Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Process Control 15 (2005) 39–52
JOURNAL OF PROCESS CONTROL
www.elsevier.com/locate/jprocont

Optimal selection of soft sensor inputs for batch distillation columns using principal component analysis

Eliana Zamprogna [a,1], Massimiliano Barolo [a,*], Dale E. Seborg [b]

[a] DIPIC—Dipartimento di Principi e Impianti di Ingegneria Chimica, Università di Padova, Via Marzolo, 9, I-35131 Padova, PD, Italy
[b] Department of Chemical Engineering, University of California, Santa Barbara, CA 93106, USA

Received 20 October 2003; received in revised form 15 March 2004; accepted 14 April 2004

Abstract

In this paper, a novel methodology based on principal component analysis (PCA) is proposed to select the most suitable secondary process variables to be used as soft sensor inputs. In the proposed approach, a matrix is defined that measures the instantaneous sensitivity of each secondary variable to the primary variables to be estimated. The most sensitive secondary variables are then extracted from this matrix by exploiting the properties of PCA, and they are used as input variables for the development of a regression model suitable for on-line implementation.

This method has been evaluated by developing a soft sensor that uses temperature measurements and a process regression model to estimate on-line the product compositions for a simulated batch distillation process. The identification of the optimal soft sensor inputs for this case study has been discussed with respect to the definition of the sensitivity matrix, the data sampling interval, the presence of measurement noise, and the size of the input set. The simulation results demonstrate that the proposed approach can effectively identify the size and configuration of the input set that leads to the optimal estimation performance of the soft sensor.

Keywords: Optimal sensor location; Principal component analysis; Measurement selection; Soft sensor; Batch distillation; Partial least squares regression

1. Introduction

Inferential estimators (or soft sensors) represent an attractive approach for estimating primary process variables, particularly when conventional hardware sensors are not available, or when their high cost or technical limitations hamper their on-line use. Inferential estimators make use of easily available process knowledge, including a process model and measurements of secondary process variables, to estimate primary variables of interest [5]. Typically, in the process industries inferential estimators are used to estimate product compositions from temperature and other secondary variables.

It is well known that an inferential estimator can be developed in the form of a Luenberger observer [16] or a Kalman filter [12] using a first-principles dynamic model of the process. However, because chemical processes are generally quite complex to model and are characterized by significant inherent nonlinearities, a rigorous theoretical modeling approach is often impractical, requiring a great amount of effort. For these reasons, recently there has been an increasing interest toward the development of inferential estimators based on heuristic models of the process. For example, the inferential estimator can be based on available measurements and multivariate regression techniques. This alternative modeling approach is advantageous because a soft sensor can provide a fast and accurate response, thus overcoming the typical limitations of hardware sensors [15]. Moreover, because soft sensors are easy to develop and to implement on-line, they are potentially more attractive than stochastic filters or deterministic observers. Artificial neural networks (ANN) and partial least squares (PLS) regression are widely used regression techniques, and

* Corresponding author. Tel.: +39-049-827-5473; fax: +39-049-827-5461.
E-mail address: max.barolo@unipd.it (M. Barolo).
[1] Current address: Corporate Technology Department, CT2; Buhlergroup AG; CH-9240 Uzwil, Switzerland.

0959-1524/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jprocont.2004.04.006

their successful application to the development of soft sensors for product composition estimation has been reported for different processes [14,22].

However, it is well known that the satisfactory performance of inferential estimators is likely to be achieved if only those secondary variables that are most sensitive to the primary variables are employed. In fact, the inappropriate selection of estimator inputs may lead to numerical problems, such as singularity and over-parameterization, or may markedly reduce the estimation accuracy [13]. Also, it is not generally possible to overcome the issue of measurement selection by using *all* available secondary variables as soft sensor inputs, because measurement redundancy generally makes the calibration of the regression model troublesome, and can undermine the accuracy of the resulting estimator.

In this paper, a systematic measurement selection methodology is proposed and demonstrated in a simulated case study for a batch distillation process. The choice of this benchmark is justified by the fact that both continuous and batch distillation have shown to benefit from the use of regression soft sensors [6,13,18,26].

To develop a composition soft sensor for a distillation process, temperature measurements are typically used as secondary variables. However, it can be difficult to select the optimal set of secondary variables to be used as estimator inputs, because there are many possible locations for temperature sensors. When continuous distillation is regarded, guidelines for optimal sensor location have been proposed on the basis of rule-of-thumb approaches [24]. Joseph and Brosilow [11] suggested an iterative selection method based on the addition of temperature measurements to the optimal set, one at a time. The procedure is repeated until satisfactory estimation accuracy is obtained, or until all measurements have been included. When the number of available secondary variables is large, however, this iterative procedure may be impractical and time consuming.

Two systematic methods have also been developed to select the best measurement location for process control purposes. Tolliver and McCune [25] proposed that the optimum temperature location be determined by evaluating the column sensitivity to the material balance and represent it as a sensitivity gain matrix. The second approach, which has been investigated by several authors [4,8,19], is based on the application of singular value decomposition (SVD) to the sensitivity gain matrix, and the determination of the sensor locations that are characterized by having the highest sensitivity and lowest mutual interaction. These locations are considered to be the most suitable choices for multivariable control purposes. Bequette and Edgar [3] have indicated that these two methods generally lead to the selection of the same tray temperature measurements, which are usually located approximately one-fourth of the distance from each end of the column. They also pointed out that neither method considers the effect of disturbance variables, which may be detrimental for the control performance.

