Ind. Eng. Chem. Res. **2009,** *48,* 5379–5387

Data-Driven Soft Sensor Design with Multiple-Rate Sampled Data: A Comparative Study

Bao Lin,[†,‡,§] Bodil Recke,[‡] Torben M. Schmidt,[‡] Jørgen K. H. Knudsen,[‡] and Sten Bay Jørgensen*,[†]

CAPEC, Department of Chemical Engineering, Technical University of Denmark and FLSmidth Automation, Valby 2500, Denmark

Multirate systems are common in industrial processes where quality measurements have slower sampling rates than other process variables. Since intersample information is desirable for effective quality control, different approaches have been reported to estimate the quality between samples, including the numerical interpolation, polynomial transformation, data lifting, and weighted partial least squares (WPLS). Two modifications to the original data lifting approach are proposed in this paper: reformulating the extraction of a fast model as an optimization problem and ensuring the desired model properties through Tikhonov Regularization. A comparative investigation of the four approaches is performed. Their applicability, accuracy, and robustness to process noise are evaluated with a single-input single-output (SISO) system. The modified data lifting and WPLS approaches are implemented to design quality soft sensors for cement kiln processes using data collected from a simulator and a plant log system. Preliminary results reveal that the WPLS approach is able to provide accurate one-step-ahead prediction. The regularized data lifting technique predicts the product quality of cement kiln systems reasonably well, demonstrating the potential to be used for effective quality control and as an advanced component of process analytical technology (PAT).

1. Introduction

In most chemical processes, quality measurements have slower sampling rates than other process variables. A major obstacle for effective quality control is the lack of real-time information, due to the time delay associated with laboratory analysis or slowly processed quality measurements of online analyzers. The Food and Drug Administration (FDA) recently introduced process analytical technology (PAT) into the pharmaceutical industry to ensure high and consistent product quality. An essential component of PAT is the real-time information of product properties. Soft sensors derived with multivariate statistical approaches can be powerful tools in the pharmaceutical industry to facilitate process understanding, to monitor process operation and quality, to detect abnormal situations, and to improve process reliability.[1,2] This paper presents a comparative study of several approaches to address the issue of multirate sampling for data-driven soft sensor development.

Plant data-log systems usually provide the operating data sampled at different frequencies. Process inputs $\{u(i)\}$ and secondary outputs $\{y^s(i)\}$ are sampled at a high frequency (with sampling interval h), while primary outputs $\{y(k)\}$ are available at slower frequency (with sampling interval $p \cdot h$, where $p > 1$ is an integer).

$$\begin{Bmatrix} u(0) & u(1) & \cdots & u(p-1 & u(p) & u(p+1) & \cdots & u(2p) & \cdots \\ y^s(0) & y^s & \cdots & y^s(p-1) & y^s(p) & & \cdots & y^s(2p) & \cdots \\ y(0) & & \cdots & & y(1) & & \cdots & y(2) & \cdots \end{Bmatrix} \quad (1)$$

From such data, it is impossible to directly identify a process quality model at the fast sampling rate. Several approaches have been developed to obtain a smooth intersample prediction and an appropriate model at the fast sample rate h. Reported approaches include numerical interpolation,[3−5] polynomial transformation,[6,7] data lifting,[8,9] and weighted regression[10] methods. For a nonuniformly sampled multirate (NUSM) system, Li et al.[11] propose a novel subspace identification to directly identify a residual model. The development of a Kalman filter and applications to fault detection and isolation (FDI) are also reported.[12,13] This paper focuses on the uniformly sampled multirate data set.

The numerical interpolation approach inserts predictions of the unavailable outputs such that the data set has a uniform sampling interval h. In contrast, data lifting develops a unirate data set at the slow sampling rate $p \cdot h$. A model in the lifted time domain is first identified, from which a model of a fast sampling domain is extracted based on the relationship between the original and lifted models. For data lifting, the key issue is how to reliably extract the fast sampling model. The direct extraction of the fast system might lead to a biased solution or even an infeasible model, i.e., with imaginary elements in the system matrix. A modified data lifting approach is proposed in this paper.

Since primary outputs are only available at the sampling instant $k = p \cdot i$, Lu and Fisher[6] proposed a polynomial transformation approach, where a predefined polynomial is multiplied on both sides of the original model such that the intersample outputs are no longer required for parameter estimation. Lin et al.[10] applied a weighted partial least-squares (WPLS) approach to develop quality estimators from multirate sampled data. A weighting value of 1 is applied to process measurements that correspond to the time instant when a new quality measurement is obtained and 0 to others. A soft sensor with a fast sampling period is derived by applying the developed regression relationship to the regressor matrix.

The main contribution of the paper is to present an improved data lifting approach and to evaluate the proposed methodology using both simulation and operating data from a cement plant. First, the extraction of the fast model is reformulated as an unconstrained optimization problem to eliminate the possibility of obtaining infeasible solutions. Second, regularization is

* To whom correspondence should be addressed. Phone: +45 4525 2872. Fax: +45 4593 2906. E-mail: sbj@kt.dtu.dk.
[†] Technical University of Denmark.
[‡] FLSmidth Automation.
[§] Current address: FLSmidth Automation, Bethlehem, PA 18017.

10.1021/ie801084e CCC: $40.75
Published on Web 05/07/2009

introduced into the objective function to enforce desired model properties. The proposed methodology is also compared with reported approaches using several case studies, regarding numerical issues, accuracy of extracted fast models, and robustness to process noise.

Section 2 reviews the numerical interpolation, polynomial transformation, and WPLS and describes regularized data lifting techniques based on the proposed optimization reformulation. A linear SISO system is used to evaluate the accuracy and robustness of the reported approaches. The regularized data lifting and WPLS approaches are compared with illustrative examples related to the design of free lime soft sensors for cement kiln processes using simulation and operating data. Conclusions are drawn in section 4.

2. Intersample Estimation Techniques

Numerical Interpolation. An intuitive solution is to insert unavailable values of output through interpolation such that a unirate data set at the fast sample frequency is available. Isaksson[3] investigated the applicability of linear interpolation for system identification problems. Ramachandran et al.[5] compared two numerical techniques, moving center Taylor series and polynomial prediction techniques, to predict unavailable output measurements. Amirthalingam and Lee[4] also suggested filling in the missing measurement values using interpolation schemes. Linear interpolation is a quick and straightforward way to predict the unavailable outputs between $[y(k\cdot p), y((k+1)\cdot p)]$. However, linear interpolation is often not sufficiently precise. In addition, the interpolant is not differentiable at the boundary points. Instead, spline interpolation approaches are utilized with a low-degree polynomial in each long sampling interval. Cubic spline is used in this study.

