Korean J. Chem. Eng., **29**(9), 1135-1143 (2012)
DOI: 10.1007/s11814-011-0289-9

Development of a novel self-validating soft sensor

Yiqi Liu†, Daoping Huang, Yan Li, and Xuefeng Zhu

School of Automation Science and Engineering, South China University of Technology, Wushan Road, 381, Room (209) Building (3), Guang Zhou 510641, P. R. China
(*Received 4 September 2011 • accepted 12 November 2011*)

Abstract–A self-validating soft sensor is proposed that not only can perform self-diagnostics and self-reconstruction, but also generate a variety of output data types, including the prediction values, input sensors status of soft sensor and the uncertainty values which represent the credibility of soft sensor's output. The input sensors are validated before performing a prediction by principal components analysis (PCA) model. These validated data are then employed for subsequent recursive partial least square (RPLS) prediction. Other than input sensor validation and modeling for prediction, a t-statistic confidence interval is created and the status of input sensors is offered. By using this self-validating soft sensor, we can determine the work condition of the soft sensor and take proper actions in real time. The usefulness of the proposed method is demonstrated through a case study of a wastewater treatment process.

Key words: Self-validating, Soft Sensor, RPLS, Wastewater, Uncertainty

INTRODUCTION

In a chemical process, the main obstacle for effective quality control is the lack of real-time measurements of the critical or quality variables, due to time-delay and slow sampling rates associated with technical or economic reasons. Thus, soft sensors, which refer to a soft approach to infer hard-to-measure variables from other easy-to-measure measurable process variables, have attracted significant attention in the process industry. Currently, partial least squares (PLS) [1,2] method is used as a modeling method for the soft sensors. Also, principal component regression (PCR) method [3], nonlinear PLS method [4], artificial neural network [5,6], support vector machine (SVM) based regression method [7], and so on, are researched as the soft sensor modeling methods. However, since soft sensors infer the measurements of the quality variables through the online measurements of input variables and a process model, if one of the input sensors fails, the soft sensor estimate will deteriorate. Furthermore, because a soft sensor uses multiple sensors as inputs, the probability that one of the sensors fails increases dramatically. Therefore, it is imperative to improve the validating ability of soft sensors. Soft sensor validation is needed to detect failed input instruments and generate validated values based on the physical relationship of input data. Although much research [8-11] has been devoted to fault detection and identification of failed input instruments, little work has been done on fault reconstruction. As a result, there is still faulty input data usage for subsequent soft sensor modeling. Martin [12] proposed a two layer neural net based method for failed input sensor validation. And operators need to alert and replace the failed measurements using process knowledge, once the predicted input is dissimilar with its corresponding measurement more than a predefined tolerance. However, manual generating of validated data is not practical in the chemical industry. Qin [13] proposed PCA based SPE and T^2 for fault identification and fault data reconstruction by minimizing the statistic SPE. Despite this effort to investigate the fault identification and reconstruction problems, there are always fault alarms happening because of using only SPE or T^2 for fault detection. In addition, Yue et al. [14] combined SPE and T^2 with fault detection and identification and got a good performance. Nonetheless, this method is difficult to implement in practice because fault reconstruction is complex and time consuming. Also, the combined index limit is hard to determine. In this paper, the reconstruction approach was a statistic SVI (sensor validation index) by adjusting SPE. In addition, the failed instrument was identified and reconstructed by minimizing SVI, and then leading to a more feasible solution than only SPE or the combined index approach.

Even though a good soft sensor is developed successfully by validating input sensors, its estimation performance always deteriorates as process characteristics change. To efficiently capture the grade-changing characteristics of the processes, much research on recursive methods [15-17] has been exclusively studied. Mu et al. [18] proposed an online dual updating strategy that combines RPLS model updating and model offset updating, to deal with the issues of time variant processes and uncertainty of the process data. In their approach, the PLS model is updated and the current bias is calculated by the sample's value and the output of the PLS once a new sample is available. On the other hand, only a limited number of papers deal with validating response measurement issue when using RPLS. Owing to its usage of response measurement for model updating, in the case of a sample from response measurement with abrupt noise, the model output would be significantly unstable since the PLS model was misled by an abnormal event from the analyzing sample. Therefore, there is a need to validate the response measurement and give the reliability of measurement value. In this work, a *t*-statistics confidence interval was created to ensure the PLS model would not be misled by outliers from online analyzers.

Qin et al. [13] have presented a brief combination of self-validating technique and soft sensors. Unfortunately, it neither provides

†To whom correspondence should be addressed.
E-mail: liuyiqi769@sina.com

useful information about input variables sensors nor offers a good indication the reliability of soft sensors output once response measurement is suffering from abnormal data. In present work, we propose a new self-validating soft sensor. Such a soft sensor had been termed "SEVA (Self-validating) soft sensor," in that they were not only able to identify and reconstruct faulty input sensors for subsequent soft sensor modeling but also provide an indication as to the validity of, or confidence in, the response measured value. In this SEVA soft sensor, with the need for assuring the hardware sensors as inputs for soft sensor modeling in the normal condition, PCA was used to validate the input sensors and reconstruct the faulty sensors' data. Furthermore, the confidence intervals of predictions, which were derived from the uncertainties of response and inputs, were created to avoid inferential models being misled by noises from response measurement, and unlike the traditional soft sensors which have been viewed as a simple signal prediction, this soft sensor generates a variety of data types, including not only prediction values, but also input sensors status and the on-line uncertainty values which represent the credibility of soft sensor's output.

The remaining sections of this paper are organized as follows. Section 2 gives some basic theories on PCA and recursive PLS. Section 3 discusses the implementation of SEVA soft sensors. Section 4 presents the performance of the SEVA soft sensor through a wastewater treatment process. Finally, the work is concluded.

