection methods and aims to find a subset of explanatory variables that best meet an objective criterion. An explanatory variable can be added or dropped at each step in a stepwise manner, until the value of a criterion can no longer be improved. In this study, Mallows' Cp (Cp),[19] Akaike's information criterion (AIC),[20] Bayesian information criterion (BIC),[21] or RMSE$_{\mathrm{CV}}$, shown in Appendix is used as a criterion so as to consider the diversity of the stepwise selection methods. In addition, we compared the forward–backward stepwise selection (FB), which begins with no explanatory variables in a model, and the backward–forward stepwise selection (BF), which begins with a model consisting of all explanatory variables.

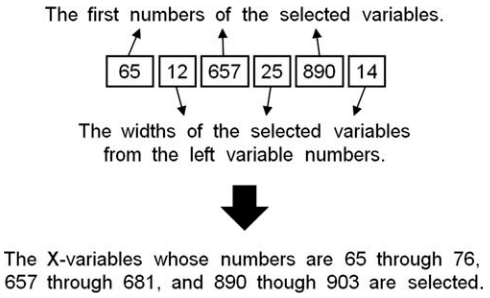
The first numbers of the selected variables.
65
12
657
25
890
14
The widths of the selected variables from the left variable numbers.
The X-variables whose numbers are 65 through 76, 657 through 681, and 890 though 903 are selected.

Figure 2. Coding rule for GAWLS.

This is the case when the number of regions is 3.

GAPLS[12]

One of the methods used to select important variables from $\mathbf{X}$-variables is GAPLS. A genetic algorithm[22] is an optimization method that is used in biology to model principles of natural evolution. Species having a high level of fitness under certain environmental conditions can prevail in the next generation, and the best species may be reproduced by crossover together with the random mutation of chromosomes in those species that survive. The solution space around superior individuals is searched for preferentially, which leads to the discovery of a solution that is close to the optimum.

GAPLS is a variable selection method that applies GA. Each of the $\mathbf{X}$-variables is assigned to a bit of the chromosome, and a set of variables that is able to construct the optimum PLS model is searched for. The r^2_{CV} value, which is calculated using a cross-validation method such as leave-one-out, is used as a fitness value of the chromosome. In this way, a model with high predictive accuracy is obtained. In this study, GAPLS is applied to data that includes all the explanatory variables that delay in the range from zero to a set/given maximum value as shown in Figure 1.

GAVDS and aGAVDS

GAWLS[17] is one of the methods that is used to select combinations of important variables from $\mathbf{X}$-variables using regions as a unit of measurement. GA is applied to select variables as it is in GAPLS. Figure 2 shows the coding method for GAWLS. Two actual values of a chromosome represent one region of variables. Hence, GAWLS can select important variables using regions as a unit of measurement. In Figure 2, the number of selected variables is 65 through 76, 657 through 681, and 890 through 903, and the number of regions is 3. The r^2_{CV} value is used as a fitness value of the chromosome as it is in GAPLS.

The GAVDS and aGAVDS methods are the modified methods of GAWLS applied to the process data represented in Figure 1 and can select process variables and dynamics simultaneously. Figure 3 shows the basic concept of these methods. In GAVDS and aGAVDS, variables are not selected across different process variables with one region. For GAVDS, selected variables are directly used as input variables of a regression model, but for aGAVDS, each average of selected variables in each region is used, that is, the number of input variables is 3, as shown in Figure 3. An

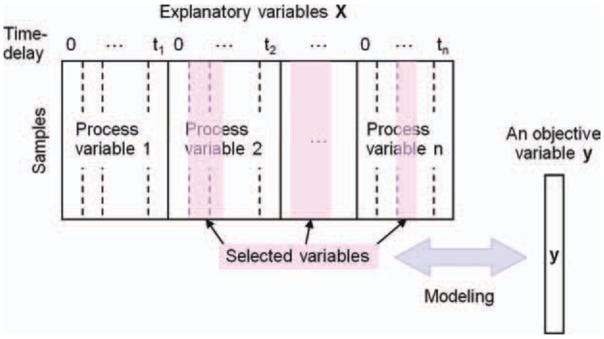

Figure 3. The basic concept of GAVDS and aGAVDS.

This is the case when the number of regions is 3. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

aGAVDS model can be simpler than that of GAVDS because the number of input variables to a regression model is the same as the number of regions by using averages of each region in the aGAVDS case. However, since information of process dynamics is summarized as an average, the aGAVDS method can put less dynamics information into a soft sensor model can the GAVDS method.

In processes that include all explanatory variables that delay in the range from zero to a set/given maximum value, the variables are highly correlated, so it is reasonable to expect that some time ranges would affect an objective variable. The regression models can, therefore, be improved without overfitting, and both the significant process variables and the time ranges that are important for objective variables can be found simultaneously. By changing the parameters of GAVDS and aGAVDS such as the maximum time-delay t_i, the maximum width of regions and the number of regions, variables such as plant properties, process knowledge, and experience of process engineers can be easily and directly incorporated into constructed models. Moreover, models with a small number of **X**-variables by using GAVDS or aGAVDS would make it easy to interpret the models and would lead to final explanatory variables for the construction of soft sensors. In addition, the amount of uncertainty in parameters and effort put into the maintenance of soft sensor models during model reconstruction would also be reduced.

Results and Discussion

We applied the proposed methods to both simulation data and real industrial data to verify their effectiveness. In the analysis of industrial data, regression models were based on the time difference of **X** and **y** for reducing the effects of deterioration with age such as drift and gradual changes in the state of plants without reconstruction of the models[23] in this case study.

Modeling of the simulation data

A pair of neighboring **X**-variables often has high correlation in the field of soft sensors. To achieve such correlation, an *i*th **X**-variable, $\mathbf{x}_i$, was set as

Table 1. Three Normal Distributions used for the Simulation Data

	Mean	Variance
A	10 delay in $\mathbf{x}_2$	40
B	15 delay in $\mathbf{x}_3$	40
C	5 delay in $\mathbf{x}_5$	60

$$\mathbf{x}_i = \begin{cases} \mathrm{U}(0,1) & (i = p) \\ 0.95\mathbf{x}_{i+1} + 0.05\mathrm{U}(0,1) & (i = p-1, p-2, \ldots p-q) \\ 0.95\mathbf{x}_{i-1} + 0.05\mathrm{U}(0,1) & (i = p+1, p+2, \ldots p+q) \end{cases} \tag{10}$$

where U(0,1) is a vector of uniform pseudorandom numbers ranging from 0 to 1; and *p* and *q* are the natural numbers. In this case study, $p = 4$, $q = 3$, and hence, seven variables were prepared. In addition, each $\mathbf{x}_i$ included time-delayed variables. The *r*-delayed variable of $\mathbf{x}_i$, $\mathbf{x}_i(r)$, is given as follows

$$\mathbf{x}_i(r) = \begin{cases} \mathbf{x}_i & (r = 0) \\ 0.95\mathbf{x}_i(r-1) + 0.05\mathrm{U}(0,1) & (r = 1, 2, \ldots t_i) \end{cases} \tag{11}$$

In this article, t_1, t_2, …, and t_7 were set as 20, that is, the number of **X**-variables was 147 (=7 × 21). Pairs of **X**-variables whose numbers of *i* and *r* are close have high correlation coefficients and those of separate variables have little correlation.