Polynomial Transformation. Lu and Fisher[6] proposed a least-squares output estimation approach for multirate data sets. A polynomial transformation is performed on the original system such that the intersample values of the output are no longer needed for parameter estimation. Assume the model with sampling interval h is

$$A(q^{-1})y(t) = B(q^{-1})u(t) \tag{2}$$

where

$$\begin{aligned} A(q^{-1}) &= 1 + a_1q^{-1} + a_2q^{-2} + \cdots + a_nq^{-n} \\ B(q^{-1}) &= b_1q^{-1} + b_2q^{-2} + \cdots + b_nq^{-n} \end{aligned} \tag{3}$$

Given λ_i are the roots of $A(q^{-1})$, then

$$A(q^{-1}) = \prod_{i=1}^{n} 1 - (\lambda_i q)^{-1} \tag{4}$$

Multiplying eq 4 on both sides with the following polynomial

$$P(q^{-1}) = \prod_{i=1}^{n} [1 + (\lambda_i q)^{-1} + (\lambda_i q)^{-2} + \cdots + (\lambda_i q)^{p-1}] \tag{5}$$

yields

$$A_p(q^{-p})y(t) = B_p(q^{-1})u(t) \tag{6}$$

where

$$A_p(q^{-p}) = \prod_{i=1}^{n} 1 - (\lambda_i q)^{-p} = 1 + a_{p,1}q^{-p} + a_{p,2}q^{-2p} + \cdots + a_{p,n}q^{-np} \tag{7}$$

$$B_p(q^{-1}) = b_{p,1}q^{-1} + b_{p,2}q^{-2} + \cdots + b_{p,np}q^{-np} \tag{8}$$

Thus, only the infrequently sampled outputs, $\{y(k\cdot p)\}$, and fast inputs are used to estimate the following parameters involved in $A_p(q^{-p})$ and $B_p(q^{-1})$

$$\theta = [a_{p,1} \quad a_{p,2} \quad \cdots \quad a_{p,n} \quad b_{p,1} \quad b_{p,2} \quad \cdots \quad b_{p,np}] \tag{9}$$

Given the relationship between eqs 4 and 7, $A(q^{-1})$ can be extracted from $A_p(q^{-p})$, followed by calculation of $B(q^{-1})$ based on $P(q^{-1})$ and $B_p(q^{-1})$ through polynomial division.

Data Lifting Technique. Process measurements $\{u(i)\}$ are sampled with the interval h, while the primary output $\{y(k)\}$ has a sampling period $p\cdot h$. Data lifting[14] reorganizes the original data set by stacking the fast sampled variables. The lifting operator is defined as

$$\tilde{\mathbf{u}} = L_p\mathbf{u} = \left\{ \underbrace{\begin{bmatrix} u(1) \\ u(2) \\ \vdots \\ u(p) \end{bmatrix}}_{\tilde{\mathbf{u}}(1)}, \underbrace{\begin{bmatrix} u(p+1) \\ u(p+2) \\ \vdots \\ u(2*p) \end{bmatrix}}_{\tilde{\mathbf{u}}(2)}, \ldots, \underbrace{\begin{bmatrix} u(k*p+1) \\ u(k*p+2) \\ \vdots \\ u((k+1)*p) \end{bmatrix}}_{\tilde{\mathbf{u}}(k)}, \ldots \right\} \tag{10}$$

Then, the lifted input sequence $\tilde{u}$ has the same period as the slowly sampled output. A process model with long sampling period, $p\cdot h$, can be identified straightforwardly. In order to obtain the intersample value of the slow-sampled process measurements, $\{\hat{y}(k)\}$, a fast model is extracted according to the relationship between the slow and the fast system as described in the sequel.

Given a discrete linear system with a sampling period of h

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k \\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k \end{aligned} \tag{11}$$

The following lifted model is obtained with a multirate data set

$$\begin{aligned} \mathbf{x}_{k+1}^L &= \bar{\mathbf{A}}\mathbf{x}_k^L + \bar{\mathbf{B}}\mathbf{u}_k^L \\ \mathbf{y}_k^L &= \bar{\mathbf{C}}\mathbf{x}_k^L \end{aligned} \tag{12}$$

The relationship between parameters can be derived

$$\begin{aligned} \bar{\mathbf{A}} &= \mathbf{A}^p \\ \bar{\mathbf{B}} &= [\mathbf{A}^{p-1}\mathbf{B} \quad \mathbf{A}^{p-2}\mathbf{B} \quad \cdots \quad \mathbf{AB} \quad \mathbf{B}] \\ \bar{\mathbf{C}} &= \mathbf{C} \end{aligned} \tag{13}$$

It is straightforward to extract matrices $\mathbf{B}$ and $\mathbf{C}$ from the lifted system. The system matrix $\mathbf{A}$ can be obtained using one of the following two approaches.[9] The first approach starts from the equation $\bar{\mathbf{A}} = \mathbf{A}^p$, where $\mathbf{A}$ is calculated directly as $\mathbf{A} = \bar{\mathbf{A}}^{1/p}$. It is necessary to obtain a real-valued matrix. Therefore, $\mathbf{A}$ is approximated with the real part of $\bar{\mathbf{A}}^{1/p}$. That is, $\mathbf{A} = \mathrm{Re}(\bar{\mathbf{A}}^{1/p})$. The second approach derives $\mathbf{A}$ through a linear regression. Given

$$\bar{\mathbf{B}} = [\mathbf{A}^{p-1}\mathbf{B} \quad \cdots \quad \mathbf{AB} \quad \mathbf{B}] \tag{14}$$

and

$$\bar{\mathbf{B}}_i = \mathbf{A}^{i-1}\mathbf{B} \quad i = 1, 2, \cdots, p \tag{15}$$

which is the ith block of $\bar{\mathbf{B}}$ from the right. The following relationships hold

$$\begin{aligned} \bar{\mathbf{B}}_{i+1} &= \mathbf{A}\bar{\mathbf{B}}_i, \quad i = 1, 2, \cdots, p-1 \\ \bar{\mathbf{A}}\mathbf{B} &= \mathbf{A}\bar{\mathbf{B}}_p \end{aligned} \tag{16}$$

rewriting in a matrix form assuming that $\mathbf{\Psi\Psi}^T$ is invertible

$$\underbrace{[\overline{\mathbf{AB}} \quad \overline{\mathbf{B}}_p \quad \overline{\mathbf{B}}_{p-1} \quad \cdots \quad \overline{\mathbf{B}}_2]}_{\Gamma} = \mathbf{A}\underbrace{[\overline{\mathbf{B}}_p \quad \overline{\mathbf{B}}_{p-1} \quad \overline{\mathbf{B}}_{p-2} \quad \cdots \quad \overline{\mathbf{B}}_1]}_{\Psi} \qquad (17)$$

$$\mathbf{A} = (\Gamma\Psi^T)(\Psi\Psi^T)^{-1} \qquad (18)$$

If $\mathbf{\Psi\Psi}^T$ is rank deficient, i.e., insufficient input excitation, a pseudoinverse may be used.