BASIC THEORIES

1. Principal Component Analysis (PCA)

PCA is a statistical method widely used in chemometrics to compress high-dimensional data into a lower-dimensional space, thus making data more comprehensible by extracting essential information. PCA uncovers combinations of the original variables (these combinations are known as latent variables or principal components, PCs) which describe the dominant patterns and the main trends in the data. The PCA model decomposes the input data as a bilinear product of scores and loadings.

$$X=TV_p+\varepsilon, \tag{1}$$

where ε is the residual matrix, which includes mainly noises under normal conditions. $T\in R^{N\times N}$ and $V_p\in R^{N\times p}$ are the score and loading matrices, respectively. Given a new sample vector x, the PCA score, prediction, and residual vectors are given as follows, respectively:

$$t=V_p x, \tag{2}$$

$$\hat{x}=Cx, \tag{3}$$

$$\tilde{x}=(I-C)x, \tag{4}$$

where $C=V_p{}^T V_p$. The sample vector x is projected onto the principal component subspace and the residual subspace, respectively, as follows:

$$x=\hat{x}+\tilde{x}. \tag{5}$$

The statistic SPE (squared prediction error) is defined as a measure of the variations of residual parts of data.

$$SPE(x)=\|\tilde{x}\|^2=x^T(I-C)x. \tag{6}$$

Under normal conditions, the residual portion of the sample is small. Jackson and Mudholkar [19] proposed the calculation of a confident limit for the SPE. If one of the sensors is faulty, the SPE will increase and thus can be used to detect sensor faults. Summarizing, PCA is a very flexible and versatile technique for extracting meaningful information from large amounts of raw data sets, resulting in easy fault detection in a wide range of industrial contexts. By minimizing SPE, we obtained a sensor validity index (SVI), (see Section 3). Before a soft sensor made a prediction, SVI was implemented to assure that the input data for the soft sensor was normal. Through this improvement, the faulty inputs were not only detected but also reconstructed with a more feasible solution.

2. Recursive Partial Least Square (RPLS)

PLS regression has been widely used for constructing soft sensors because of its simplicity and ability to cope with collinear problems. One of the most popular PLS algorithms was proposed by Qin [20]. Given a matrix {X, Y}, {T_p, W, V_p, B, Q} is the parameters that are processed by PLS, and is shown as follows:

$$\{X, Y\}\rightarrow\{T_p, W, V_p, B, Q\} \tag{7}$$

where $T_p\in\Re^{N\times p}$ is the latent variable (LV) matrix, $V_p\in\Re^{M\times p}$ and $Q\in\Re^{L\times p}$ are the loading matrices of X and Y, respectively. B is the diagonal matrix of inner model coefficients. M, N, L and p (≤M) denote the numbers of input variables, samples, output variables and LVs retained in the PLS model, respectively. And the regression coefficient matrices from PLS are:

$$C^{PLS}=(X^TX)+X^TY=W*BQ^T, \tag{8}$$

where

$$W^*=[w_1^*, w_2^*, \ldots w_M^*], \tag{9}$$

$$w_i^*=\prod_{j=1}^{i-1}(I_M-w_jV_{pj}^T)w_i. \tag{10}$$

When a new data pair $\{x_q, y_q\}$ is available, the PLS model is updated using the augmented data matrices. Then it can be used for Eq. (8) to calculate the regression coefficient matrices.

$$X_{NEW}=[V_p^T; x_q^T], Y_{NEW}=[Q^T; y_q^T]. \tag{11}$$

To improve the computational ability of PLS model, we can add a forgetting factor into the above equation.

$$X_{NEW}=[\beta V_p^T; x_q^T], Y_{NEW}=[\beta Q^T; y_q^T]. \tag{12}$$

where $0<\beta\leq 1$.

Estimation performance of soft sensors always deteriorates as process characteristics change. In chemical processes, for example, process characteristics are changed by catalyst deactivation or fouling. Such a situation may deteriorate product quality. Modeling soft sensors using recursive PLS method can overcome process characteristic variation. Furthermore, due to its simplicity, a heavy computation load can be avoided easily. In addition, the response measurement y_q is used for model updating, so there is a need to create confidence intervals to assure it in the normal condition.

IMPLEMENTATION OF SEVA SOFT SENSORS

In this paper, we propose an integrated framework known as SEVA soft sensors that is shown in Fig. 1 and will do the following:

(1) Validate the input sensors before making a prediction. If an

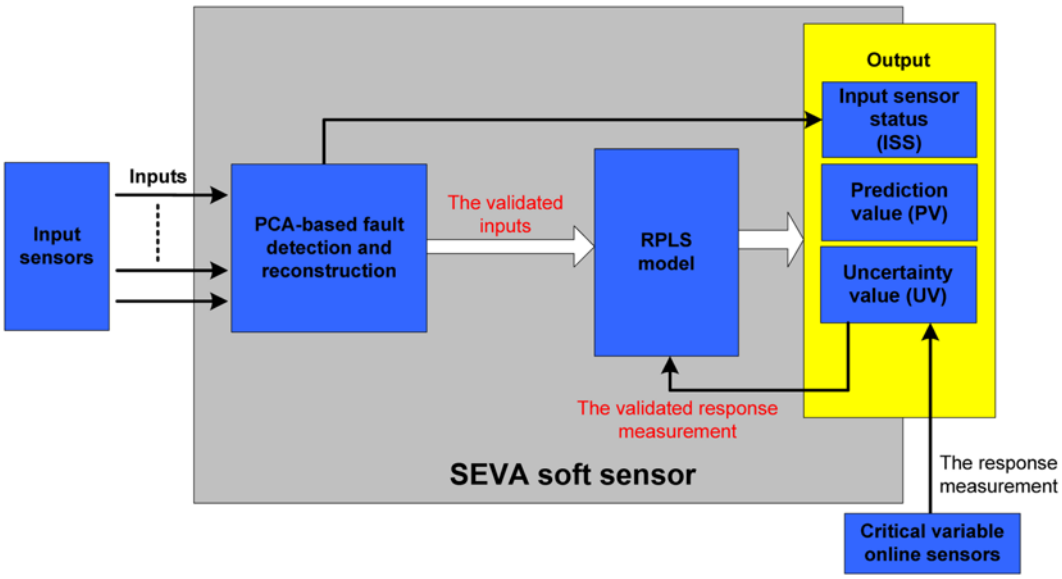

Fig. 1. The SEVA soft sensor framework.

input sensor is faulty, it will be detected, identified and reconstructed by an estimate. A PCA model is obtained to achieve this goal using the sensor validation method proposed by Dunia et al.