Optimal temperature measurement selection for a batch distillation process entails additional significant difficulties, because the location of the most sensitive trays may change during the operation due to the inherent dynamic nature of the process. In fact, a continuous shift of the column temperature profile occurs during the batch (from the bottom of the column to the top), which makes it difficult to determine a priori which tray temperature measurements can be used to reliably infer product compositions during the entire operation. Thus, the optimal location of "sensitive" trays may change during a batch. For example, Oisiovici and Cruz [21] showed that the optimal sensor configuration obtained by applying the SVD approach to a batch column is time-varying because of the dynamic behavior of the process. Furthermore, it is important to recall that the SVD approach has been developed to select optimal measurement locations for process control purposes. For batch distillation columns, the critical issue is monitoring rather than control. Thus, the SVD approach cannot in principle be used in this case, and could even lead to misleading results. Indeed, when the optimal number and location of temperature measurement points have to be selected for batch columns, no systematic guidelines are presently available.

Quintero-Marmol et al. [23] suggest that $N_C + 2$ temperature measurements be considered, where N_C is the number of chemical components in the feed. They also recommend locating one sensor in the still pot while distributing the remaining sensors evenly throughout the column. While this appears to be a sound guideline, it may nevertheless lead to a sensor configuration where some of the most informative locations are omitted. Oisiovici and Cruz [20] considered a high-purity batch distillation column, and investigated the influence of the temperature sensor locations on the estimation accuracy of an extended Kalman filter. They found that the estimation performance depend markedly on the sensor locations and claimed that it is advisable to place the temperature sensors away from the top stages. Barolo et al. [2] found that measurement noise can have a great impact on the appropriateness of measurement locations for a middle-vessel batch distillation column separating a highly nonideal ternary mixture.

From the above discussion, it appears that a systematic approach for the selection of the optimal number and location of temperature measurements for composition estimation in batch distillation is still lacking. In this paper, a novel input selection method-

ology is proposed based on principal component analysis (PCA) [10].

2. Selection of the optimal sensor location for monitoring purposes

In order to select the most suitable secondary variables to be used for process monitoring via soft sensor, a sensitivity index is proposed that measures the degree of sensitivity of each available secondary variable (tray temperature) with respect to changes in each primary variable (product composition). This sensitivity index is defined as the partial derivative of each secondary variable with respect to each variable to be estimated. The sensitivity indexes calculated for all the available process variables are collected in a gain matrix $\mathbf{K}$:

$$\mathbf{K} = \begin{bmatrix} \frac{\partial T_1}{\partial x_1} & \cdots & \frac{\partial T_1}{\partial x_i} & \cdots & \frac{\partial T_1}{\partial x_m} \\ \vdots & & \vdots & & \vdots \\ \frac{\partial T_j}{\partial x_1} & \cdots & \frac{\partial T_j}{\partial x_i} & \cdots & \frac{\partial T_j}{\partial x_m} \\ \vdots & & \vdots & & \vdots \\ \frac{\partial T_n}{\partial x_1} & \cdots & \frac{\partial T_n}{\partial x_i} & \cdots & \frac{\partial T_n}{\partial x_m} \end{bmatrix}^{\mathrm{T}}, \tag{1}$$

where T_j is the jth secondary variable, x_i represents the ith primary variable, n is the number of available secondary variables, and m is the number of primary variables to be estimated. Because the units of the sensitivity gains should be chosen to reflect the operability range of sensors, both T_j and x_i are expressed as a percentage of the maximum sensor signal, as:

$$T_j = \frac{T_j(t) - T_0}{\Delta T} \times 100\ [\%], \tag{2}$$

$$x_i = \frac{x_i(t) - x_0}{\Delta x} \times 100\ [\%], \tag{3}$$

where $T_j(t)$ and $x_i(t)$ indicate the signals obtained at each sampling instant from the sensors measuring the jth secondary variable and the ith variable to be estimated, respectively; T_0 and x_0 represent the corresponding instrument zeroes; and ΔT and Δx denote the corresponding instrument spans.

The $m \times n$ sensitivity matrix $\mathbf{K}$ can be determined from simulations based on a first-principles process model. In principle, a sensitivity gain matrix can be calculated for both continuous and batch processes. For continuous processes, $\mathbf{K}$ is time-invariant, and can be obtained by applying "small" perturbations of the primary variables around the reference steady state of the system. Conversely, for batch processes $\mathbf{K}$ is time-varying. In this case, an instantaneous pseudo-steady state sensitivity matrix is calculated at different time instants t during the batch by the following approximation:

$$\widehat{\mathbf{K}}(t) = \begin{bmatrix} \frac{\Delta T_1}{\Delta x_1} & \cdots & \frac{\Delta T_1}{\Delta x_i} & \cdots & \frac{\Delta T_1}{\Delta x_m} \\ \vdots & & \vdots & & \vdots \\ \frac{\Delta T_j}{\Delta x_1} & \cdots & \frac{\Delta T_j}{\Delta x_i} & \cdots & \frac{\Delta T_n}{\Delta x_m} \\ \vdots & & \vdots & & \vdots \\ \frac{\Delta T_n}{\Delta x_1} & \cdots & \frac{\Delta T_j}{\Delta x_i} & \cdots & \frac{\Delta T_n}{\Delta x_m} \end{bmatrix}^{\mathrm{T}}, \tag{4}$$

where $\Delta x_i = x_i(t + \Delta t) - x_i(t)$ indicates the variation of the ith primary variable during the selected time interval Δt, and $\Delta T_j = T_j(t + \Delta t) - T_j(t)$ represents the variation of the jth secondary variable in the same period. It should be noted that, because batch processes are inherently dynamic, all variables are time varying during the time interval Δt. Consequently, each element $\Delta T_j/\Delta x_i$ of $\widehat{\mathbf{K}}$ is only an approximation of the corresponding partial derivative $\partial T_j/\partial x_i$.