Two modifications are incorporated into the original data lifting technique: reformulating the extraction of the fast model as an unconstrained optimization problem to eliminate the possibility of obtaining infeasible solutions and adding the regularization term to the objective function to enforce desired model properties.

Modified Data Lifting Approach. Since the direct extraction of the fast system might lead to an infeasible system, i.e., with imaginary elements in the system matrix, the extraction of a fast model is reformulated within an optimization framework, where a penalty term is added to eliminate the possibility of obtaining infeasible solutions. Assume the system matrix of the fast system A is expressed as

$$\mathbf{A} = \mathrm{Re}(\mathbf{A}) + i\mathrm{Im}(\mathbf{A}) \qquad (19)$$

the objective function is expressed as the deviation between $\mathbf{A}^p$ and $\bar{\mathbf{A}}$, plus the norm of the imaginary part of $\mathbf{A}$

$$\arg\min_{\mathrm{A}}(\|\mathbf{A}^p - \bar{\mathbf{A}}\|_l + \lambda\|\mathrm{Im}(\mathbf{A})\|_l) \qquad (20)$$

where λ is a positive real scalar, $\| \|_l$ can be taken as the 1, 2, or ∞ norm.

The objective function can also be defined as the deviation between the impulse response of the lifted and fast models

$$\arg\min_{\mathbf{A}} \sum_i (h^p(i) - h(p\cdot i))^2 \qquad (21)$$

The regularization technique has been widely used to address ill-conditioned identification problems due to the lack of excitation in process data. In addition to reducing the variance of the model parameter estimate, regularization techniques can be employed to enforce that the estimated model possesses some desired properties. One efficient method to incorporate model properties into least-squares estimation is Tikhonov Regularization (TR).[15] Bonné[16] applies the TR technique in the identification of an interdependent grid of linear models (GoLM) for batch processes. Model parameters are estimated by solving the TR problem

$$\hat{\theta}_{TR} = \arg\min(\|\mathbf{Y} - \mathbf{X}\theta\|^2_{\mathbf{W}} + \|\mathbf{L}\theta\|^2_{\Lambda^2}) \qquad (22)$$

where the structured penalty matrix $\mathbf{L}$ maps the parameter vector θ into the desired properties. For example, a smooth impulse response of the estimated model is enforced by defining $\mathbf{L}\theta$ to approximate the second-order derivative of the impulse response

$$\|h(i-1) - 2h(i) + h(i+1)\| \qquad (23)$$

The objective function to extract fast model is defined as

$$\arg\min_{\mathbf{A}} \sum_i (h^p(i) - h(p\cdot i))^2 + \lambda\sum_i (h(i-1) - 2h(i) + h(i+1))^2 \qquad (24)$$

Enforcing model properties inevitably introduces bias into the model parameter estimates, which is a function of the weighting factor λ. Therefore, there will be a trade off between the bias and variance of the model parameter estimates. This trade off is to be used to provide optimal predictive capability of the estimated model by determining λ on validation data.

The extraction of fast model is reformulated as an unconstrained optimization problem, which can be solved with any minimizer.

WPLS Approach. Since the primary outputs are sampled at a slower rate than other process measurements, a zero-order hold is commonly used, i.e., the constant value is inserted in the log system until a new laboratory analysis is obtained. Therefore, a weighting vector $\{w(k)\}$

$$w(k) = \begin{cases} 1 & y(k) \neq y(k-1) \\ 0 & \text{otherwise} \end{cases} \qquad (25)$$

is used to downweight the intersample value of the primary output as follows

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ \vdots \end{bmatrix} * \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(p-1) \\ y(p) \\ y(p+1) \\ \vdots \\ y(k) \\ \vdots \end{bmatrix} = f\left(\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ \vdots \end{bmatrix} * \begin{bmatrix} u(1) \\ u(2) \\ \vdots \\ u(p-1) \\ u(p) \\ u(p+1) \\ \vdots \\ u(k) \\ \vdots \end{bmatrix}\right) \qquad (26)$$

where the asterisk "*" defines element by element production. A regression relationship is developed using the weighted matrices of regressors and the dependent variable, which is then applied to all available regressor samples to obtain intersample estimation.

PLS addresses the collinearity that commonly exists among process measurements by projection onto a lower dimensional subspace. A second advantage of the PLS regression is its inherent regularization via dimension reduction to stabilize the solution of ill-conditioned problems and to limit the overfit of data.[17]

3. Case Studies

The numerical interpolation, polynomial transformation, and data lifting are evaluated with a linear SISO system. The WPLS and the data lifting techniques are compared on designing free lime soft sensors of cement kiln processes with both simulation and operation data.

SISO System. The first test case is a simple SISO system[6]

$$(1 - 1.5q^{-1} + 0.7q^{-2})y(t) = (q^{-1} + 0.5q^{-2})u(t) \qquad (27)$$

The following comparisons are performed: noise free simulation and white noise with a noise-to-signal ratio (NSR) of 0.10, 0.15, and 0.20 added to the output.

For each case, the following p values are evaluated: 2, 4, 6, and 10. The modeling and validation data blocks both contain 3000 samples. Since the extracted fast models of the proposed data lifting approach depend on the choice of weight coefficient, λ, the values of 0, 10^{-2}, and 10^{-1} are evaluated.

The optimal model is obtained by minimizing the prediction error of the validation data

$$\mathrm{PRESS} = \sum_{i=1} (\hat{y}(p\cdot i) - y_m(p\cdot i))^2 \qquad (28)$$

where $\hat{y}(p\cdot i)$ denotes the one-step-ahead prediction. The performance is then evaluated with the sum of squared errors

Table 1. SSES of the SISO System with Noise-Free Data

p	numerical interpolation	polynomial transformation	lifting technique $\lambda = 0$	lifting technique $\lambda = 10^{-2}$	lifting technique $\lambda = 10^{-1}$
2	0.20 ± 0.05	3.3e-7 ± 5.4e-7	1.5e-3 ± 2.3e-3	3.0e-3 ± 3.5e-3	0.12 ± 0.02
4	8.38 ± 3.67	6.5e-6 ± 1.8e-5	2.4e-3 ± 3.7e-3	4.0e-3 ± 4.6e-3	0.12 ± 0.03
6	1.6e2 ± 1.1e2	2.0e-3 ± 3.8e-3	2.1e-3 ± 1.0e-2	3.5e-2 ± 8.9e-3	0.13 ± 0.02
10	5.2e2 ± 2.9e2	8.3e3 ± 3.6e2	0.06 ± 0.10	0.05 ± 0.08	0.10 ± 0.04

between the step responses (SSES) of the extracted fast model ($\tilde{S}(i)$) and the original system ($S(i)$), defined as

$$\mathrm{SSES} = \sum_{i=1}^{N_S} (\tilde{S}(i) - S(i))^2 \tag{29}$$

where the step response time $N_S = 30$. For noise-free data, 100 Monte Carlo simulations are performed. The mean SSES and the standard deviation of 100 runs are summarized in Table 1.