To consider the dynamics of the contribution of **X**-variables to **y**, we examined the case where **y** is a linear combination of **X** and its coefficients are set as parts of three normal distributions of the probability density, as given in Table 1. The coefficients of normal distributions *A*, *B*, and *C* are 2, 3, and 2, respectively. Figure 4 shows the coefficients of the contributions of each **X**-variable to **y**.

Random numbers from a normal distribution with standard deviation of 0.1 and mean of zero were finally added to **X** and **y** as noise. The numbers of training data and test data were both set as 100.

The modeling and predictive results of PLS, Lasso, stepwise, GAPLS, GAVDS, and aGAVDS are given in Table 2. The details of r^2 and r^2_{CV} are explained in Appendix. r^2_{pred} represents values of r^2 calculated from test data. The results of $\mathrm{RMSE_{CV}}$ in the FB stepwise and Cp in the BF stepwise were the same as those of Cp in the FB stepwise and the only one variable was selected from 147 variables. Meanwhile, in the BF stepwise methods, all variables were selected from 147 variables for AIC, BIC, and $\mathrm{RMSE_{CV}}$, that is, the results were identical to those before variable selection.

We used the Genetic Algorithm Optimizing Toolbox for MATLAB5[24] for the calculations of GAPLS, GAVDS, and aGAVDS. A q^2 value calculated with five-fold cross-validation was used as an evaluation function of the chromosome. The number of generations was set to 1000, and the number of populations was set to 500. The probability of crossover and the probability of mutation were given as default values. In this case study, the maximum size of regions was set as 20 and the number of regions was set as 3 for GAVDS and

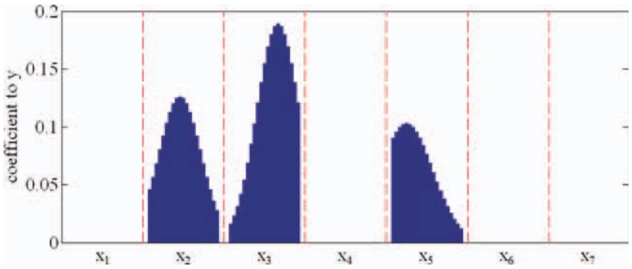

Figure 4. Coefficients of the contributions of each X-variable to y.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 2. Modeling and Prediction Results for the Simulation Data

		#selvar*	$A^{\dagger}$	r^2	r^2_{CV}	r^2_{pred}
No variable selection		147	3	0.989	0.986	0.983
Lasso		54	1.5	0.996	0.982	0.975
FB	Cp	1	1	0.954	0.953	0.948
	AIC	99	1	0.962	0.960	0.966
	BIC	15	1	0.979	0.978	0.975
GAPLS	Average	73	3.6	0.991	0.987	0.983
	Std. dev.	6.7	0.7	9.9×10^{-4}	9.2×10^{-4}	1.1×10^{-3}
GAVDS #reg‡ = 3	Average	26	1.7	0.988	0.987	0.984
	Std. dev.	6.6	0.8	6.1×10^{-4}	7.8×10^{-4}	9.6×10^{-4}
aGAVDS #reg‡ = 3	Average	26	1.1	0.988	0.988	0.985
	Std. dev.	2.3	0.5	3.1×10^{-4}	4.8×10^{-4}	4.8×10^{-4}

*The number of selected variables.
†The η value for Lasso and the number of latent variables for the others.
‡The number of regions.

aGAVDS. We performed 50 calculations and the averages and standard deviations of these 50 calculations are given in Table 2 for GAPLS, GAVDS, and aGAVDS.

As shown in Table 2, the r^2, r^2_{CV}, and r^2_{pred} values of Cp and AIC in the FB stepwise were lower than those obtained before variable selection. Appropriate variable selection was not performed by using those methods. The performance of the models for Lasso and BIC in the FB stepwise were slightly lower than that of PLS, that is, before variable selection, whereas the models for the other variable selection methods had almost same accuracy and predictive ability as seen with PLS.

Figure 5 shows the standard regression coefficients of each variable about the PLS model and the Lasso model and the VIP values of each variable about the PLS model. Figure 5a is complicated and the standard regression coefficients of some **X**-variables are negative though true coefficients to **y** are zero or positive. Figure 5b is completely different from Figure 4 and the VIP values of the PLS model could not reflect the contribution of **X**-variables to **y**. For the Lasso model, the variable selection was carried out for all kinds of $\mathbf{x}_i$, and, therefore, unnecessary variables remained. Figure 5c is also too complicated to analyze and search for important variables.

The frequency of each variable about the 50 GAPLS, GAVDS, and aGAVDS models are shown in Figure 6. The selected variables were not consistent using the GAPLS method. Meanwhile, by using GAVDS and aGAVDS, the results were consistent and only important variables that have high peaks in Figure 4 were selected with consideration of the dynamics.

We investigated the performance of each method, changing the amount of standard deviation of noise. The inferior-to-superior relationships among the methods shown in Table 2 were almost identical. The PLS method was one of the methods having the highest accuracy and the best predictive ability in this case study. Not all **X**-variables contributed to **y**, but many **X**-variables have strong correlation with the **X**-variables having high coefficients of the contribution to **y**. Hence, the **X**-variables that did not contribute to **y** had relationships with **y**. The PLS method could model the relationship between **y** and all **X**-variables, considering the correlation among **X**-variables appropriately.

Even in this situation, the GAVDS and aGAVDS methods could select the significant variable regions where the contributions to **y** were high in Table 2, as shown in Figure 6. We confirmed that the proposed methods can achieve appropriate selection of process variables and the dynamics in a simultaneous manner.