The properties of principal component analysis are exploited in order to identify the most appropriate set of secondary variables for monitoring purposes from the information contained in the sensitivity matrix $\widehat{\mathbf{K}}$. The $\widehat{\mathbf{K}}$ matrix is first scaled in such a way that each row is normalized to zero mean and unit variance [10]

$$\tilde{k}_{ij} = \frac{\hat{k}_{ij} - \bar{k}_i}{\sigma_i} \tag{5}$$

with

$$\bar{k}_i = \frac{1}{n} \sum_{j=1}^{n} \hat{k}_{ij}, \tag{6}$$

$$\sigma_i^2 = \frac{\sum_{j=1}^{n} (\hat{k}_{ij} - \bar{k}_i)^2}{n - 1}, \tag{7}$$

where $\hat{k}_{ij}$ indicates an element of $\widehat{\mathbf{K}}$, $\tilde{k}_{ij}$ is its normalized value, $\bar{k}_i$ and σ_i are the mean and standard deviation of the ith row of $\widehat{\mathbf{K}}$, respectively. This normalization procedure was the most suitable to pre-process the information contained in $\widehat{\mathbf{K}}$ over alternative scaling methods [27].

In a PCA analysis, the normalized gain matrix $\widetilde{\mathbf{K}}$ is factored into two matrices [10]:

$$\widetilde{\mathbf{K}}(t) = \mathbf{T}\mathbf{P}^{\mathrm{T}}, \tag{8}$$

where $\mathbf{T}(m \times s)$ is the score matrix and $\mathbf{P}(n \times s)$ is the orthonormal loading matrix, whose rows are the s principal components.

In the proposed approach, the PCA decomposition is performed in such a way that the original information contained in $\widetilde{\mathbf{K}}$ is summarized into a single principal component ($s = 1$). Thus, the loading matrix $\mathbf{P}$ becomes a vector, which represents the direction that is most sensitive to the primary variables, and the jth element of $\mathbf{P}$ can be interpreted as a measure of the contribution of the jth secondary variable to that high-sensitivity direction. Therefore, the largest value of the principal components identifies the secondary variable that is

most sensitive to the primary variables, thus resulting the most profitable to be used as soft sensor input. The second largest value of the loadings identifies the second most sensitive measurement location, and so on.

The PCA transformation of the sensitivity matrix also indicates the number of measurements that need to be taken into account, because all secondary variables that correspond to loadings with much smaller value than the largest one can be disregarded.

For batch processes, the sensitivity gain matrix $\widehat{\mathbf{K}}$ calculated at each time sample t and the PCA-based sensitivity analysis identify the most sensitive secondary variables at the current sampling instant. The overall optimal configuration for the soft sensor inputs is then determined by calculating the cumulative PC index, CUMPC, for each secondary variable:

$$\mathrm{CUMPC}_j = \sum_{t=1}^{N_s} p_j(t), \tag{9}$$

where $p_j(t)$ represents the value of the principal component obtained at time t for the jth secondary variable, and N_s indicates the total number of samples. The set of secondary variables that have the highest CUMPC values are considered as the optimal soft sensor inputs.

3. Process description and data generation

The separation of a hypothetical zeotropic ternary mixture in a conventional batch rectifier with 20 trays is used to verify the effectiveness of the proposed measurement selection method.

The batch column, which is shown in Fig. 1, is operated according to the constant-reflux strategy described by Luyben [17]. In this strategy, the column is initially operated at total reflux. When the distillate composition meets the desired quality specification, the distillate withdrawal is started, and products (P_1 and P_2) and slop cuts (S_1 and S_2) are sequentially collected from the top and segregated in separate tanks. The heaviest product (P_3) is extracted from the reboiler at the end of the batch. The process objective is to recover each component of the feed at a given minimum purity level. In particular, the mole fraction of the key component in each product must be greater than or equal to 0.95. The physical model of the process consists of a system of differential and algebraic equations that have been obtained by considering conventional simplifying assumptions (i.e., theoretical stages, negligible vapor hold-up, constant-reflux drum holdup, constant vapor boilup rate and internal vapor flow, constant pressure, constant relative volatilities, perfectly-mixed capacities, and total condensation with no sub-cooling). The model parameters reported by Barolo and Berto [1] and a tray holdup of 5 mol are used in this study.

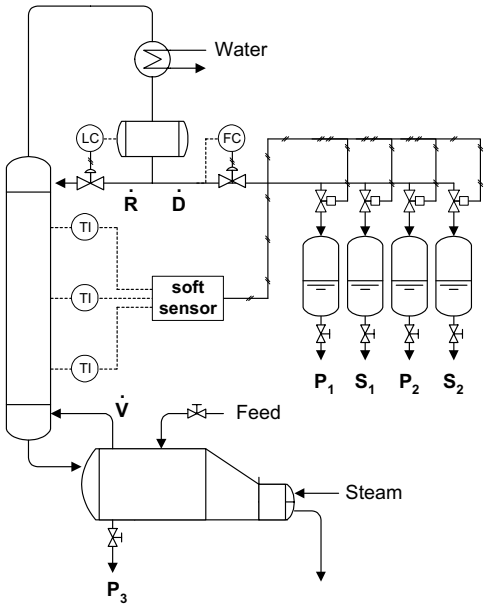

Fig. 1. Schematic diagram of the batch distillation column and its control configuration.