(2) Build an RPLS-based prediction model for output variables on the basis of the validated inputs rather than the raw input variables. If one sensor is faulty, the principal components are reconstructed based on the PCA model, thus making the prediction model well conditioned.

(3) Validate the prediction values of RPLS model by using *t*-statistics confidence intervals. These confidence intervals can characterize the uncertainty of the prediction model and provide useful information about prediction quality whether or not a fault has occurred. This is imperative because the response measurement is also used for model updating.

(4) Generate three types of outputs for a soft sensor, including input sensor status (ISS), prediction values (PV) and uncertainty values (UV).

1. PCA-based Fault Detection and Reconstruction

This sub-section presents an automatic and online FDD (fault detection and diagnosis) and sensor reconstruction scheme that can be used in chemical processes and control systems to detect and diagnose sensor faults and reconstruct faulty sensors. The scheme is based on principal component analysis to build a model that captures the correlation among input sensors installed in the processes. The model is employed to reconstruct an assumed faulty sensor. Different from only SPE usage for fault detection, the square prediction error (SPE) based on the model and the sensor validity index (SVI) based on the construction are employed, respectively, to detect the sensor fault and identify the faulty sensor. On the other hand, index unreconstructed variance (URV) is utilized to determine the number of PCs. In short, all the motivation behind the PCA based fault detection and reconstruction is to provide correct signals for subsequent soft sensor modeling. The detailed procedure is summarized in Fig. 2.

1-1. Fault Reconstruction

The prompt identification and correct reconstruction of a faulty

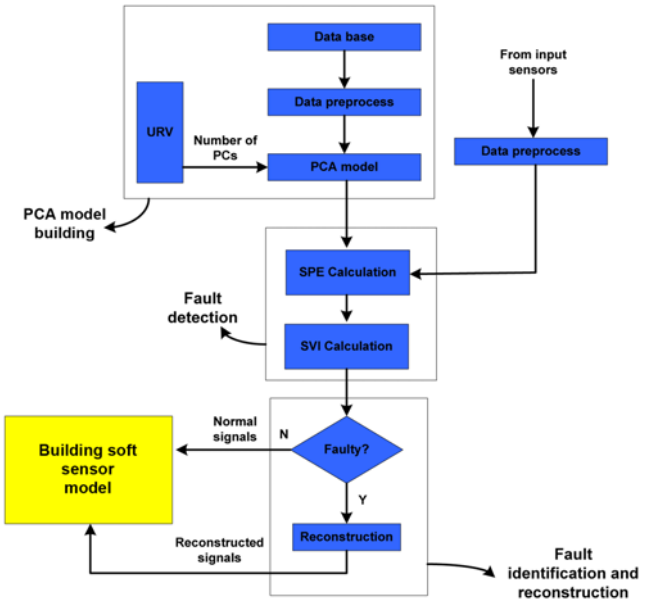

Fig. 2. PCA-based fault detection and reconstruction scheme.

sensor are critical in order to bring the operation of soft sensor back to normal. The approach to sensor identification assumes that one sensor fails and is reconstructed based on the measurements of the remaining sensors. The assumption is then validated by calculating an SVI to examine the residuals before and after reconstruction. If the faulty sensor is validated, a significant decrease in the residuals is expected. Since the tasks for identification and reconstruction are coupled together, we first present the reconstruction approaches and principal determination assuming we know the faulty sensor, and then discuss fault identification and the SVI.

One may estimate the ith variable from x using Eq. (3), where the prediction $\hat{x}_i$ is used as a reconstruction of x_i. Yet, the drawback of this approach is that the faulty sensor contained in x is used in the estimate. To eliminate the effect of the faulty sensor, the prediction

of the ith variable to the input is fed back and iterated until it converges to a value z_i, which can be employed as an estimate of faulty input. The iteration can be represented by the following expression:

$$z_i^{new} = c_{ii} z_i^{old} + [x_{-i}^T \;\; 0 \;\; x_{+i}^T] c_i = [x_{-i}^T \;\; z_i^{old} \;\; x_{+i}^T] c_i, \tag{13}$$

where $C=V_p V_p^T=[c_1\ c_2\ \cdots\ c_m]$, $c_i^T=[c_{1i}\ c_{2i}\ \cdots\ c_{mi}]$ and x^T represents a row of the matrix X and the subscripts −i, +i denote a vector formed by the first i − 1 and the last m − i elements of the original vector, respectively. Dunia et al. [21] show that the iteration always converges and the converged value z_i can be calculated with one formula without actually iterating as follows:

$$z_i = \frac{[x_{-i}^T \;\; 0 \;\; x_{+i}^T] c_i}{1 - c_{ii}}. \tag{14}$$

In practice of PCA, the number of PC's retained in the model is rather subjective; however, the model must have a unique number of PCs for the PCA. If the faulty inputs are reconstructed by a large number of PCs, the predictions with the reconstructed inputs will contain unexpected noises induced by the excess PCs. On the other hand, if fewer PCs are used to reconstruct the faulty data, the incorrect information of the faulty inputs will deteriorate the predictions from the smaller number of PCs. Furthermore, in applying PCA for sensor validation, the number of PC's has a significant impact on each step of the sensor validation procedure, such as the ability to detect small faults, the degree of freedom for fault identification and the accuracy of reconstruction. Dunia et al. [21] also found that when the PCA model is used to reconstruct faulty sensors, the reconstruction error is a non-monotone function of the number of principal components. An URV, which is the variance of the reconstruction error, is proposed to determine the number of principal components. An important feature of this approach is that the proposed index is the minimum corresponding to the best reconstruction of the variable. The unreconstructed variance of variable i, denoted by u_i, is given by the variance of $x_i - z_i$:

$$u_i = \mathrm{var}(x_i - z_i) = \mathrm{var}\left(\varepsilon_i^T x - \frac{[x_{-i}^T \;\; 0 \;\; x_{+i}^T] c_i}{1 - c_{ii}}\right)$$
$$\approx \frac{1}{(1-c_{ii})^2} \varepsilon_i^T (I-C) R (I-C) \varepsilon_i = \frac{\varepsilon_i^T R_e \varepsilon_i}{(\varepsilon_i^T V_p V_p^T \varepsilon_i)^2}. \tag{15}$$

Where $R=E(xx^T)$ is the covariance matrix of the normal data and ε_i is the direction vector of unite length for the faulty sensor. Since each variable x_i is normalized to unit variance, $u_i=1$ if we let $z_i=\bar{x}_i$. Therefore, $u_i>1$ is the case for poor reconstruction.