The numerical interpolation approach inserts unavailable values between output samples with a spline function. The larger the ratio between the slow and the fast sampling interval, the larger the interpolation error. When $p = 2$, 4, and 6, SSES of the numerical interpolation approach are significantly larger than those of the polynomial transformation and the data lifting technique.

The SSES of models derived with the polynomial transformation method are smallest when $p = 2$, 4, and 6. However, the values are excessively large when $p = 10$, which is caused by numerical difficulties. First, directly extracting $A(q^{-1}) = \prod_{i=1}^{n} 1 - (\lambda_i q)^{-1}$ from $A_p(q^{-p}) = \prod_{i=1}^{n} 1 - (\lambda_i q)^{-p}$ might be erroneous when p is large. For example, this procedure may lead to unpaired imaginary roots λ_i that result in $A(q^{-1})$ with complex coefficients. In order to obtain a feasible model ($A(q^{-1})$ with all real parameters), it is necessary to perform an approximation. In this case study, the real part of the coefficients is kept and the imaginary part is neglected, which inevitably degrades the estimation of $A(q^{-1})$. Since the polynomial $P(q^{-1})$ is constructed based on the eigenvalues of $A(q^{-1})$, the approximation error in obtaining $A(q^{-1})$ then propagates in the second step of extracting $B(q^{-1})$ from $B_p(q^{-1})$. Therefore, the sequential procedure in extracting the fast model from $A_p(q^{-p})$ and $B_p(q^{-1})$ leads to degraded models when the ratio between the slow and fast sampling intervals is large. Reliable rational function approximation approaches should be used.

The weighting factor $\lambda = 10^{-1}$ leads to biased models since the second term of eq 24 is dominant in the objective function. The result of $\lambda = 10^{-2}$ is comparable with that of

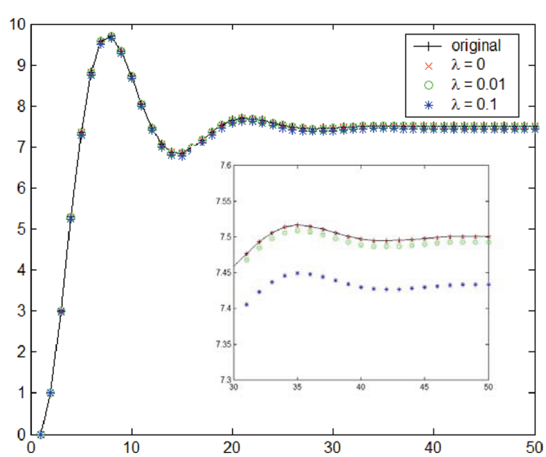

Figure 1. Step response of the fast model extracted with different λ values ($p = 10$).

Table 2. SSES of the the SISO System with Data at Different Noise Levels

p	NSR	polynomial transformation	data lifting technique $\lambda = 0$	data lifting technique $\lambda = 10^{-2}$	data lifting technique $\lambda = 10^{-1}$
	0.10	3.20 ± 4.09	0.025 ± 0.008	0.026 ± 0.008	0.148 ± 0.009
2	0.15	2.79 ± 3.34	0.049 ± 0.002	0.049 ± 0.002	0.159 ± 0.017
	0.20	5.29 ± 8.04	0.124 ± 0.026	0.126 ± 0.026	0.251 ± 0.080
	0.10	2.93 ± 5.07	0.045 ± 0.048	0.046 ± 0.051	0.168 ± 0.144
4	0.15	9.55 ± 12.58	0.122 ± 0.131	0.121 ± 0.129	0.230 ± 0.237
	0.20	23.7 ± 26.1	0.190 ± 0.190	0.189 ± 0.180	0.295 ± 0.251
	0.10	1.1e3 ± 1.1e3	0.069 ± 0.070	0.070 ± 0.077	0.194 ± 0.166
6	0.15	1.2e3 ± 0.9e3	0.136 ± 0.152	0.132 ± 0.138	0.211 ± 0.160
	0.20	1.6e3 ± 1.1e3	0.192 ± 0.190	0.183 ± 0.169	0.236 ± 0.192

$\lambda = 0$ for $p = 2$ and 4. The step response of an extracted model ($p = 10$) is shown in Figure 1. With $\lambda = 10^{-1}$, a biased model is obtained with a small but visible steady-state error. A zoomed plot reveals that $\lambda = 10^{-2}$ leads to an improved model compared to $\lambda = 10^{-1}$.

The investigation with noise-free data reveals the unsatisfactory performance of the numerical interpolation approach, which is not considered in the case study of simulated data with different noise levels. One hundred Monte Carlo simulations are performed for each case.

Table 2 reveals that data lifting is more robust to process noise. When $p = 2$ and NSR = 0.1, the mean SSES of the polynomial transformation is 3.20, which is much higher than that of the lifting technique, varying from 0.025 ($\lambda = 0$) to 0.148 ($\lambda = 0.1$). When $p = 6$, the mean SSES of the polynomial transformation is over 1000 while that of regularized lifting technique is less than 0.25.

When the noise level in the data is small, i.e., NSR = 0.10, the regularization term leads to a noticeably biased model. For example, the mean SSES with $\lambda = 10^{-1}$ is 0.148 at $p = 2$, which is about 6 times that of the model without regularization ($\lambda = 0$, SSES = 0.025). When NSR = 0.20 and $p = 2$, the bias caused by the regularization term is about 200% (0.251 vs 0.124). When $p = 4$ and 6 and NSR = 0.15 and 0.20, the mean SSES obtained with the regularization term of $\lambda = 10^{-2}$ is slightly smaller than that of $\lambda = 0$. The mean SSES of $\lambda = 10^{-1}$ is larger than that of $\lambda = 10^{-2}$ and 0. This investigation clearly demonstrates the trade-off effect of the regularization term in the data lifting technique.

The step response of a fast model extracted with the regularized data lifting approach is compared with that of the original system. Figure 2 plots the relationship between the SSES and the regularization coefficient λ.

In the case of noise-free data, SSES increases with λ due to a biased estimation. In the case of noise contamination, L-shaped curves are obtained. An optimal λ value can be clearly identified. For NSR = 0.10, the optimal λ value is around 10^{-2}. When NSR increases to 0.20, a larger λ is preferred. As shown in Figure 2, the knee point of the curve is at 10^{-1}. This investigation demonstrates the necessity of regularization, especially for noisy data, such as those collected from most chemical processes. For industrial applications, the optimal λ can be determined from validation data and is dependent on the actual NSR.