A soft sensor is developed for this process in order to estimate the instantaneous product compositions using temperature measurements, as schematically shown in Fig. 1. The soft sensor estimates the mole fraction of the light and intermediate components in the distillate stream ($x_{D,1}$ and $x_{D,2}$, respectively), and the mole fraction of the heavy component in the reboiler ($x_{B,3}$). They are the key compositions needed for process monitoring. Partial least squares regression and artificial neural networks are used to obtain the empirical models for the soft sensor because these methods can provide an accurate representation of the process behavior and require low computational load [27]. A detailed description of the PLS algorithm and its mathematical formulation are provided by Geladi and Kowalski [7]. Theoretical background on ANN can be found in the book by Haykin [9].

The data needed to calibrate and validate the composition estimator are generated using the nonlinear physical model of the batch column and the operating conditions reported in Table 1. The time-varying tra-

Table 1
Operating conditions for the batch distillation column

Mixture relative volatility, $\alpha_1/\alpha_2/\alpha_3$	9/3/1
Feed composition, $x_{F,1}/x_{F,2}/x_{F,3}$	0.45/0.50/0.05
Feed charge, F	300 mol
Vapor boilup rate, V	110 mol/h
Distillate withdrawal rate, D	50 mol/h
Reflux drum holdup, H_D	10 mol
Tray hold up, H_i	5 mol
Tray hydraulic time constant	0.001 h
Number of ideal trays, N	20
Nominal composition setpoint, $x_{P1}^{sp}/x_{P2}^{sp}/x_{P3}^{sp}$	0.95/0.95/0.95

jectories of all process variables are monitored throughout the entire duration of the batch, and recorded using a sampling period of 18 s. At each time instant, the sensitivity matrix $\widehat{\mathbf{K}}$ (3×21) is computed from the temperatures for all 20 column trays and the reboiler, and from the "measurements" of $x_{D,1}$ and $x_{D,2}$, and $x_{B,3}$. The proposed PCA sensitivity analysis is employed to identify the most informative temperature measurements to be used as inputs for the composition estimator.

The estimation performance of the soft sensors that are obtained for different input sets are evaluated and compared. The estimation accuracy is assessed in terms mean squared (MSQ) error, which is calculated as:

$$\mathrm{MSQ}_i = \sqrt{\frac{(\mathbf{x}_i - \hat{\mathbf{x}}_i)(\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}}}{N_s}}, \tag{10}$$

where $\mathbf{x}_i$ is the row vector of measurements for the ith variable x_i, and $\hat{\mathbf{x}}_i$ is the corresponding estimate from the soft sensor. The most effective measurement selection strategy is the one leading to the composition estimator with the lowest value of MSQ.

4. Optimal temperature sensor location using conventional methods

Based on practical considerations, Quintero-Marmol et al. [23] suggested that $N_C + 2$ temperature measurements should be considered, where N_C is the number of components in the feed mixture. They also recommended that one sensor should be placed in the still pot, while the remaining ones should be distributed evenly along the column.

Alternatively, information on the most sensitive temperature measurements can be extracted from the sensitivity gain matrix by means of direct analysis of the matrix or through the extension of the SVD analysis proposed by Moore [19] and by Oisiovici and Cruz [20]. The results obtained from these two approaches are reported in the next two subsections.

4.1. Optimal sensor configuration from direct analysis of the sensitivity matrix

The sensitivity matrix $\widehat{\mathbf{K}}$ can be used directly to establish the instantaneous optimal sensor configuration. Because $\hat{k}_{ij}$ is a measure of the sensitivity of the jth secondary variable to the variation in the ith primary variable, the secondary variable having the largest value of $\hat{k}_{ij}$ could be considered as the most suitable soft sensor input. Similarly, the location having the second largest value of $\hat{k}_{ij}$ is the second most appropriate soft sensor input, and so on.

Fig. 2 shows the locations of the three most sensitive temperatures identified from the direct analysis of $\widehat{\mathbf{K}}$ (a bottom-to-top tray numbering scheme is used, with "B" corresponding to the reboiler and "20" denoting the top tray). The location of the three most sensitive trays along the columns varies considerably during a batch. It is interesting to note that, at a given time instant, the three most sensitive measurement points are located in the same section of the column. This "sensitive region" is initially located at the top of the column, but drops suddenly to the bottoms section at $t \approx 0.5$ h. Subsequently, it shifts towards the top of the column, before dropping again to the bottoms toward the end of the batch ($t \approx 3.5$ h). This trend suggests that no particular region of the column shows consistently high sensitivity for the entire duration of the batch, and all column trays can be considered equally important from the sensitivity point of view. Therefore, these results seem to support the even distribution proposed by Quintero-Marmol et al. [23].