1-2. Fault Detection and Identification

The approach to identifying a single faulty sensor is discussed below. The algorithm can be serially implemented to identify multiple faults that do not occur simultaneously because it is unusual that sensors fail at the same moment. When a sensor fault has occurred, the sample vector *x* can be represented as follows:

$$x = x^* + f\varepsilon_i \tag{16}$$

Where x^* denotes the portion of normal data, ε_i is the direction vector of unit length for the faulty sensor, and f is the magnitude of the fault which can be negative or positive. To illustrate this method, we denote the reconstructed sample as:

$$x_j^T = [x_{-j}^T \;\; z_j \;\; x_{+j}^T] \tag{17}$$

The reconstructed vector can be projected onto the model and residual subspaces then a SPE can be calculated as

$$SPE(x_j) = \|\tilde{x}\|^2 = x_j^T (I-C) x_j \tag{18}$$

If the i sensor is faulty and j≠i, there will be no much decrease in $SPE(x_j)$. Yet, if a true sensor is chosen, a large reduction in $SPE(x_j)$ is expected. Furthermore, SPE will increase significantly, due to the fault. Therefore, the SVI is used to detect faults, which is more sensitive to the fault and is defined as the sensor validity index (SVI),

$$SVI = \eta_j = \frac{SPE(x_j)}{SPE(x)}. \tag{19}$$

Note that $0 \le \eta_j \le 1$. A validity index close to 1 indicates that the sensor validations follow the variations experienced by the remaining sensors. When the sensor is faulty, η_j is close to zero. Dunia et al. [21] discuss the validity index in detail. Additionally, based on the correlated relationship among variables, other variables that are normal are used to reconstruct the faulty inputs and thus will lead to a reliable prediction.

2. Output Parameters of the SEVA Soft Sensor

Traditionally, the soft sensor has been viewed as a simple prediction signal generator. The application of fault detection techniques, coupled with increasing demands for measurement quality assurance, has rendered inadequate such a simplistic view. In this subsection, a new function of soft sensors is proposed which not only encompasses self-diagnostics capabilities as mentioned above, but also generates a variety of data types, including the prediction value (PV), the uncertainty value (VU), and input sensor status (ISS).

2-1. Prediction Value (PV)

This corresponds to the conventional prediction of soft sensor under normal conditions. Given some easy-to-measure sensors, the soft sensor makes a prediction according to these inputs and models.

2-2. Uncertainty Value (UV)

This is the metrological uncertainty, or probability error of the PV. The metrological definition is used here [22,23]: UV gives a confidence interval for a true prediction. For example, if prediction value is 3.8 units, and UV is 0.19, then there is 95% chance that the prediction value lies within the interval 3.8±0.19 units. Uncertainty is based on existing metrological standards, and its value is calculated based upon all error sources affecting the online measurement, such as: the models, process noise and the effect of any fault. Thus, the VU provides useful information about prediction quality whether or not a fault has occurred.

In the application of RPLS, generally there is the need to provide metrics to measure the quality and reliability of a prediction. In practice, it is common to report population-weighted metrics such as standard error of calibration (SEC) and standard error of prediction (SEP) as a measure of PLS model prediction uncertainty. However, these metrics are typically estimated during model building and validation stages to evaluate general model performance and cannot be directly translated into prediction uncertainty for future samples. Thus, to characterize the UV the variance of RPLS algorithm presented is taken into account. First, considering a convention regression model for a single output, $y=X\beta+\xi$ and all data have been normalized with zero mean and unit variance. ξ is the vector

of residuals identically and independently distributed with mean zero and variance σ^2. In practice, X and y are often contaminated by non-negligible prediction errors. In other words, to simplify the presentation, the presence of prediction error is not indicated by additional notation unless specifically mentioned. The corresponding model to predict response for object o based on observation x_o^T is as follows:

$$\hat{y}_o = \bar{y} + x_o^{*T}\hat{\beta} + \varepsilon_0 \bar{x}^T, \quad (20)$$

where $\bar{y}$ and $\bar{x}^T$ are the measured average response and predictor vector in the calibration data set. $x_o^{*T} = x_o^T - \bar{x}^T$ is mean-centered predictor row vector.

The general procedure to estimate the uncertainty for *y*-response $\hat{y}_o$ of a new individual observation o with predictor row-vector x^T using an empirical model consists of two steps. First, obtain an estimation of standard deviation of the predictor error s. Second, establish a *t*-statistics confidence intervals by the following equation [24]:

$$CI = \hat{y}_o \pm t_{\alpha/2, N-df}\sigma, \quad (21)$$

where N is the number of calibration samples used in convention model building, df is the degree of freedoms used by the model and α is the significance level for the interval. Since the true regression matrix cannot be observed directly, the variance of the regression vector is shown in the following equation when it is estimated by ordinary least-squares (OLS):

$$var(\hat{\beta}_{OLS}) = (x^T x)^{-1} x^T var(y) x (x^T x)^{-1} = (x^T x)^{-1}\sigma^2. \quad (22)$$

Herein we extend the prediction uncertainty by the OLS model to the RPLS model. The variance of the regression vector by the RPLS model can be written as

$$var(\hat{\beta}_{RPLS}) = (x^T x) + x^T var(y) x (x^T x)^{+} = (x^T x)^{+}\sigma^2. \quad (23)$$

The prediction uncertainty of the new data by the RPLS model can be obtained.

$$var(\hat{y}_o) = var(\bar{y} + x_o^{*T}\hat{\beta} + \varepsilon_0) = var(\bar{y}) + x_o^{*T} var(\hat{\beta}_{RPLS}) x_o^{*} + var(\varepsilon_0)$$
$$= \frac{\sigma^2}{N} + x_o^{*T}(x^T x) + x_o^{*}\sigma^2 + \sigma^2 = \sigma^2\left(1 + h_0 + \frac{1}{N}\right) = \sigma^2_{RPLS}. \quad (24)$$

where $h_0 = x_o^{*T}(x^T x) + x_o^{*}$ is the leverage for individual observation o and it measures the distance of an observation to model center in the *x*-space. Note that the term 1/N is due to mean-centering. In this paper, confidence intervals are used to examine the reliability of new observations from online analyzers. The inferential model would not be updated if observations were out of the confidence intervals.