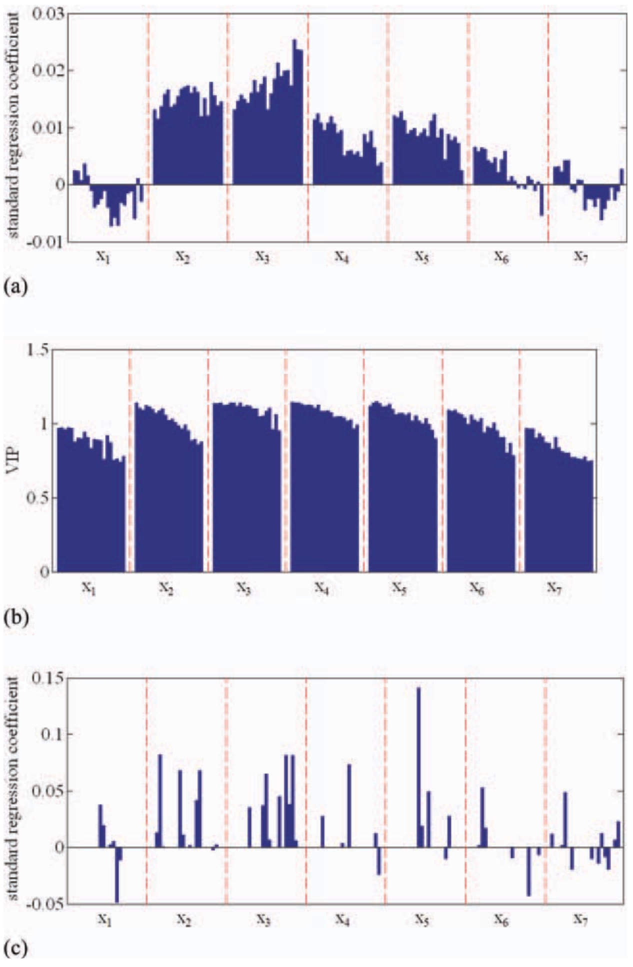

Figure 5. Standard regression coefficients of each variable about the PLS model and the Lasso model and the VIP values of each variable about the PLS model for the simulation data.

(a) Standard regression coefficients about the PLS model. (b) VIP values. (c) Standard regression coefficients about the Lasso model. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

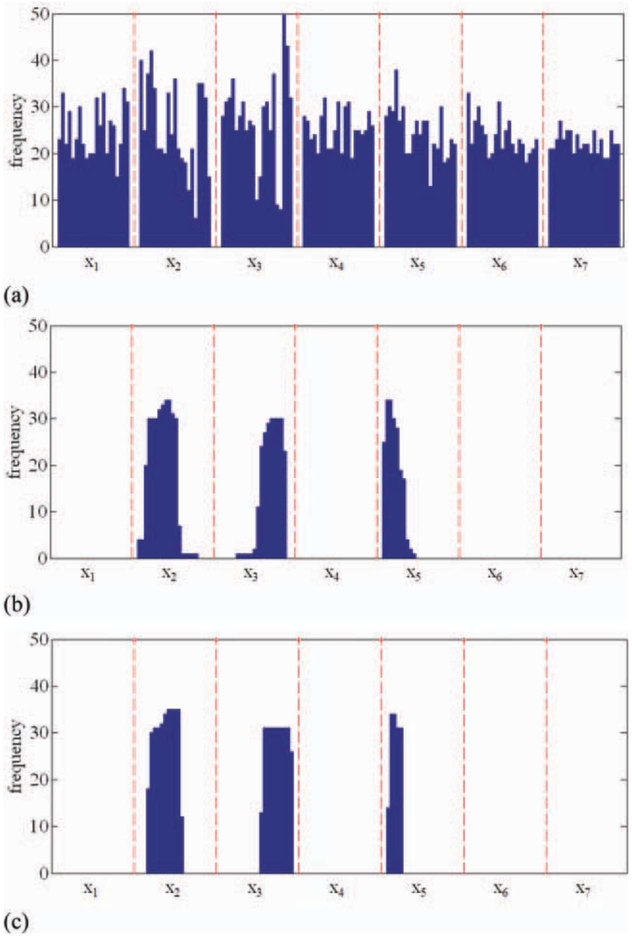

Figure 6. Frequency of each variable about the 50 GAPLS, GAVDS, and aGAVDS models.

(a) GAPLS. (b) GAVDS. (c) aGAVDS. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Application to Distillation Column

We analyzed the data obtained from the operation of a distillation column at Mizushima works, Mitsubishi Chemical Corporation. Figure 7 shows a schematic representation of the distillation column and Table 3 shows the process variables. An objective variable that represents the concentration of the bottom product having a lower boiling point, and explanatory variables that represent 19 variables such as temperature and pressure were used. The input variables are F_3 and F_4, and the operational variables are F_1 and F_2. The measurement interval of **y** is 30 min and that of **X** is 1 min. We collected data from monitoring that took place from 2002 to 2006, and used data from January to March 2003 for training data because plant tests took place and data from April 2003 to December 2006 for test data. Basically, a plant inspection took place every year. Data that reflects variations caused by **y**-analyzer faults were eliminated in advance.

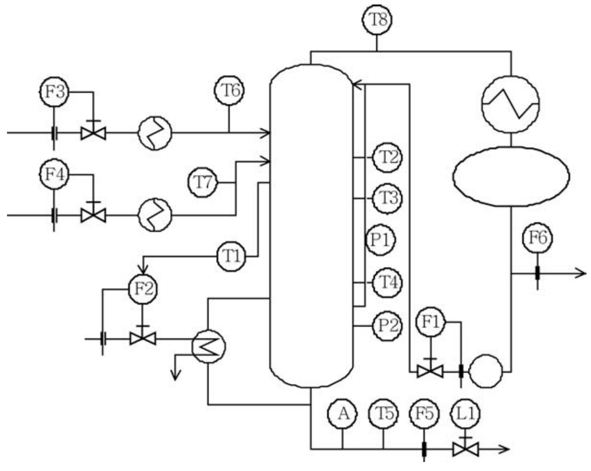

Figure 7. A schematic representation of the distillation column.

Table 3. Process Variables

No.	Symbol	Objective Variable
	A	Bottom product concentration
No.	Symbol	Explanatory variables
1	F_1	Reflux flow
2	F_2	Reboiler flow
3	F_3	Feed 1 flow
4	F_4	Feed 2 flow
5	F_5	Bottom flow
6	F_6	Top flow
7	L_1	Liquid level
8	P_1	Pressure 1
9	P_2	Pressure 2
10	T_1	Temperature 1
11	T_2	Temperature 2
12	T_3	Temperature 3
13	T_4	Temperature 4
14	T_5	Bottom temperature
15	T_6	Feed 1 temperature
16	T_7	Feed 2 temperature
17	T_8	Top temperature
18	$F_4/F_3 = R$	Reflux ratio
19	$F_1/F_6 = F$	Feed flow ratio

To incorporate the dynamics of process variables into soft sensor models, **X** included each process variable that was time-delayed as shown in Figure 1. We considered the two cases described as follows. First, the 1159 explanatory variables are the ones including each process variable that was delayed for durations ranging from 0 min through 60 min in steps of 1 min, that is, $t_1 = t_2 = \ldots = t_n = 60$ and the time interval of each process variable is 1 min in Figure 1 ($1159 = 19 \times 61$). Next, the 133 explanatory variables are the ones including in each process variable that was delayed for durations ranging from 0 min through 60 min in steps of 10 min, that is, $t_1 = t_2 = \ldots = t_n = 60$ and the time interval of each process variable is 10 min in Figure 1 ($133 = 19 \times 7$), which means rough variable selection is performed in advance. For the GA-based methods, the results of only the 1159 variables were compared.