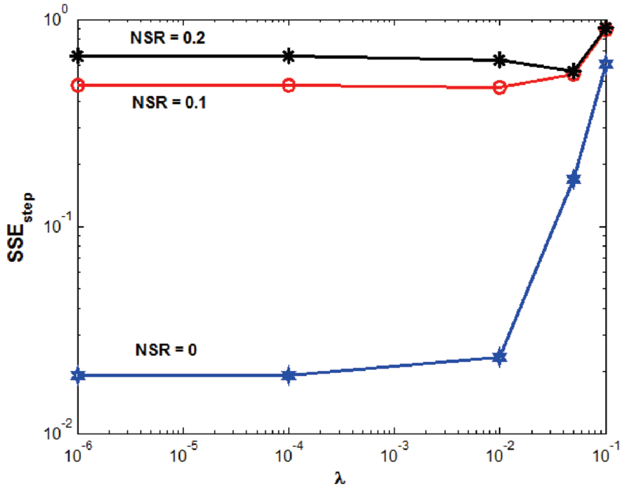

Figure 2. Relationship between the SSES and regularization coefficient λ ($p = 10$).

Since the polynomial transformation approach is based on the transfer function between each input and output, extension to the MIMO system is not as straightforward as the data lifting technique. It is necessary to identify a set of MISO systems to obtain the MIMO model. In addition, the number of independent variables increases significantly with the dimensions of inputs/outputs.

In summary, the data lifting technique outperforms the numerical interpolation and polynomial transformation approaches regarding the robustness to process noise and the straightforwardness to handle MIMO systems. The next two case studies compare it with the WPLS approach for designing free lime soft sensors using simulation and operation data of cement kiln processes.

Cement Kiln Process. The rotary kiln is operationally the most complex and energy-consuming equipment in the cement industry. For most dry processes (as shown in Figure 3), the feed materials are preheated by hot gas from the rotary kiln. A fuel combustion chamber, called a precalciner, is integrated in the preheating tower to improve energy efficiency. The mixture of preheated and precalcined materials enters the rotary kiln, where fuel together with air enters from the opposite end. Several exothermic and endothermic reactions take place in both solid and gas phases. The solid feed is heated to an extremely high temperature (about 1500 °C) in the burning zone such that raw materials react and form the nodular clinker. The clinker exits the kiln at about 1200 °C; then it is cooled down by cross-flowing air in a separate clinker cooler. Partial heat integration is achieved by feeding part of the heated air back into the kiln and part to the precalciner.

The product quality of a cement kiln is indicated by the amount of free lime (CaO) in clinker. The direct off-line measurement is commonly available once every 1 or 2 h with a time delay of about 1 h due to the residence time of the cooler, sampling, and analysis in the laboratory. The measurement is also very sensitive to operating perturbations within the kiln system, which result in uncertain indication of the average quality. Therefore, it is desirable to develop a soft sensor that can accurately predict the content of CaO in the burning zone real timely for effective quality control.

Case Study with Simulation Data. Data are collected from a cement kiln process simulator, Cemulator,[18] which is based on first-principles models[19] and solved with gPROMS.[20] Pseudo Random Binary Sequence (PRBS) signals are applied to the kiln feed, fuel, and rotary speed, fuel to the precalciner, tertiary air damper, and induced draft fan power to sufficiently excite the process. More importantly, the free lime is calculated based on a first-principles model, which renders it possible to compare the deviation of the smoothed intersample prediction with the true simulated value.

The 10 min averaged process data are used to derive the free lime soft sensor, which include flow rates of feed and fuel, temperatures, and gas composition measurements around the kiln system. A data block of 1200 samples are collected: 900 for modeling and 300 for validation. The first investigation compares the WPLS and the data lifting techniques at $p = 6$, which corresponds to a free lime measurement every hour as in many cement plants.

A CaO soft sensor is developed using the WPLS approach. Weighting vectors are multiplied onto inputs and outputs of modeling data. During the model validation phase, this regression model is applied to all input samples. Therefore, the intersample behavior of the soft sensor can be calculated. Following the procedure described in Lin et al.,[21] a 10th-order

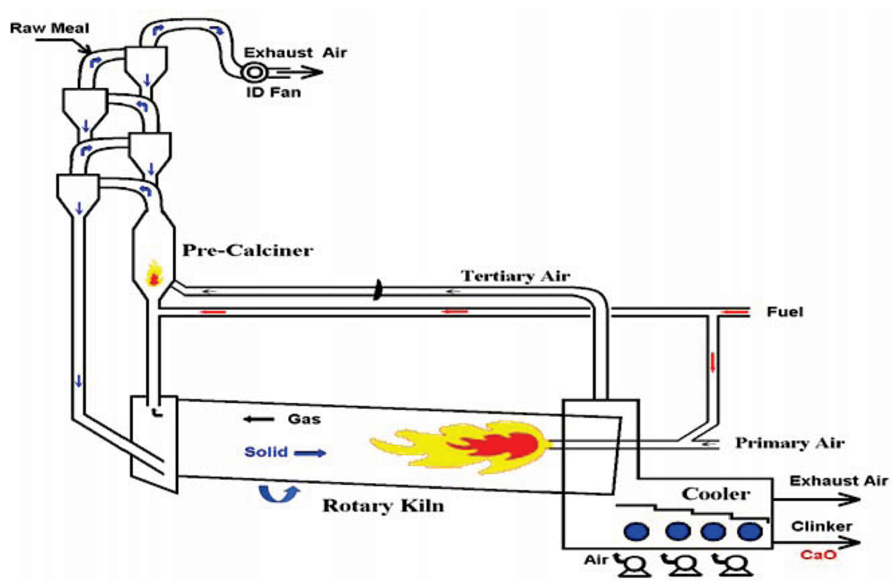

Figure 3. Dry cement kiln process.

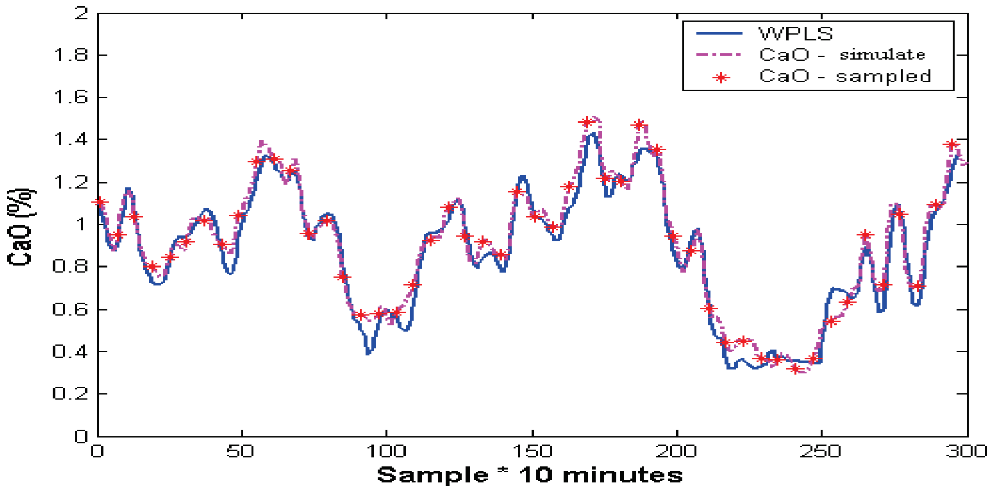

Figure 4. Free lime estimation with 10th-order WPLS model with three latent variables during the validation period.