Table 1. Input sensor status values

Status	Explanation
CLEAR	The ISS is derived normally
BLURRED	The ISS is derived from live data, but is being corrected for a fault
DAZZLED	A transient state: the ISS is based on historical data while fault is accessed
BLIND	No credible live data is available
OFFLINE	The instrument is off-line

2-3. Input Sensor Status (ISS)

The input sensor status (ISS) is a discrete-valued flag indicating how the input sensor status of soft sensor has been calculated. The basic categories defined are shown in Table 1. The ISS assists users (human or automated) to determine whether the input sensor measurement is acceptable in the particular application.

SIMULATION AND APPLICATION STUDIES

This is a particularly interesting case study for the proposed methodology. The biological wastewater treatment plant investigated in this work is Activated Sludge, which is a common example of a wastewater treatment process. It was designed for the removal of organic matter and nutrients. In this process the influent rate and composition are variables, the population of microorganisms varies over time (both in quality and number of species), process knowledge is very limited, and BOD_5 online analyzer tends to be unreliable. The amount of organic matter presented is measured as BOD_5. It is very desirable to have a reasonably accurate inferential model for BOD_5 prediction due to a five-day delay that is inherent in the laboratory measurement. In addition, the aeration bank also has a significant hydraulic time delay. And in fact, the process has been observed to go through frequent variations due to seasonal variability and several manual operations. Consequently, sometimes, the experimentation BOD_5 data is not useful for purpose of process control. As shown in Fig. 3, the proposed wastewater plant process [25] is comprised of four elements: pretreatment, primary settlers, aeration tankers and secondary settlers. Also, the parameters needing to be measured are presented.

The data for the proposed soft sensor correspond to data collected daily on the operation of the WWTP (Wastewater treatment plant) in about two-year period. A total of 400 data records have been used, each consisting of 38 process variables. At the same time, the design of the soft sensor requires the application of preprocessing techniques to select the relevant variables. Thus, the automatic clustering algorithms based on Kohomen's self-organizing maps (SOM) [26] are used to detect redundant and irrelevant features. Through this preprocess, nineteen process variables are selected as inputs of models, which are shown as Table 2. And 200 samples are utilized for training; the remaining 200 samples are used for test the performance of the proposed soft sensor. An objective variable y is the concentration of BOD_5, and explanatory variables X are 19 variables, which are biological oxygen demand, suspended solids,

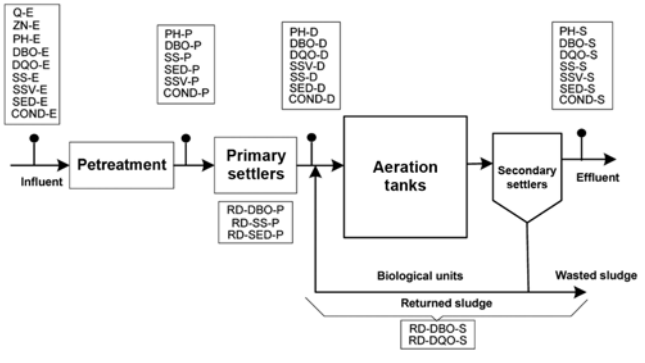

Fig. 3. Wastewater plant process.

Table 2. The parameters selected for inferring

Position	Parameters
Overall plant	Sedimentable solids [RD-SED-G], Suspended solids [RD-SS-G], Biological oxygen demand [RD-DBO-G)], Chemical oxygen demand [RD-DQO-G]
Primary settlers	Biological oxygen demand [RD-DBO-P)], Suspended solids [RD-SS-P]
Secondary settlers	Biological oxygen demand [RD-DBO-S)], Chemical oxygen demand [RD-DQO-S]
Influent to WWTP	Biological oxygen demand and Chemical oxygen demand
Secondary treatment	DQO, Biological oxygen demand, Suspended solids, pH and Sedimentable solids
Output	Chemical oxygen demand, Sedimentable solids, Suspended solids and pH

sedimentable solids, and so on.

1. Validation of the Input Sensor Fault for a SEVA Soft Sensor

1-1. Number of PCs Determination and PCA Model Used for Normal Condition

The prediction performance of a soft sensor is not only dependent on the capability of the inferential model, but also on the data quality of inputs. Thus, inputs of soft sensor validation are also important. The first step of validation is to design a PCA model. Then on the basis of the PCA model designed, the indices are used to detect sensors faults, and faulty sensor identification can be calculated online. In the first test, no fault is added to any sensor in the WWTP. Of the data collected from the database, about one-half of the data points are used for training the PCA model and the other half of the points are used for testing the indices for sensor fault detection and identification under normal operation conditions. The first step to sensor validation is to obtain the number of principal components for best reconstruction. The total URVs are calculated by reconstructing each sensor from others. As can be noted in Fig. 4, when the number of principal components is 9, the total URV is minimum. Therefore, nine principal components are used as the optimal number for sensor reconstruction.