First, the PLS method was used to construct the regression models with all **X**-variables because the support vector regression[25] model had almost the same predictive accuracy as that of PLS for this distillation column.[4] Table 4 shows the modeling results. The details of the statistics are explained in Appendix. The number of **X**-variables did not significantly affect the statistics and the models were predictive to some extent. The standard regression coefficients of each **X**-variable by each PLS model are shown in Figure 8. The figures seem to be complicated, and in addition, the pattern of positive and negative coefficients is not consistent even for the same variables. Indeed, the regression

Table 4. Modeling Results for the Industrial Data

		#var*	#selvar†	A‡	r^2	RMSE	q^2	RMSE_{CV}
No variable selection		1159	1159	8	0.987	0.135	0.986	0.142
		133	133	12	0.985	0.144	0.985	0.149
Lasso		1159	380	4.5	0.979	0.127	0.986	0.142
		133	98	3.5	0.985	0.144	0.985	0.149
FB	Cp	1159	1	1	0.968	0.214	0.968	0.214
		133	1	1	0.967	0.217	0.967	0.218
	AIC	1159	227	20	0.988	0.129	0.987	0.138
		133	53	10	0.985	0.146	0.985	0.148
	BIC	1159	39	8	0.986	0.141	0.985	0.141
		133	25	10	0.984	0.151	0.984	0.152
GAPLS	Average	1159	360.7	9.2	0.987	0.137	0.986	0.144
	Std. dev.		5.4	0.63	1.4×10^{-4}	7.4×10^{-4}	1.8×10^{-4}	8.7×10^{-4}
GAVDS #reg§ = 5	Average	1159	87	4.9	0.983	0.157	0.983	0.158
	Std. dev.		6.6	0.3	4.3×10^{-4}	2.0×10^{-3}	4.4×10^{-4}	2.0×10^{-3}
GAVDS #reg§ = 10	Average	1159	126	7.1	0.985	0.147	0.984	0.149
	Std. dev.		18	1.1	4.0×10^{-4}	1.9×10^{-3}	4.1×10^{-4}	1.9×10^{-3}
GAVDS #reg§ = 15	Average	1159	157	7.2	0.985	1.45	0.985	0.147
	Std. dev.		30	0.9	3.2×10^{-4}	1.6×10^{-3}	3.0×10^{-4}	1.5×10^{-3}
aGAVDS #reg§ = 5	Average	1159	5	4.8	0.983	0.158	0.983	0.158
	Std. dev.		0	0.3	6.6×10^{-4}	3.0×10^{-3}	6.7×10^{-4}	3.0×10^{-3}
aGAVDS #reg§ = 10	Average	1159	10	9	0.985	0.149	0.984	0.149
	Std. dev.		0	0.9	2.3×10^{-4}	1.1×10^{-3}	2.2×10^{-4}	1.1×10^{-3}
aGAVDS #reg§ = 15	Average	1159	15	10.4	0.985	0.145	0.985	0.146
	Std. dev.		0	3.0	2.9×10^{-4}	1.4×10^{-3}	3.0×10^{-4}	1.5×10^{-3}

*The number of **X**-variables.
†The number of selected variables.
‡The η value for Lasso and the number of latent variables for the others.
§The number of regions.

coefficients did not conform to process knowledge and they changed dramatically if the models were updated because of the collinearity or multicollinearity of **X**-variables. We judge that these models were, therefore, impractical and it is difficult to extract information on process knowledge from them.

Figure 9 shows the VIP values of each **X**-variable by each PLS model. In almost all process variables, there seemed to be multiple peaks of the VIP values. This arises from the effect of autocorrelation because time-delayed variables are highly correlated with each other, as is often observed in process data. For the point of view of reducing the number of **X**-variables, some variables around just one of the peaks should be selected from certain process variables. However, if the threshold of VIP values is set as 1,[7] variables around more than two peaks are selected from a process variable, and then, the number of **X**-variables is still large. We, therefore, attempted to select important variables with consideration of the process dynamics.

First, Lasso was used to select variables from 1159 or 133 variables and Table 2 shows the results. The selected variables imply that their regression coefficients were not zero. The r^2, root-mean-square error (RMSE), r^2_{CV}, and RMSE_{CV}

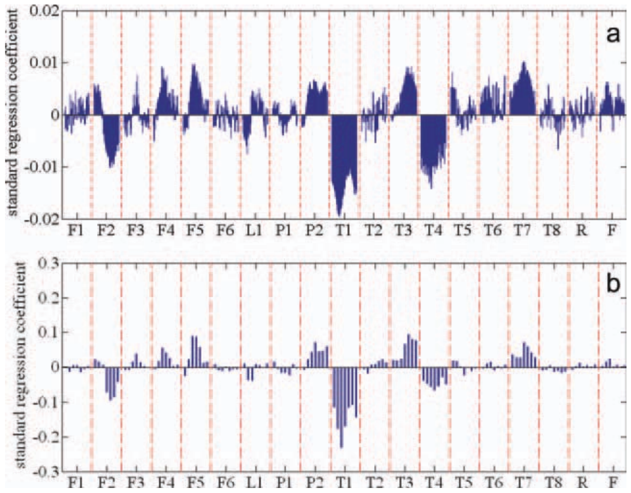

Figure 8. The standard regression coefficients of each variable about the PLS models before variable selection.

(a) 1159 variables. (b) 133 variables. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

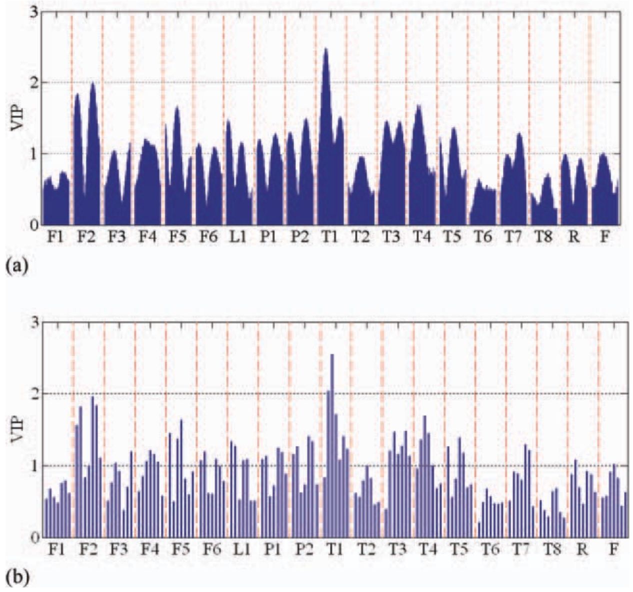

Figure 9. The VIP values of each variable about the PLS models before variable selection.