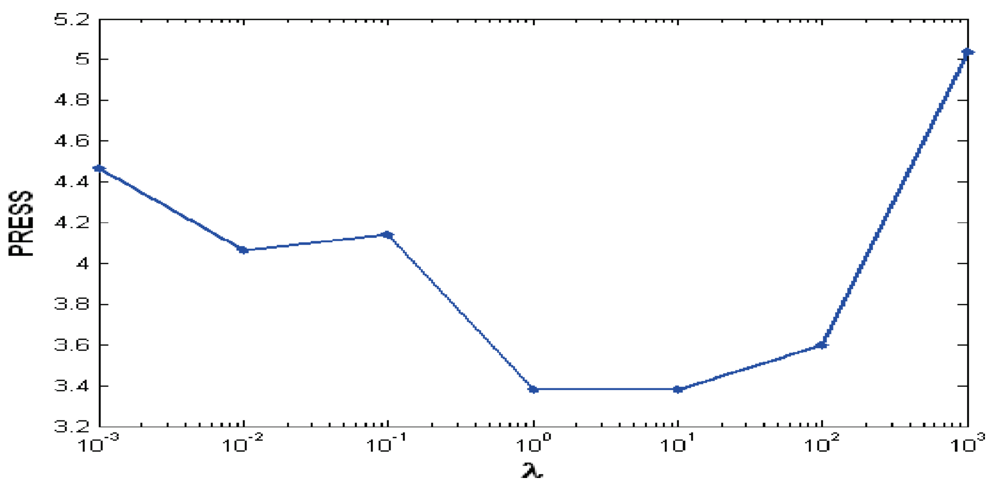

Figure 5. Relationship between PRESS and the regularization coefficient λ.

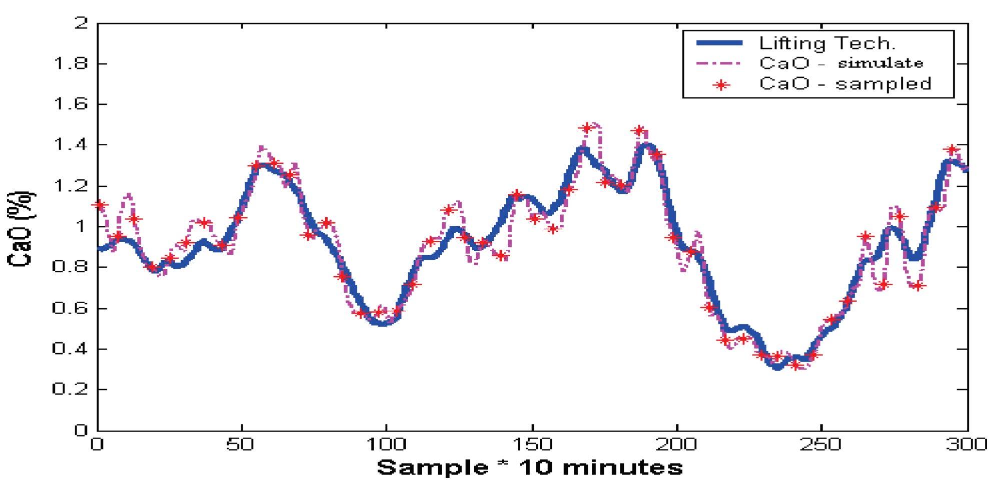

Figure 6. Free lime estimation with regularized data lifting technique in the validation period.

dynamic PLS model with three latent variables achieves the minimum PRESS value of 3.21. Figure 4 shows the calculated CaO (dotted line), sampled CaO used to derive the soft sensor (star), and model prediction (solid line). The WPLS soft sensor provides reasonable one-step-ahead predictions.

A second free lime soft sensor is derived with the regularized data lifting technique, with regularization parameters from 0.001 to 1000 varying by a factor of 10. The average PRESS of 100 Monte Carlo simulations of each regularization parameter are shown in Figure 5, which reveals the trade-off effect of the regularization parameter. An optimal value is expected to be in the range between 1 and 10.

As shown in Figure 6, the free lime soft sensor of the regularized data lifting technique is able to predict the trend of the quality measurement reasonably well. However, deviations between the calculated free lime (dotted line) and the model prediction are slightly larger than those of the WPLS model, especially when fast variations occur in the process between

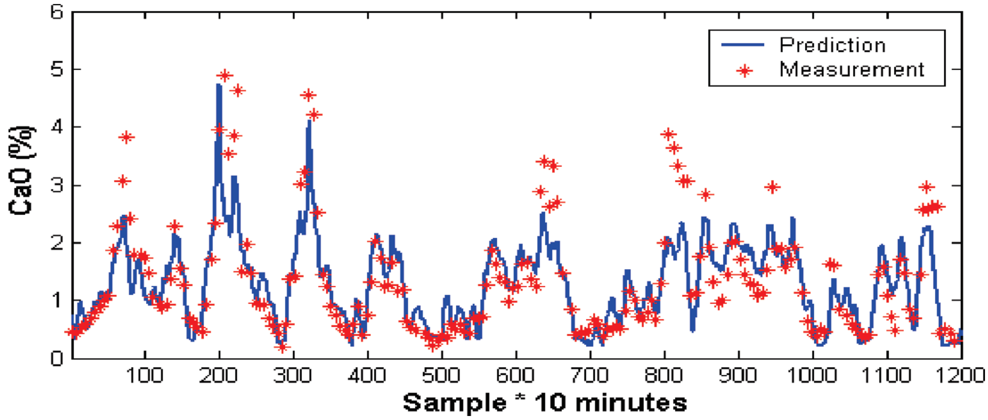

Figure 7. Free lime estimation with the WPLS approach during the modeling period.