Based on the PCA model obtained, the indices for sensor faults and faulty sensor identification are calculated by using the other half of the data points. Fig. 5 shows the two indices for the test data set in normal condition. Due to no fault being added, the SPE and SVI are within their control limit, and SVI is close to 1. Since we treat points that continue to get over the control limit less than three times as false alarms, some points that exceed the control limit in SPE are ignored. These results demonstrate that the test data are

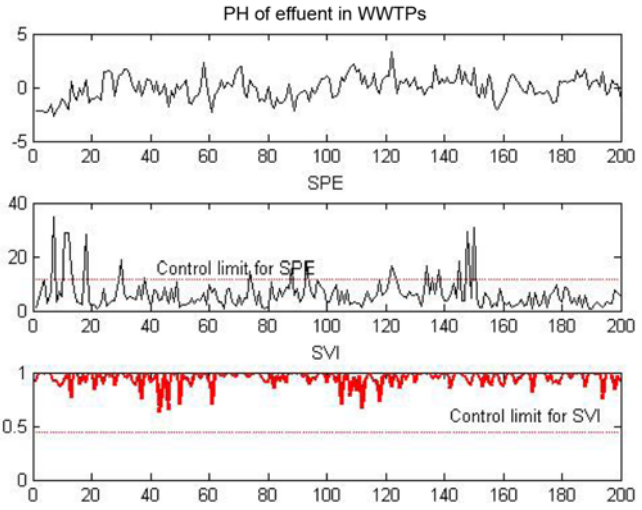

Fig. 5. Indices for fault detection at normal condition, their confidence intervals are 95%.

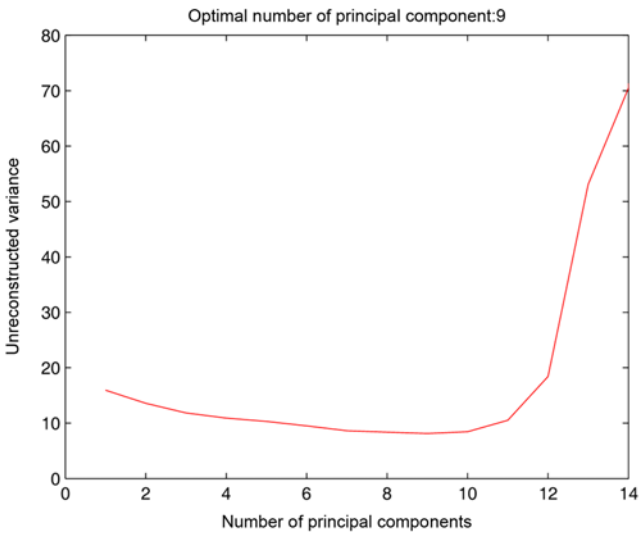

Fig. 4. Total URV versus the number of principal components.

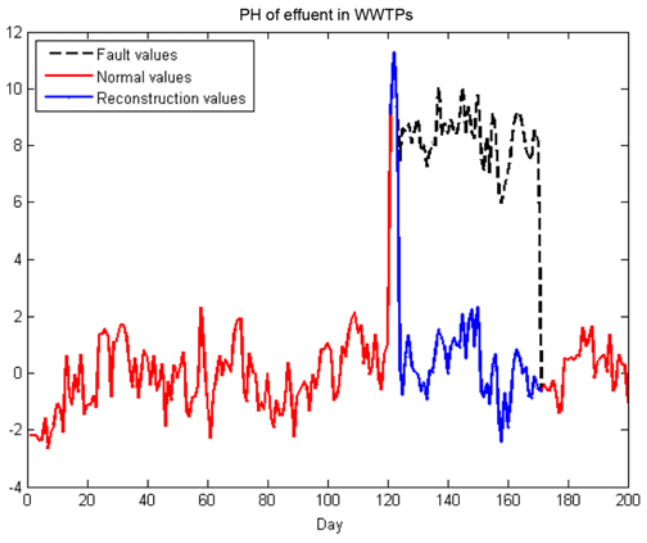

Fig. 6. Fault, normal and Reconstructed pH.

within the region of the training data and involve no fault.

1-2. PCA Model Employing in Abnormal Condition

The pH of effluent in the WWTP is one of the most important variables for soft sensor input. To verify the senor validation method for a soft sensor, it is assumed that the measured values of the pH in the secondary settler have a bias value (8) from the 120th day to 170th in the test data, as shown in Fig. 6. The SPEs were calculated online and compared with the threshold to monitor the sensor status in the wastewater treatment system. The SVI of each sensor was also calculated by reconstructing it using the measurements of the remaining sensors, in order to identify the faulty sensor. Fig. 7 shows the sensor fault detection index and the identification index for the pH sensor. Before the fault was introduced, the SPE was below the threshold, and the SVIs of all sensors were close to 1. After the presence of the bias fault, the SPE was significantly greater than the threshold. At this time, the SVI for the pH sensor began to be close to 0. However, the SVIs of the other sensors were still close to 1 (not shown in Fig. 7). The SPE indicates that a fault occurred among the sensors of the wastewater plant from 120th to 170th day, and the SVI shows that the fault occurred on the pH sensor. Fig. 6 shows the measured and recovered measurements of the pH sensor under the bias fault. The measurements reconstructed from the remaining sensors before and after the fault presence approach the actual value with very high accuracy. It is also necessary to note that faulty alarms occurred when using SPE from Fig. 7. Yet, when performing SVI, no faulty alarms happened.

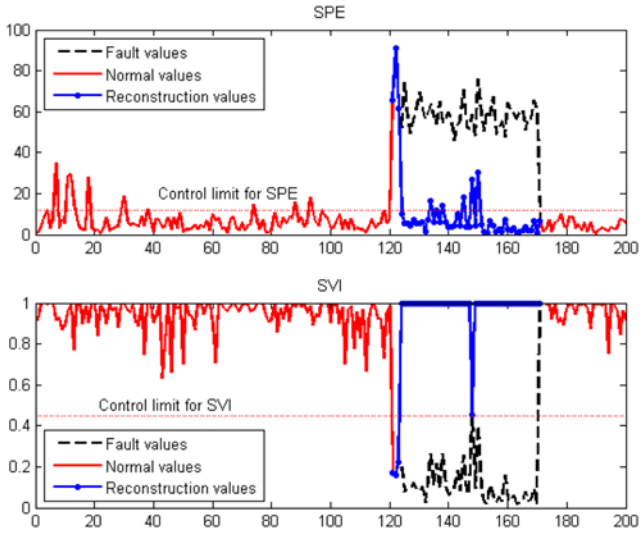

Fig. 7. SPE and SVI for a sensor fault that occurs and disappears after a period of time.