(a) 1159 variables. (b) 133 variables. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

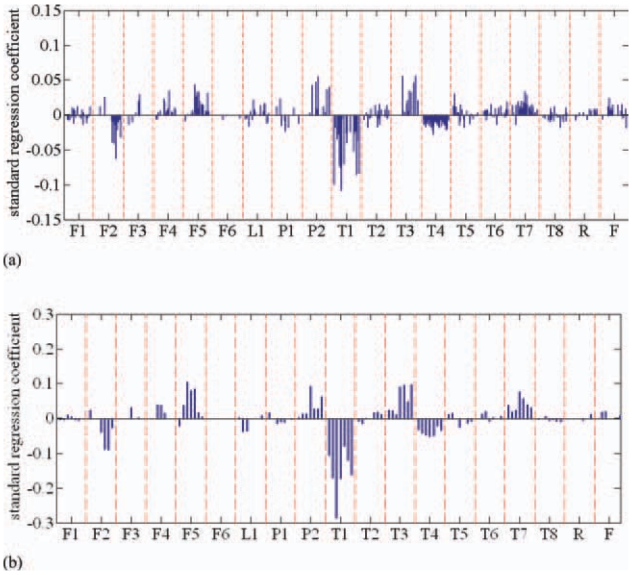

Figure 10. The standard regression coefficients of each variable about the Lasso models.

(a) 1159 variables. (b) 133 variables. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

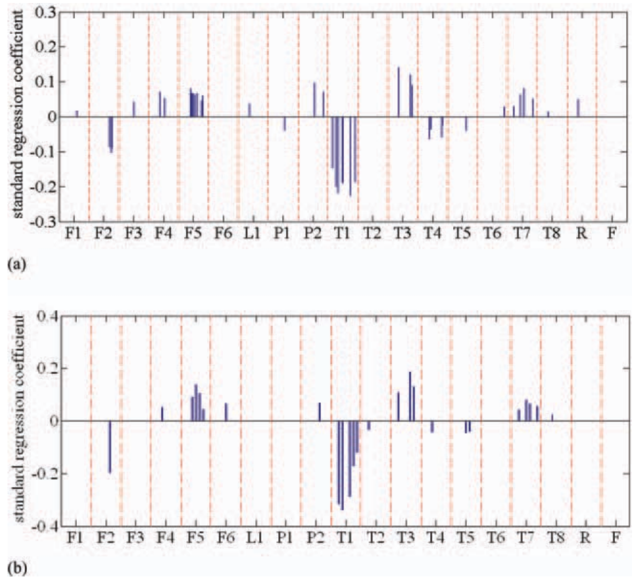

Figure 12. The standard regression coefficients of each variable about the PLS models after FB stepwise selection using BIC.

(a) 1159 variables. (b) 133 variables. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

values did not change much compared with those found before variable selection. By using Lasso, the **X**-variables used for each regression model could be reduced from 1159 to 380 and from 133 to 98 while the predictive accuracy of each model did not decrease. The standard regression coefficients of each Lasso model are shown in Figure 10. The results were simpler than those of Figure 8 that represents the results before variable selection. However, time-delayed variables were selected from almost all process variables and the number of variables was still large. Furthermore, the pattern of positive and negative coefficients is not consistent even for the same variables as was also seen before variable selection.

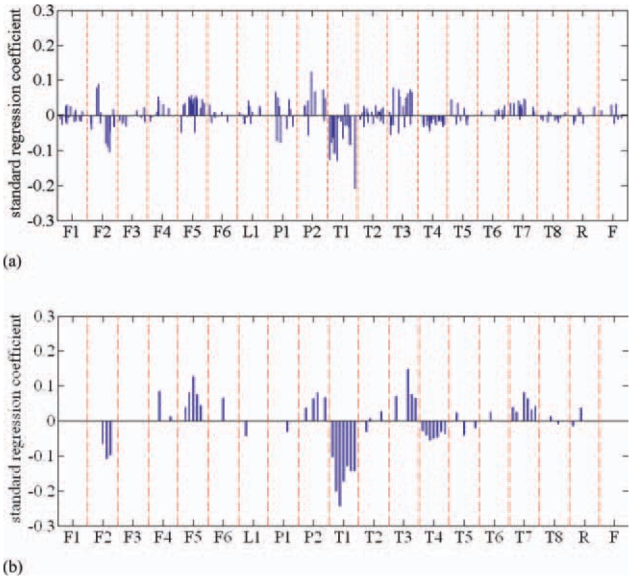

Figure 11. The standard regression coefficients of each variable about the PLS models after FB stepwise selection using AIC.

(a) 1159 variables. (b) 133 variables. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The results of the variable selection with the stepwise methods are shown in Table 4. In the FB stepwise method, the result of $\mathrm{RMSE_{CV}}$ was the same as that of Cp and the one variable was selected from 1159 or 133 variables. The selected variables were 18 min delayed Temperature 1 (T_1) for 1159 variables and 20 min delayed T_1 for 133 variables. The RMSE and $\mathrm{RMSE_{CV}}$ values were larger than those of the other methods, and the number of variables must be too small in those cases. When AIC was used as a criterion, 227 variables were selected from 1159 variables and 53 variables from 133 variables; when BIC was used as a criterion, 39 variables were selected from 1159 variables and 25 variables from 133 variables as shown in Table 4. Smaller sets of variables were selected by the FB stepwise with AIC or BIC than by Lasso while the predictive accuracy was almost the same. Figures 11 and 12 show the standard regression coefficients of each variable by the PLS models after the FB stepwise selection using AIC and BIC, respectively. The figures of AIC were simpler than those found before variable selection and almost same as those of Lasso. While no time-delayed variables were selected from Reflux flow (F_1), Feed 1 flow (F_3), and Feed flow ratio (F) for 133 variables from Figure 11b, the time-delayed variables were selected from all process variables for 1159 variables from Figure 11a. Using BIC, the simpler models could be constructed than was the case with the use of AIC. However, many kinds of process variables were selected; the time-delays were chosen discontinuously; and the results were not consistent. Pressure 1 (P_1) was selected from 1159 variables, but was not selected from 133 variables and Top flow (F_6) was selected from 133 variables, but not from 1159 variables, for example. Meanwhile, in the BF stepwise methods, all variables were selected from either 1159 or 133 variables for Cp, AIC, BIC, and $\mathrm{RMSE_{CV}}$, that is, the results are identical to

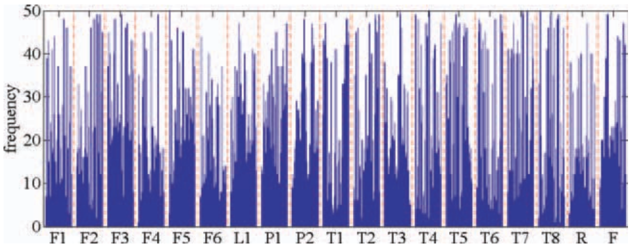

Figure 13. The selected frequency of each variable about 50 GAPLS models.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary. com.]

those obtained before variable selection. By using the stepwise methods, reasonable selection was not achieved for the construction of proper soft sensor models.