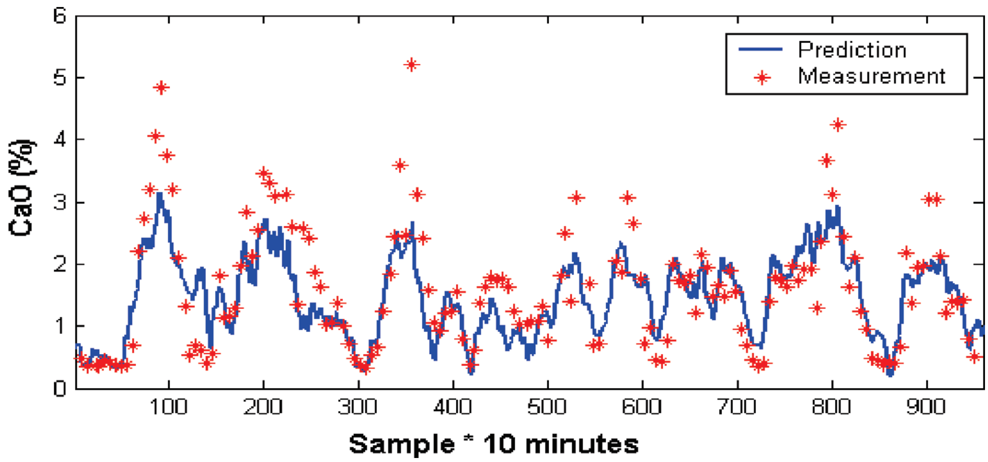

Figure 8. Free lime estimation with the WPLS approach during the validation period (PRESS = 81.19).

samples 270 and 290. It is clear from this comparison on simulated data that WPLS is better to follow higher frequency dynamics than the regularized data lifting approach.

Case Study with Operational Data. Operating data from a cement kiln log system are used to derive a soft sensor for free lime in the clinker. There are in total 39 process measurements logged every 10 min, including kiln feed, kiln motor power, fuel flow rates to calciner and kiln, plus several temperature measurements within the kiln system. Twelve process variables are selected as inputs based on process knowledge as well as considering the reliability of process measurements. The 10 min averaged standard measurements are used as regressors to estimate the free lime, which is logged approximately every 1 h. One thousand two hundred samples are used to derive the soft sensor, which is then validated on another 900 samples.

The 12th-order dynamic PLS model with two latent variables is obtained through the cross-validation procedure described in Lin et al.[21] The comparison between the laboratory measurement and model prediction during the modeling and validation periods are shown in Figures 7 and 8, respectively.

Large deviations between the model prediction and process measurement are observed between samples 800 and 850. On the basis of the results of data analysis and discussions with experienced process engineers, the deviations are most possibly caused by two factors: the missing of key measurements and nonlinearity. First, the chemistry of raw feed and the fuel characteristics of coal burned in the kiln vary continuously. In most cement plants, both measurements are available once a day or even less frequently. It is thus impossible in this case study to pinpoint the exact contribution of each factor. It is also difficult to incorporate the information into the soft sensor. Second, although it is observed that a linear model describes the nominal responses of a cement kiln process reasonably well, it is well known that the response of the process is extremely nonlinear when the process is operated in a zone away from the nominal conditions, for example, when the free lime is very high (>3.5%) or low (<0.5%). Furthermore, a linear PLS model has a limited extropolability, which might not be able to precisely predict the peak values caused by large disturbances to normal operations. Several nonlinear regression models are also evaluated using this data set. However, the improvements over a linear model are marginal. It might be due to the limited availability of the data and the increased number of parameters involved in nonlinear models, comparing to linear ones.

In comparison with the simulation data, the deviation between the laboratory measurement and the WPLS model prediction is noticeably larger. There are several reasons. First, plant data is obtained during normal operation that is not as informative as simulator data, which is sufficiently excited with the designed PRBS signals. Second, the chemical compositions of the feed and fuels are continuously varying in cement plant, while white-noise-like disturbances are applied to the simulator. In the cement plant, other process equipment and instrument introduce frequent disturbances to normal operation. Incorporating non-linear functionalities into the soft sensor model may also be helpful.

A second free lime soft sensor is derived with the regularized data lifting technique. The modeling and validation results are shown in Figures 9 and 10. It should be noted that Figure 10 reveals significant variations of product quality in the plant due

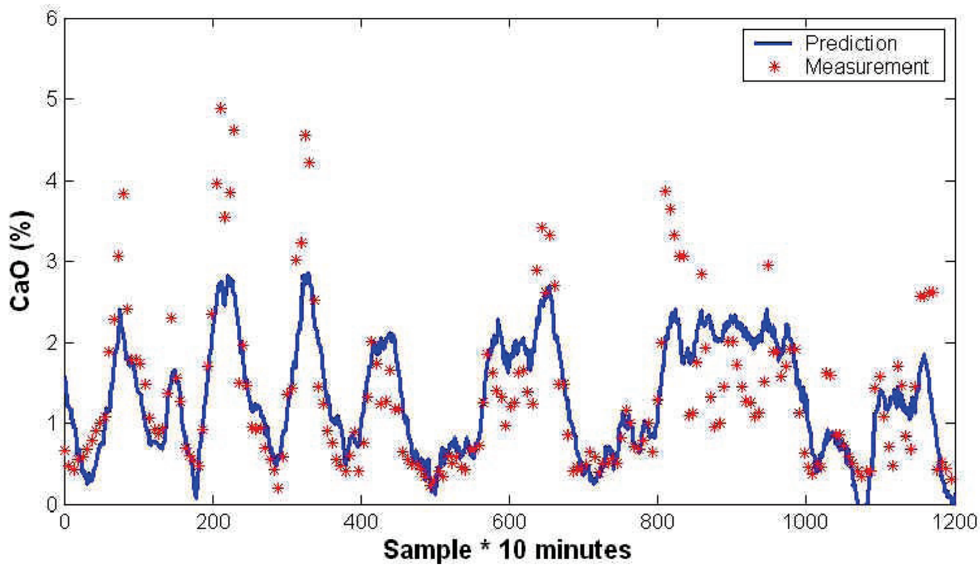

Figure 9. Free lime estimation with the regularized data lifting technique during the modeling period.

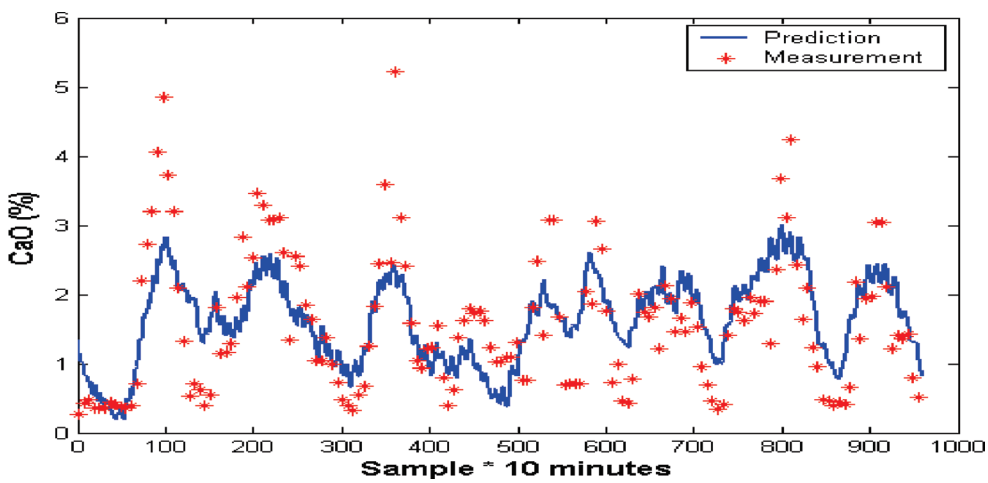

Figure 10. Free lime estimation with regularized data lifting technique during validation period (PRESS = 85.78).

to the lack of real-time information of product quality. The off-line investigations demonstrate that the soft sensor correctly captures the trend of variations in laboratory measurement, which is the key requirement of a predictive quality model. Although deviations between the laboratory measurement and the model prediction are observed during several periods, the integration of a predictive quality model enables the process control system to take prompt actions to achieve improved quality control, i.e., preventing off-specification products and minimizing the variations of the product quality.