4.2. Optimal sensor configuration from SVD analysis

The sensitivity information of $\widehat{\mathbf{K}}$ can be extracted by exploiting the properties of Singular Value Decomposition, as proposed by Moore [19]. As mentioned earlier, this SVD approach was originally developed for control purposes in order to select the column temperatures that have lowest mutual interaction and highest sensitivity to the manipulated variables. In principle, this method can

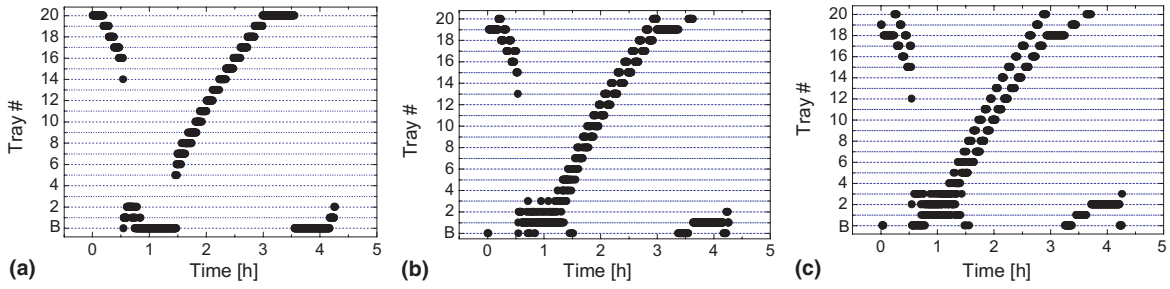

Fig. 2. Tray sensitivity information obtained from direct analysis of $\widehat{\mathbf{K}}$: variation of the (a) most sensitive location, (b) second most sensitive location, and (c) third most sensitive location during the batch of Table 1.

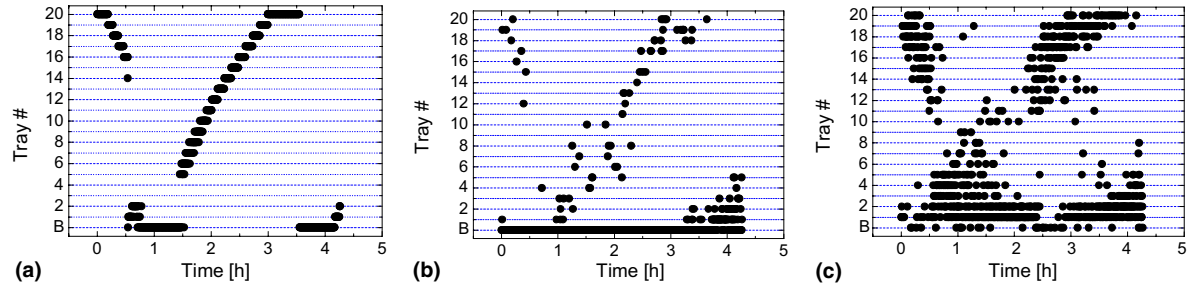

Fig. 3. Tray sensitivity information obtained from SVD analysis of $\widehat{\mathbf{K}}$: variation of the (a) most sensitive location, (b) second most sensitive location, and (c) third most sensitive location during the batch.

be extended to process monitoring. Because of the properties of the SVD analysis, the application of this approach to the sensitivity gain matrix $\widehat{\mathbf{K}}$ leads to the identification of the secondary variables that are least interacting and most sensitive to the primary variables.

Fig. 3 shows the location of the three most informative temperatures determined from this approach. Similarly to the results obtained when using direct analysis of $\widehat{\mathbf{K}}$, the location of the most sensitive measurement point (Fig. 3a) changes during the batch, and all column trays seem to be equally suitable as temperature sensor locations. Only the reboiler and the top column tray could be considered slightly more relevant, since they correspond to the most sensitive measurement point for a longer period of time compared to the other available locations.

The results obtained for the second and third most sensitive measurements are more difficult to interpret. As for the second most sensitive measurement location (Fig. 3b), the SVD method suggests that it corresponds to the reboiler for almost the entire duration of the process. This result can be explained considering the fact that one of the estimated variables is the bottoms composition, and therefore the temperature obtained from the reboiler is inherently very informative during the entire operation. Furthermore, sensor interaction is taken into account in this approach, and biases the choice of the optimal sensor configuration. Thus, since the location of the most sensitive temperature usually corresponds to one of the column trays, the reboiler is selected as the second most sensitive point because the corresponding temperature measurement is likely to be the least interacting with the first sensor. The latter remark also suggests that the determination of the third most sensitive measurement location, which is required to have low interaction with measurements obtained from both a column tray *and* the reboiler, could be difficult. This conjecture is confirmed by the results reported in Fig. 3c, in which it can be observed that the third most sensitive location tends to change at each time instant. From these observations it is possible to conclude that the SVD analysis of $\widehat{\mathbf{K}}$ suggests placing one temperature sensor at the top tray and one in the reboiler. All the other sensors allowed should be evenly distributed along the column, since all the remaining possible locations result to be equally important from the sensitivity point of view.

5. Optimal sensor configuration from PCA sensitivity analysis

The results obtained from the proposed PCA sensitivity analysis of the sensitivity matrix $\widehat{\mathbf{K}}$ are shown in Fig. 4. At any given instant, the three most sensitive trays are located in the same section of the column, as occurred for the direct analysis of $\widehat{\mathbf{K}}$. However, in contrast to the results obtained from the direct analysis and