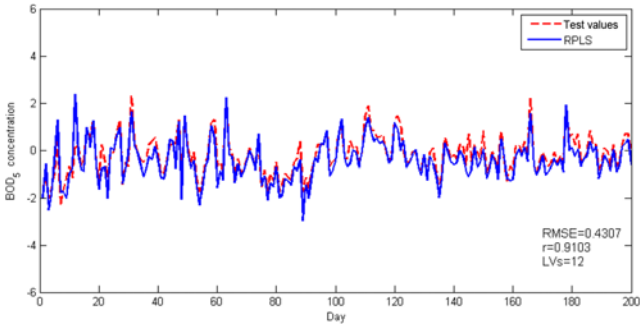

Fig. 8. Prediction result of BOD_5 using RPLS, RMSE=0.4307, r=0.9103, LVs=12.

2. Prediction Model of the SEVA Soft Sensor and its Output Parameters

The root mean square error (RMSE) and coefficient (r) are used to access the prediction performance of inferential model.

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_{i,actual}-Y_{i,mod}}{Y_{i,actual}}\right)^2}, \tag{25}$$

where $Y_{i,actual}$ and $Y_{i,mod}$ are the measurement and the output of prediction model, respectively.

2-1. Building an RPLS Model for the SEVA Soft Sensor

A soft-sensor model that estimates BOD_5 of the Barcelona WWTP is constructed by using recursive PLS, and it is updated every 24 h when BOD_5 is measured. To take into account process dynamics, the input data consist of the present samples. And no fault occurs herein. The number of latent variables used in the RPLS model is determined by trial and error to maximize the prediction performance. The estimation result is shown in Fig. 9. The estimates shown in the bottom figure are calculated every day whenever the input variables are observed, and they fluctuate by measurement noise. In this figure, r denotes the correlation coefficient between measurements and estimates, RMSE is the root-mean-squares error, and LVs is the number of latent variables. Next, the forgetting factor β is used to adapt the model to the changes in process characteristics more rapidly. The estimation result with forgetting factor β=0.95

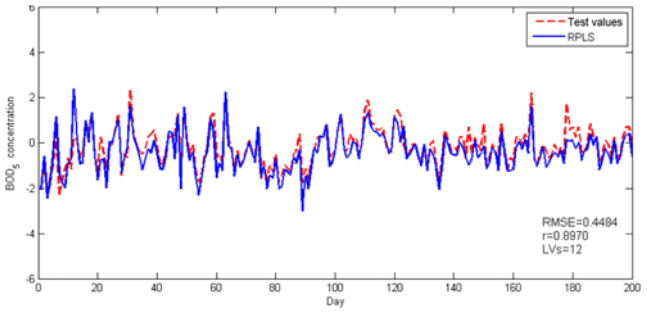

Fig. 9. Prediction result of BOD_5 using RPLS with β=0.95, RMSE =0.4484, r=0.8970, LVs=12.

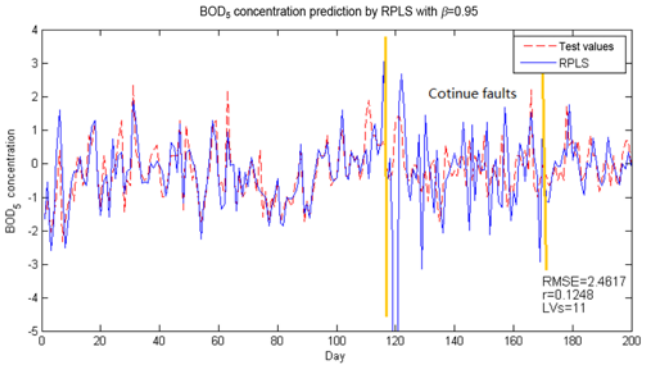

Fig. 10. Soft sensor prediction using RPLS as model when it is subjected to continuing faults.

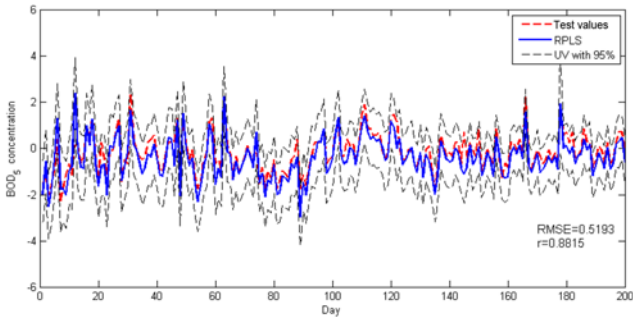

Fig. 11. BOD_5 prediction with 95% confidence intervals when suffering from continuing abrupt changes in pH over 50 days.

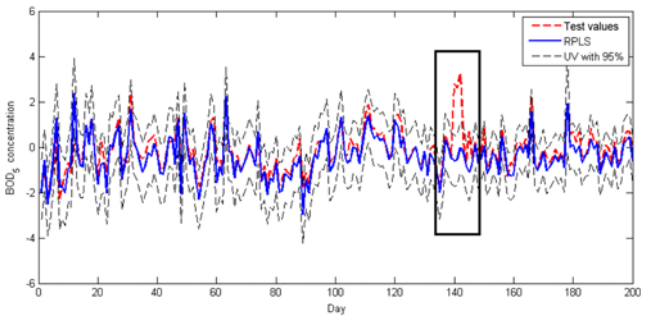

Fig. 13. BOD_5 prediction with 95% confidence intervals when suffering from some undetected abrupt changes by PCA.

is shown in Fig. 10. The estimation accuracy deteriorates a little bit. It is obvious that RPLS can capture the nonlinear relationship between the inputs and output when a suitable forgetting factor and LVs are selected.

2-2. Verify an SEVA Soft Sensor and its Output Parameters

When pH suffers from continuing faults with biases (8) between the 120th and 170th sample, the soft sensor using RPLS model and raw (not be validated) data makes unstable predictions for BOD_5 in Fig. 10. The fit between the model prediction values and test values is poor, where RMSE is 2.4617. This is because the faulty data are not identified and still utilized for subsequent RPLS modeling.