Figures 8 and 10 reveal that the model obtained with the regularized data lifting approach is comparable to that of the WPLS. As described previously, the regularized data lifting approach consists of two steps: a state space model with a slow sampling interval is first identified, from which a fast model is then extracted. It should be noted that the lifting technique involves a larger number of model parameters than the WPLS approach. It is difficult to obtain an accurate lifted model when the ratio between long/short sampling intervals is large and the available data is limited. Figure 10 shows that the free lime soft sensor obtained with the regularized data lifting approach is not as smooth as that of the WPLS. The intersample prediction of the state space model obtained through the regularized lifting approach could be improved with increasing data length.

4. Conclusions

This paper evaluates several approaches to develop soft sensors from multirate sampled data sets. The original data lifting approach are improved: reformulating the extraction of a fast model as an optimization problem and ensuring the desired model properties through Tikhonov Regularization. Comparative investigations are performed with a SISO system and free lime soft sensors for cement kiln processes using data collected from a simulator and a plant log system.

The numerical interpolation technique is only applicable to data in which the slow sampling interval is still relatively fast compared to the process dynamics. The polynomial transformation approach is a feasible choice if the ratio between the slow and fast sampling intervals is small and the noise level of the data is insignificant. Case studies of a cement kiln process reveal that soft sensor designed with the WPLS approach provides reasonably accurate one-step-ahead prediction, which can be a good candidate for open-loop advisory performance indicator for kiln operation.

The modified data lifting technique including explicit regularization demonstrates improved performance in extracting a fast model. Obtaining a feasible fast system is ensured and a smooth prediction is enforced through regularization. Preliminary application to cement kiln process demonstrates that the soft sensor is able to model the trend of the quality measurement, demonstrating potential to be used as a predictive quality model for effective quality control and optimizing process operation.

Literature Cited

(1) Kourti, T. The Process Analytical Technology initiative and multivariate process analysis, monitoring and control. *Anal. Bioanal. Chem.* **2006**, *384* (5), 1043–1048.

(2) Hinz, D. C. Process analytical technologies in the pharmaceutical industry: The FDA's PAT initiative. *Anal. Bioanal. Chem.* **2006**, *384* (5), 1036–1042.

(3) Isaksson, A. J. *On the use of linear interpolation in identification subject to missing data*; Report EE9202; Department of Electrical and Computer Engineering, University of Newcastle: Australia, 1992.

(4) Amirthalingam, R.; Lee, J. H. Subspace identification based inferential control applied to a continuous pulp digester. *J. Process Control* **1999**, *9* (5), 397–406.

(5) Ramachandran, P.; Young, G. E.; Misawa, E. A. Intersample output estimation with multirate sampling. *Proc. IEEE Int. Conf. Control Appl.* **1996**, 576–81.

(6) Lu, W.; Fisher, D. G. Least-squares output estimation with multirate sampling. *IEEE Trans. Automat. Control* **1989**, *34* (6), 669–672.

(7) Fisher, D. G.; Lu, W. Multirate adaptive inferential estimation. *IEEE Proc. Control Theor. Appl.* **1992**, 181–189.

(8) Srinivasaro, M.; Gudi, R. D.; Patwardhan, S. C. *Identification of fast-rate nonlinear output error models from multi-rate data*; IFAC World Congress: Prague, 2005.

(9) Wang, J.; Chen, T.; Huang, B. Multirate sampled-data systems: computing fast-rate models. *J. Process Control* **2004**, *14* (1), 79–88.

(10) Lin, B.; Recke, B.; Knudsen, J.; Jørgensen, S. B. A Systematic Approach for Soft Sensor Development. European Symposium on Computer Aided Process Engineering−15, Barcelona, Spain, 2005.

(11) Li, W.; Han, Z.; Shah, S. L. Subspace Identification for FDI in Systems with Non-uniformly Sampled Multirate Data. *Automatica* **2006**, *42*, 619–627.

(12) Shah, S. L.; Li, W.; Xiao, D. *Kalman Filters for Non-uniformly Sampled Multirate Systems*; 16th IFAC World Congress: Prague, 2005; p 2486.

(13) Li, W.; Shah, S. L.; Xiao, D. Kalman Filters in Non-uniformly Sampled Multirate Systems: For FDI and Beyond. *Automatica* **2008**, *44* (1), 193–198.

(14) Khargonekar, P. P.; Poolla, K.; Tannenbaum, A. Robust control of linear time-invariant plants using periodic compensation. *IEEE Trans. Automatic Control* **1985**, *30* (11), 1088–1096.

(15) Johansen, T. A. On Tikhonov Regularization, Bias and Variance in Nonlinear System Identification. *Automatica* **1997**, *33* (3), 441–446.

(16) Bonné, D. Modelling for Optimal and Reliable Operation of Batch Processes. Ph.D. Thesis, Technical University of Denmark, Lyngby, 2005.

(17) Bennett, K. P.; Embrechts, M. J. An Optimization Perspective on Partial Least Squares. In *Advances in Learning Theory: Methods, Models and Applications*; IOS Press: Amsterdam, 2003; Vol. 227, p 250.

(18) FLSmidth Automation, ECS/CEMulator-the Most Realistic Environment for Operator Training in the Cement Industry. http://www.flsautomation.com/pdf/cemulator.pdf.

(19) Delfter, O. Simulation for Start-up and Operation of a Rotary Dry Process Kiln. Ph.D. Thesis, Technical University of Denmark, 1982.

(20) Process System Enterprise Limited, gPROMS overview. http://www.psenterprise.com/gproms/index.html.

(21) Lin, B.; Recke, B.; Knudsen, J. K. H.; Jørgensen, S. B. A Systematic Approach for Soft Sensor Development. *Comput. Chem. Eng.* **2006**, *31*, 419–425.

Received for review July 15, 2008
Revised manuscript received March 20, 2009
Accepted April 7, 2009

IE801084E