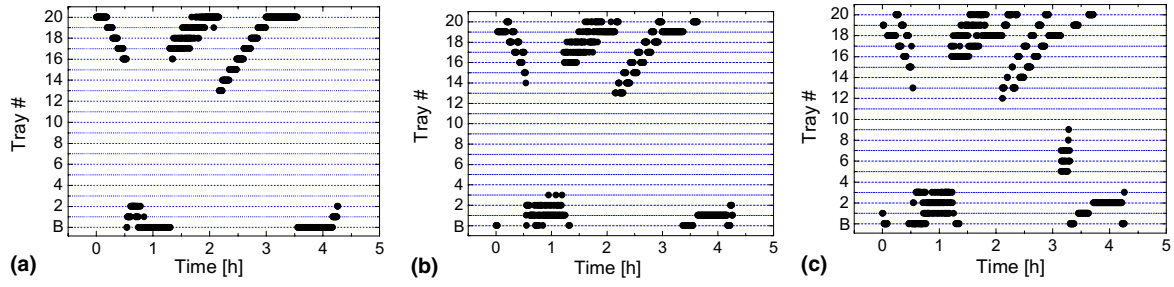

Fig. 4. Tray sensitivity information obtained from PCA analysis of $\widehat{\mathbf{K}}$: variation of (a) the most sensitive location, (b) the second most sensitive location, and (c) the third most sensitive location during the batch.

from the SVD approach, the optimal sensor locations identified using the PCA sensitivity analysis are clustered into regions of the column corresponding to its upper and lower sections only. The trays located in the central section of the column are indeed never designated as "important" measurement locations.

The proposed PCA sensitivity analysis also makes it possible to determine the *number* of measurement points that should be used as inputs to the soft sensor. In fact, the optimal size of the input measurement set corresponds to the number of the loadings of larger absolute value. In fact, because each loading represents a measure of the sensitivity of the corresponding temperature measurement to composition changes, a measurement should be selected only if its corresponding loading has a large absolute value, while all measurements whose loadings are much smaller that the largest one should be disregarded.

Fig. 5 reports the absolute values of the loadings calculated during the batch from the largest (p_1) to the smallest (p_{21}). The value of each loading changes during the process. However, only the first few loadings (from p_1 to p_5) have consistently large values relative to the others, thus indicating high-sensitivity locations (that may correspond to different column locations from sample to sample). Loadings p_6 to p_{15} can be considered of equally low importance, because their values are small and their time evolution is very similar, while loadings p_{16} to p_{21} are very small and indicate measurement points that contain very little information about the primary variables.

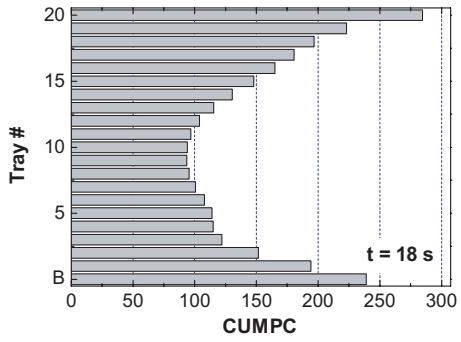

Fig. 6. Cumulative sensitivity index CUMPC for each sensor location for the reference batch.

These remarks suggest that five measurement points could be considered a suitable choice for this case study. According to this analysis, using a larger number of measurements should not result in any significant improvement in the accuracy of the composition estimator.

Within the high-sensitivity column regions determined above, the most suitable locations for sensor placement can be identified by calculating the cumulative index CUMPC in Eq. (9). The results in Fig. 6 suggest that the sensitive temperature measurement locations are the reboiler ("B"), and trays # 1, 18, 19, 20.

In the next sections, the effect of the duration of the sampling period Δt and of measurement noise on the characterization of the optimal input set via PCA

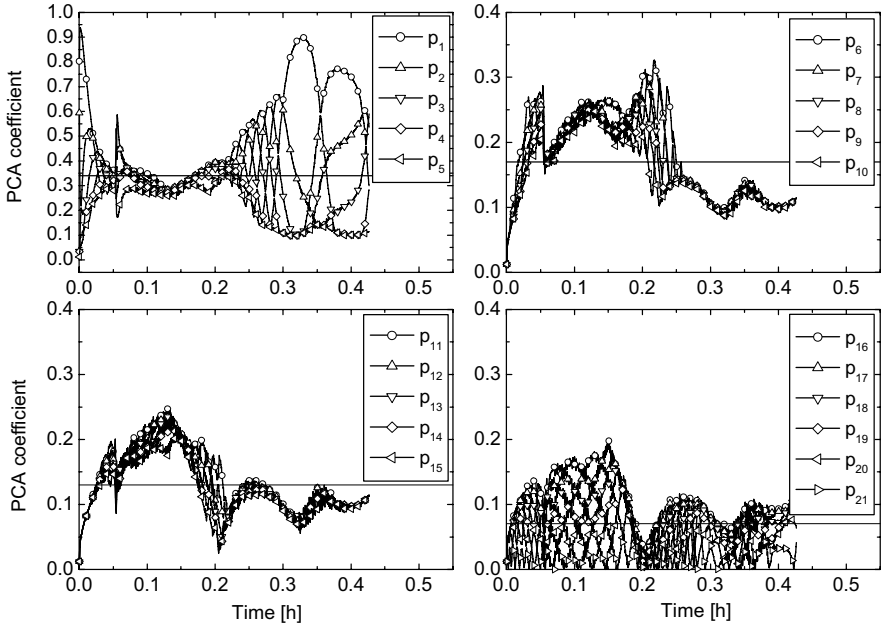

Fig. 5. Absolute value of the loadings calculated during the batch (for each set of loadings, the average value of the time trajectories is also indicated).

sensitivity analysis are evaluated. The results obtained using a different analytical formulation for the sensitivity matrix are also investigated.