GAPLS was carried out to select variables from 1159 variables. The number of generations and populations were set to 1000 and 500, respectively. The other set values of GA were the same as those used in the analysis of the simulation data. Table 4 shows the averages and the standard deviations of 10 calculation results for GAPLS. The accuracy and predictive accuracy of GAPLS models were almost the same as those before variable selection as reflected in the r^2 and r^2_{CV} values. About 360 variables were selected on average and the number of selected variables did not vary widely. Subsequently, additional calculations of GAPLS were performed, and the selected frequency of each variable about the 50 GAPLS models is shown in Figure 13. This figure indicates that the selected variables were not consistent using the GAPLS method. The diverse solutions of GAPLS derive from the large number of **X**-variables and the collinearity or the multicollinearity of them, as is commonly seen with process data. Therefore, GAPLS results and those of the traditional methods that we described above make it difficult for process engineers to consider the final explanatory variables used to construct a detailed soft sensor model.

Lastly, variable selection was performed by using GAVDS and aGAVDS. The number of generations was set to 500 and the other setting values of GA were the same as those used for GAPLS. In this study, the maximum size of regions was set as 20 and the numbers of regions was 5, 10, or 15, with 10 calculations being performed in each setting. The modeling results of GAVDS and aGAVDS are shown in Table 2. For both GAVDS and aGAVDS, the r^2_{CV} values seemed to be relatively small and the $\mathrm{RMSE}_{\mathrm{CV}}$ values seemed to be relatively large compared to other results when the number of regions was small, but the difference was very small and the GAVDS and aGAVDS models whose number of regions is 5 displayed almost the same degree of accuracy and predictive accuracy as those before variable selection. Using GAVDS and aGAVDS, predictive models could be constructed with only 5 time-regions.

The examples of standard regression coefficients of each variable about a GAVDS model and an aGAVDS model are shown in Figure 14. These are the cases in which the number of windows is 5. For the GAVDS model, the figure is clear and the coefficients are almost consistent for each variable, but the pattern of positive and negative coefficients is not consistent for T_1, which probably arose as a result of the high collinearity of the time-delayed variables. The standard regression coefficients in the same regions of an aGAVDS model are constant because the aGAVDS method uses averages of regions as input variables. Of course, it is dangerous to simply consider these regression coefficients as important for each variable because there are correlations in **X**-variables; however, it is clear that the models are easier to interpret and make it easier for process engineers to consider the final process variables and their time-delay used in constructing a soft sensor model in detail than do those obtained before variable selection. The proposed methods could reduce 1159 variables to only 5 regions without decreasing the predictive ability.

Figure 15 shows the selected frequency of each variable about the 50 GAVDS models after additional calculations when the numbers of regions are 5, 10, or 15. The kinds of selected process variables increased as the number of regions increased, but the peak positions and the forms of the frequencies were similar among these cases, which supported the robustness and the consistency of GAVDS. In Figures 15a–c, there seemed to be two peaks for T_1, which is more correlated with **y** than the other process variables in this case study. This high correlation probably reflects the autocorrelation inherent in this study. Meanwhile, Feed 2 flow (F_4) and Feed 2 temperature (T_7) were selected many times while F_3 and Feed 1 temperature (T_6) were rarely selected. These high and low frequencies of both of the variables measured at the same feed input confirmed that the proposed method could produce significant and reasonable results. Figure 16 shows the frequency of each variable about the 50 aGAVDS models after the additional calculations, when the numbers of regions are 5, 10, or 15 as well as GAVDS. The forms of the frequencies were similar to those of GAVDS, which confirmed that the GAVDS and aGAVDS methods were robust regarding the change in the number of regions. In this case study, we used 5, 10, or 15 as the number of regions for GAVDS and aGAVDS, but by changing the number of regions, it is easy for process engineers to produce many kinds of soft sensor models into which plant properties, process knowledge, experience of process engineers, and others can be directly incorporated.

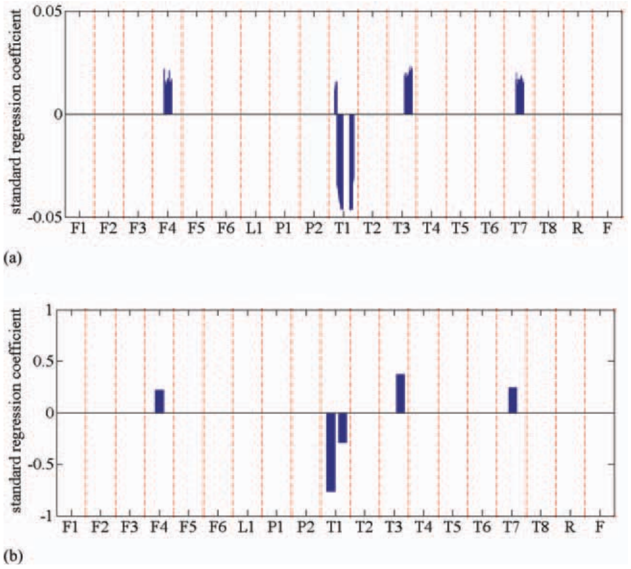

Figure 14. The standard regression coefficients of each variable about a GAVDS model and an aGAVDS model.

The number of regions is 5. (a) GAVDS. (b) aGAVDS. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

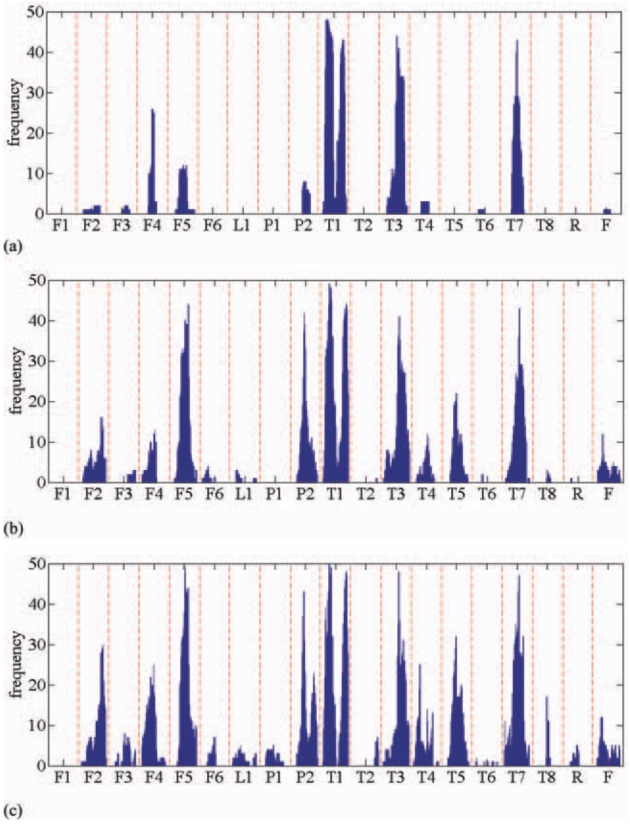

Figure 15. The selected frequency of each variable about the 50 GAVDS models.