However, as can also be seen in Fig. 11, the SEVA soft sensor can still track the variation of BOD_5 because of PCA fault detection and reconstruction method using SPE and SVI. Not only can this soft sensor identify and reconstruct the faults as described in Fig. 6 and Fig. 7, but also indicate the status of the soft sensor changing from BLURRED to DAZZLED in Fig. 12. Because the faults which are detected by PCA model are not permanent and the pH sensor returns to normal, the self-validating soft sensor declares the measurement CLEAR again after the 170th sample. All the status changes are shown in Fig. 12. Note that the two dotted lines represent uncertainty values of model, which will described in the following part.

Since the proposed method adapts the model once a new measurement of the online output analyzer is available, the model may be misled by samples with noise or abnormal data. In an industrial process, a common example of encountering abrupt noise occurs when the online analyzer is calibrating, so there is a need to perform checking in the output of the soft sensor. In this work, the confidence intervals of the response measurements have been created using prediction variance to ensure that the RPLS model will not be misled. As Fig. 13 shows, some test values from the online analyzer deviate from the 95% control limit. These data are labeled as abnormal and should not be used to update the RPLS model. Thus, the RPLS model cannot be misled by samples with noises or abnormal data, and the confidence intervals of predictions are capable of preventing the model from some undetected noises.

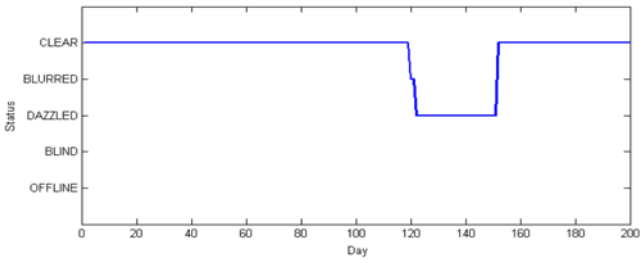

Fig. 12. The input sensors status of soft sensor.

CONCLUSIONS

The SEVA soft sensor approach proposed does not just perform self-diagnostics and self-reconstruction, but also generates a variety of data types, including the prediction value, input sensors status of soft sensor and the uncertainty values which represent the credibility of soft sensor's output. The integrated framework is successfully applied to a wastewater treatment process to predict BOD_5. The SPE and SVI which are generated from PCA are derived and shown to be effective in detecting, identifying and reconstructing the single sensor fault. Also, the RPLS has been proven to accommodate the time-varying nature of the process. Even if suffering from some abnormal events, the RMSE of the SEVA soft sensor is improved by 474% in comparison with the general RPLS-based soft sensor because of using input sensors validation. Additionally, this resulting soft sensor framework may validate the output of soft sensor and give statuses of its input sensors, which were illustrated through the wastewater treatment example. The proposed approach has the potential for the implementation of soft sensors in the process industry. Further study of uncertainty description on model and noises will be carried out in the near future.

ACKNOWLEDGEMENTS

This work is supported by the Fundamental Research Funds for the Centrol Universities, SCUT (2012zm0102, 2009zm0161) and the National Natural Science Foundation of China (60704012).

REFERENCES

1. P. Facco, F. Doplicher, F. Bezzo and M. Barolo, *J. Process Control*, **19**, 520 (2009).
2. H. Martens, M. Hoy, F. Westad, D. Folkenberg and M. Martens, *Chemometrics and Intelligent Laboratory Systems*, **58**, 151 (2001).
3. D. Aguado, T. Montoya, L. Borras, A. Seco and J. Ferrer, *Engineering Applications of Artificial Intelligence*, **21**, 919 (2008).
4. T. Khayamian, *Chemometrics and Intelligent Laboratory Systems*,

88, 35 (2007).
5. L. Fortuna, S. Graziani and M. G. Xibilia, *Control Engineering Practice*, **13**, 499 (2004).
6. E. Alhoniemi, *Integrated Computer Aided Engineering*, **6**, 3 (1999).
7. W. Yan, H. Shao and X. Wang, *Comput. Chem. Eng.*, **28**, 1489 (2004).
8. W. Li, H. H. Yue, S. V. Cervantes and S. J. Qin, *J. Process Control*, **10**, 471 (2000).
9. N. Pessel, J. F. Balmat, F. Latont and J. Bonnal, 9th WSEAS International Conference on Automatic Control, Modeling and Simulation, Istanbul, Turkish (2007).
10. D. X. Doan, X. Tien, K.-W. Lim and L. Jun, The 30th Annual Conference of the IEEE Industrial Electronics Society, Busan, South Korea (2004).
11. F. Zhang, Proceedings of the 2004 American Control Conference, Boston, U.S.A. (2004).
12. G. Martin, *Chem. Eg. Prog.*, **93**, 66 (1997).
13. S. J. Qin, H. Yue and R. Dunia, *Ind. Eng. Chem. Res.*, **36**, 1675 (1997).
14. H. H. Yue and S. J. Qin, *Ind. Eng. Chem. Res.*, **40**, 4403 (2001).
15. F. Ahmed, S. Nazir and Y. K. Yeo, *Korean J. Chem. Eng.*, **26**, 14 (2009).
16. J. Ciba, P. Dydo and K. J. Tags, *Chemosphere*, **76**, 565 (2009).
17. P. Facco, F. Doplicher, F. Bezzo and M. Barolo, *J. Process Control*, **19**, 520 (2009).
18. S. Mu, Y. Zeng, R. Liu, P. Wu, H. Su and J. Chu, *J. Process Control*, **16**, 557 (2006).
19. J. E. Jackson and G. Mudholkar, *Technometrics*, **21**, 341 (1979).
20. J. S. Qin, *Comput. Chem. Eng.*, **22**, 503 (1998).
21. R. Dunia, S. J. Qin, T. F. Edgar and T. J. McAvoy, *AIChE J.*, **42**, 2797 (1996).
22. S. Kline and F. McClintock, *Mech. Eng.*, **3**, 8 (1853).
23. T. Q. Management and S. S. P. Committee, British Standards Institution, London, 3 (1995).
24. J. Liu, D. S. Chen and J.-F. Shen, *Ind. Eng. Chem. Res.*, **49**, 11530 (2010).
25. C. L. Blake and C. J. Merz, UCI Repository of Machine Learning Databases (1998).
26. R. Rallo, J. Ferré-Giné and F. Giralt, Report, Spain (1998).