5.1. Effect of the sampling interval

In principle, only small sampling intervals should be used when applying the PCA sensitivity analysis to the batch distillation process. Small values of Δt are desirable because each element of the instantaneous sensitivity gain matrix is approximated by a finite difference value. Using a large Δt leads to a less accurate estimation of $\mathbf{K}$, thus could potentially affect the results of the PCA sensitivity analysis.

Despite these concerns, Fig. 7 demonstrates that the results obtained from PCA sensitivity analysis are only marginally influenced by the length of the sampling interval. The value of Δt does affect the value of the cumulative sensitivity index, since the value of CUMPC for each tray changes with increasing length of Δt. However, the relative importance of each sensor location with respect to the other locations remains essentially unchanged.

The robustness of PCA sensitivity analysis to the sampling interval confirmed by these results is advantageous, as it guarantees that this method provides consistent results even when a fairly large sampling interval is used to collect the temperature and composition measurements required to calculate the instantaneous sensitivity matrix.

5.2. Effect of measurement noise

Normally distributed noise with zero mean and standard deviation σ was added to the temperatures in order to determine whether measurement noise can bias the results of the PCA sensitivity analysis.

Fig. 8a–c shows the results obtained for the CUMPC index when the soft sensor inputs are corrupted by noise (from low-level noise, $\sigma = 0.1$ °C, to high-level noise, $\sigma = 0.5$ °C). This measurement noise does affect the outcome of the PCA sensitivity analysis, because the profile of CUMPC tends to flatten at increasing noise levels, thus making it more difficult to rank the available temperature measurements and identify the most sensitive ones. As confirmed by Fig. 8c, when all temperatures are affected by relatively high-level noise, the PCA sensitivity analysis suggests that the secondary measurements are almost equally sensitive to the product compositions.

The detrimental effect of measurement noise can however be easily and effectively counteracted through appropriate adjustment of the sampling interval. As can be observed in Fig. 8a, e and i, the CUMPC profile remains practically unaltered when a larger sampling interval is adopted for larger measurement noise level, and the PCA sensitivity analysis provides the same indications obtained when noise free data were used. Due to the inherent robustness of PCA sensitivity analysis to the sampling interval, which was shown in Section 5.1, no disadvantage occurs for the selection of a larger sampling interval that is appropriate to the level of the measurement noise.

5.3. Effect of the sensitivity matrix formulation

In Eq. (1), the sensitivity gain has been defined as the partial derivative of a secondary variable with respect to a primary variable. As an alternative to this characterization, the sensitivity gain could be expressed as the

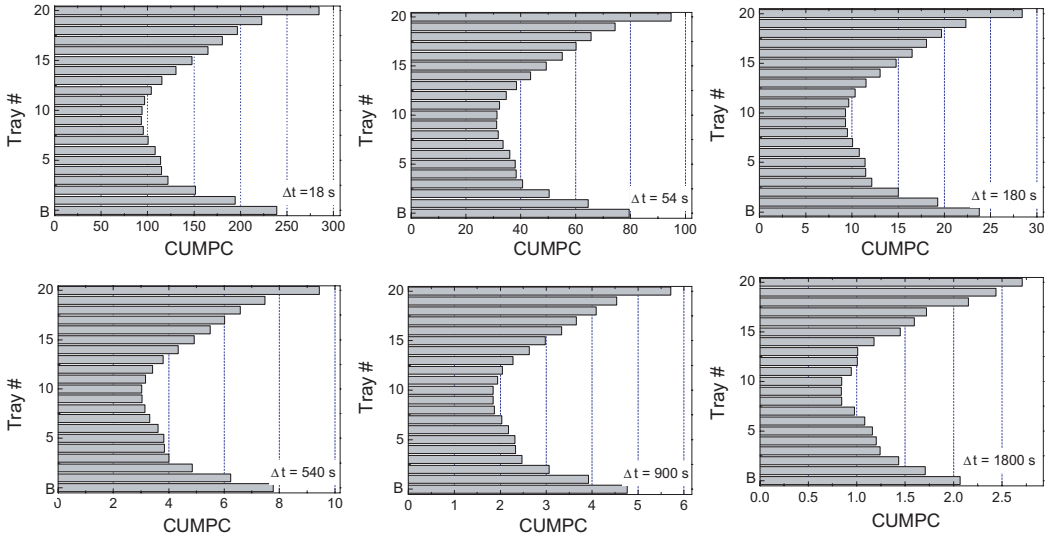

Fig. 7. Effect of the sampling interval Δt on cumulative sensitivity index CUMPC.