The number of regions is 5, 10, or 15 as indicated. (a) The number of regions is 5. (b) The number of regions is 10. (c) The number of regions is 15. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

We checked the predictive accuracy for each year of each method. The prediction results are shown in Table 5. When the FB stepwise selection method with Cp was used and only 1 explanatory variable was selected for construction of regression models, the RMSE values in each year were larger than the others. However, the predictive accuracy of the other models were comparable in whole years as shown in the RMSE values in Table 5. The reduction of explanatory variables could be achieved without decreasing predictive accuracy by using each variable selection method. In addition, the GAVDS and aGAVDS models had the same predictive accuracy as others with only 5 regions of **X**-variables. We can, therefore, conclude that these models were easy to interpret and handle for further evaluation leading to the final soft sensor model, and could also be readily maintained for construction of appropriate models through use of the GAVDS and aGAVDS methods.

The PLS method was one of the methods having the highest accuracy and the best predictive ability in this analysis. This situation is probably the same as that found with the simulation data. Not all **X**-variables will contribute to **y**, but many **X**-variables have a strong correlation with the **X**-variables having high contribution to **y**, and, therefore, the **X**-variables that did not contribute to **y** had relationships with **y**. The PLS method could model the relationship between **y** and all **X**-variables, considering the correlation among **X**-variables appropriately. Even with the industrial data, the GAVDS and aGAVDS methods could achieve the appropriate selection of process variables and the dynamics, simultaneously.

Conclusion

In this article, we have proposed the GAVDS and aGAVDS methods for selecting important variables and evaluating process dynamics simultaneously by modifying the GAWLS method. These methods can select time-regions of process variables as a unit by using process data that includes process variables that delay in the range from zero to a set/given maximum value. Through the analysis of the simulation data, we verified that the proposed methods could construct the regression models with high accuracy and high predictive ability and select important variable-regions even in the presence of the collinearity among **X**-variables and noise. Additionally, the modeling results of real industrial data confirmed that predictive accuracy of the proposed models was comparable with those of the traditional methods, but were easy-to-interpret, and appropriate models were constructed using the proposed methods with only 5 time-regions. This will lead to improved interpretation and lower measurement costs. By making changes in the number of regions and checking the standard regression coefficients, the accuracy of the models, and other statistics, the soft sensor model can be easily optimized with implications for plant

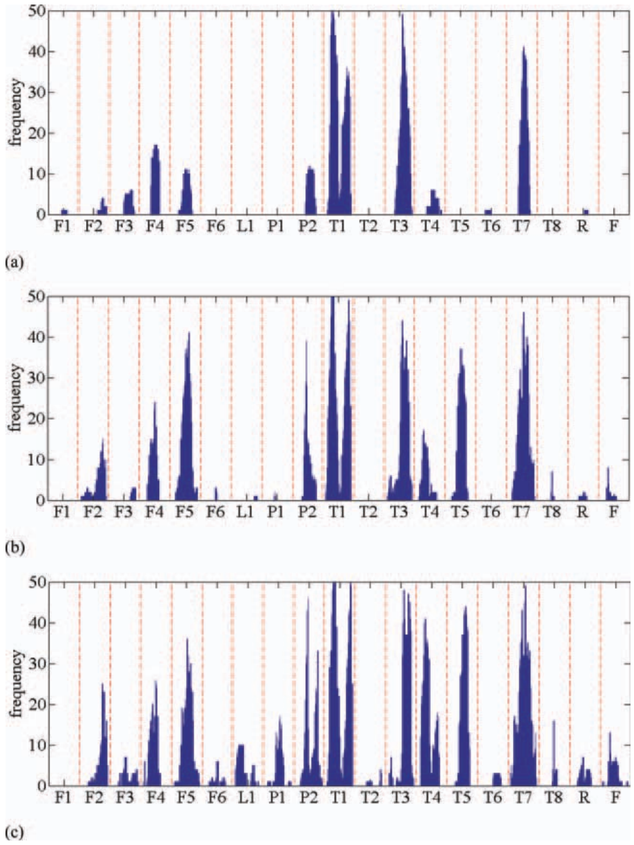

Figure 16. The selected frequency of each variable about the 50 aGAVDS models.

The number of regions is 5, 10, or 15 as indicated. (a) The number of regions is 5. (b) The number of regions is 10. (c) The number of regions is 15. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 5. Prediction Results for the Industrial Data

		#var*	#selvar†	2003	2004	2005	2006
No variable selection		1159	1159	0.277	0.634	0.482	0.375
		133	133	0.276	0.636	0.483	0.375
Lasso		1159	380	0.277	0.635	0.482	0.375
		133	98	0.274	0.633	0.482	0.373
FB	Cp	1159	1	0.315	0.667	0.523	0.405
		133	1	0.316	0.671	0.528	0.406
	AIC	1159	227	0.277	0.633	0.481	0.373
		133	53	0.276	0.631	0.480	0.374
	BIC	1159	39	0.277	0.640	0.485	0.373
		133	25	0.280	0.640	0.490	0.378
GAPLS	Average	1159	360.7	0.276	0.632	0.481	0.376
	Std. dev.		5.4	1.9×10^{-4}	8.8×10^{-4}	3.1×10^{-4}	2.5×10^{-4}
GAVDS #reg‡ = 5	Average	1159	87	0.285	0.657	0.496	0.385
	Std. dev.		6.6	1.8×10^{-3}	4.7×10^{-3}	2.5×10^{-3}	2.8×10^{-3}
GAVDS #reg‡ = 10	Average	1159	126	0.280	0.646	0.489	0.378
	Std. dev.		18	3.1×10^{-4}	1.6×10^{-3}	1.0×10^{-3}	5.4×10^{-4}
GAVDS #reg‡ = 15	Average	1159	157	0.279	0.642	0.489	0.377
	Std. dev.		30	3.7×10^{-4}	2.0×10^{-3}	9.9×10^{-4}	3.3×10^{-4}
aGAVDS #reg‡ = 5	Average	1159	5	0.283	0.646	0.493	0.381
	Std. dev.		0	5.9×10^{-4}	2.0×10^{-3}	1.4×10^{-3}	5.0×10^{-4}
aGAVDS #reg‡ = 10	Average	1159	10	0.281	0.643	0.491	0.379
	Std. dev.		0	8.4×10^{-4}	1.1×10^{-3}	1.4×10^{-3}	4.6×10^{-4}
aGAVDS #reg‡ = 15	Average	1159	15	0.280	0.639	0.489	0.377
	Std. dev.		0	8.4×10^{-4}	2.0×10^{-3}	1.1×10^{-3}	9.8×10^{-4}

The RMSE values are calculated for each year.
*The number of **X**-variables.
†The number of selected variables.
‡The number of regions.

properties, process knowledge, experience of process engineers, and other pertinent factors.

In this case study, there was no significant difference in accuracy and predictive ability between the GAVDS and aGAVDS models. In such a situation, the aGAVDS method is recommended because we can construct simpler regression models using the aGAVDS method than with the GAVDS approach. However, it is conceivable that the GAVDS method should be used if process dynamics is complicated and averages of time-delayed variables cannot extract the information.

The number of process variables was not high in this analysis, but the proposed method will work effectively in situations where there are more process variables, for example, thousands of process variables that have to be considered. The modeling results can make it easy to interpret the models and facilitate further assessment of the final explanatory variables for the construction of soft sensors. Moreover, adoption of the proposed methods would lead to a decrease in the error associated with a given parameter and a reduction in the maintenance effort in soft sensor models during model reconstruction, and probably contribute to the optimal positioning of sensors.

Since the algorithm of GAVDS and aGAVDS is independent of the regression method, it can be used with any nonlinear modeling methods. It can be expected that the problems of variable selection and maintenance of soft sensor models will be reduced by using our proposed methods.

Acknowledgments

The authors acknowledge the support of Mizushima works, Mitsubishi Chemical Corporation and the financial support of the Japan Society for the Promotion of Science.

Literature Cited

1. Kano M, Nakagawa Y. Data-based process monitoring, process control, and quality improvement: recent developments and applications in steel industry. *Comput Chem Eng*. 2008;32:12–24.
2. Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry. *Comput Chem Eng*. 2009;33:795–814.
3. Kaneko H, Arakawa M, Funatsu K. Development of a new soft sensor method using independent component analysis and partial least squares. *AIChE J*. 2009;55:87–98.
4. Kaneko H, Arakawa M, Funatsu K. Applicability domains and accuracy of prediction of soft sensor models. *AIChE J*. 2011;57:1506–1513.
5. Wold S. Principal component analysis. *Chemom Intell Lab Syst*. 1987;2:37–52.
6. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst*. 2001;58:109–130.
7. Chong IG, Jun CH. Performance of some variable selection methods when multicollinearity is present. *Chemom Intell Lab Syst*. 2005;78:103–112.
8. Kaneko H, Arakawa M, Funatsu K. Development of a new regression analysis method using independent component analysis. *J Chem Inf Model*. 2008;48:534–541.
9. Andersen CM, Bro R. Variable selection in regression—a tutorial. *J Chemom*. 2010;24:728–737.
10. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc*. 1996;58:267–288.
11. Hocking RR. The analysis and selection of variables in linear regression. *Biometrics*. 1976;32:1–49.
12. Hasegawa K, Miyashita Y, Funatsu K. GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *J Chem Inf Comput Sci*. 1997;37:306–310.
13. Kano M, Miyazaki K, Hasebe S, Hashimoto I. Inferential control system of distillation compositions using dynamic partial least squares regression. *J Process Control*. 2000;10:157–166.
14. Rallo R, Ferre GJ, Arenas A, Giralt F. Neural virtual sensor for the inferential prediction of product quality from process variables. *Comput Chem Eng*. 2002;26:1735–1754.
15. Zamprogna E, Barolo M, Seborg DE. Optimal selection of soft sensor inputs for batch distillation columns using principal component analysis. *J Process Control*. 2005;15:39–52.
16. Ni W, Brown SD, Man R. Stacked partial least squares regression analysis for spectral calibration and prediction. *J Chemom*. 2009;23:505–517.
17. Arakawa M, Yamashita Y, Funatsu K, Genetic algorithm-based wavelength selection method for spectral calibration. *J Chemom*. 2011;25:10–19.
18. Faber K, Kowalski BR. Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. *J Chemom*. 1997;11:181–238.
19. Mallows CL. Some comments on Cp. *Technometrics*. 1973;15:661–675.
20. Akaike H. Factor analysis and AIC. *Psychometrika*. 1987;52:317–332.

21. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6:461–464.
22. Whitley D. A genetic algorithm tutorial. *Stat Comput*. 1994;4:65–85.
23. Kaneko H, Funatsu K. Maintenance-free soft sensor models with time difference of process variables. *Chemom Intell Lab Syst*. 2011;107:312–317.
24. Houck CR, Joines JA, Kay MG. A Genetic Algorithm for Function Optimization: A Matlab Implementaion. NCSU-IE TR 95–09. Metaheuristic Research and Applications Group: North Carolina State University, Raleigh, NC, 1995.
25. Vapnik VN. *The Nature of Statistical Learning Theory*. New York: Springer, 1999.

Appendix

To construct a highly predictive model, the number of components in the PLS models, the η value in the Lasso model, and other values must be chosen appropriately. The r^2 and r^2_{CV} values are used as measured and defined as follows

$$r^2 = 1 - \frac{\sum \left(y_{\mathrm{obs}} - y_{\mathrm{calc}}\right)^2}{\sum \left(y_{\mathrm{obs}} - \bar{y}\right)^2} \tag{A1}$$

$$r^2_{\mathrm{CV}} = 1 - \frac{\sum \left(y_{\mathrm{obs}} - y_{\mathrm{pred}}\right)^2}{\sum \left(y_{\mathrm{obs}} - \bar{y}\right)^2} \tag{A2}$$

where y_{obs} is the measured y value, y_{calc} is the calculated y value, and y_{pred} is the predicted y value in the cross-validation procedure. In this study, the five-fold cross validation method is used in the calculation of y_{pred}. In the above equations, r^2 represents the fitting accuracy of the constructed models and r^2_{CV} represents the predictive accuracy of the constructed models. Values close to unity for both r^2 and r^2_{CV} are favorable. The r^2 and r^2_{CV} values must both be compared using models constructed with the same objective variables data.

The RMSE of y_{calc} and y_{pred} is defined as follows

$$\mathrm{RMSE} = \sqrt{\frac{\sum \left(y_{\mathrm{obs}} - y_{\mathrm{calc}}\right)^2}{n}} \tag{A3}$$

$$\mathrm{RMSE}_{\mathrm{CV}} = \sqrt{\frac{\sum \left(y_{\mathrm{obs}} - y_{\mathrm{pred}}\right)^2}{n}} \tag{A4}$$

The lower the RMSE and $\mathrm{RMSE}_{\mathrm{CV}}$ values, the higher will be the accuracy and predictive accuracy obtained with the constructed model.

Manuscript received Mar. 4, 2011, revision received Jan. 31, 2012, and final revision received Apr. 4, 2012.