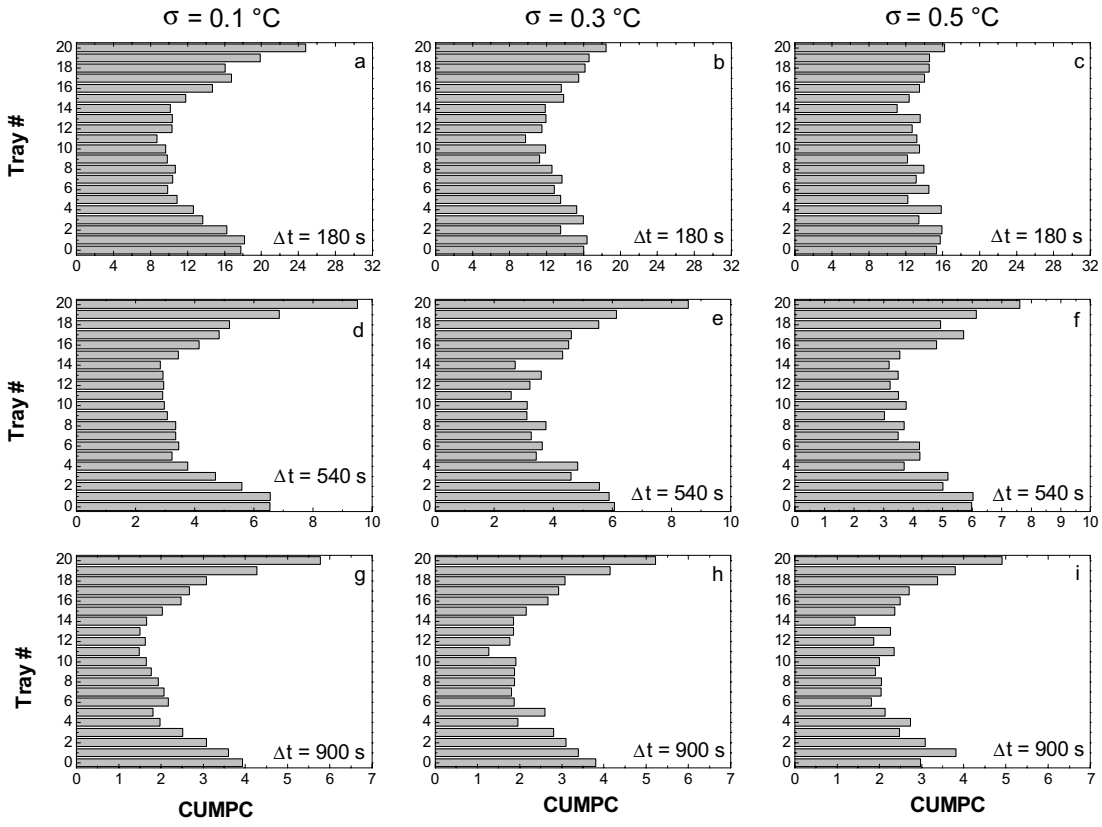

Fig. 8. Effect of measurement noise on the cumulative sensitivity index CUMPC.

partial derivative of a primary variable with respect a secondary variable. In this case, the resulting sensitivity matrix is:

$$\mathbf{K}^{\text{inv}} = \begin{bmatrix} \frac{\partial x_1}{\partial T_1} & \cdots & \frac{\partial x_1}{\partial T_j} & \cdots & \frac{\partial x_1}{\partial T_n} \\ \vdots & & \vdots & & \vdots \\ \frac{\partial x_i}{\partial T_1} & \cdots & \frac{\partial x_i}{\partial T_j} & \cdots & \frac{\partial x_i}{\partial T_n} \\ \vdots & & \vdots & & \vdots \\ \frac{\partial x_m}{\partial T_1} & \cdots & \frac{\partial x_m}{\partial T_j} & \cdots & \frac{\partial x_m}{\partial T_n} \end{bmatrix}, \tag{11}$$

which can be approximated at each time instant t as:

$$\widehat{\mathbf{K}}^{\text{inv}}(t) = \begin{bmatrix} \frac{\Delta x_1}{\Delta T_1} & \cdots & \frac{\Delta x_1}{\Delta T_j} & \cdots & \frac{\Delta x_1}{\Delta T_n} \\ \vdots & & \vdots & & \vdots \\ \frac{\Delta x_i}{\Delta T_1} & \cdots & \frac{\Delta x_i}{\Delta T_j} & \cdots & \frac{\Delta x_i}{\Delta T_n} \\ \vdots & & \vdots & & \vdots \\ \frac{\Delta x_m}{\Delta T_1} & \cdots & \frac{\Delta x_m}{\Delta T_j} & \cdots & \frac{\Delta x_m}{\Delta T_n} \end{bmatrix}. \tag{12}$$

We will refer to $\widehat{\mathbf{K}}^{\text{inv}}$ as the "inverse" sensitivity matrix, to distinguish it from the "direct" sensitivity matrix $\widehat{\mathbf{K}}(t)$ in Eq. (4).

As shown in Fig. 9, the results obtained from the application of the PCA analysis to the inverse sensitivity

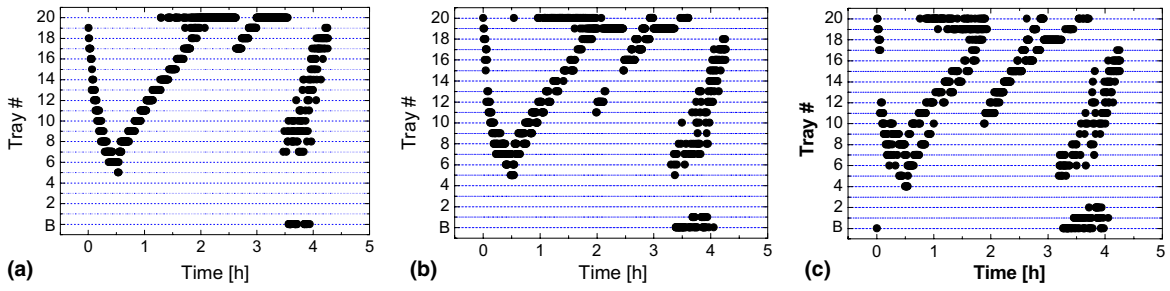

Fig. 9. Tray sensitivity information obtained from PCA analysis of $\widehat{\mathbf{K}}^{\text{inv}}$: variation of the (a) most sensitive location, (b) second most sensitive location, and (c) third most sensitive location during the batch.

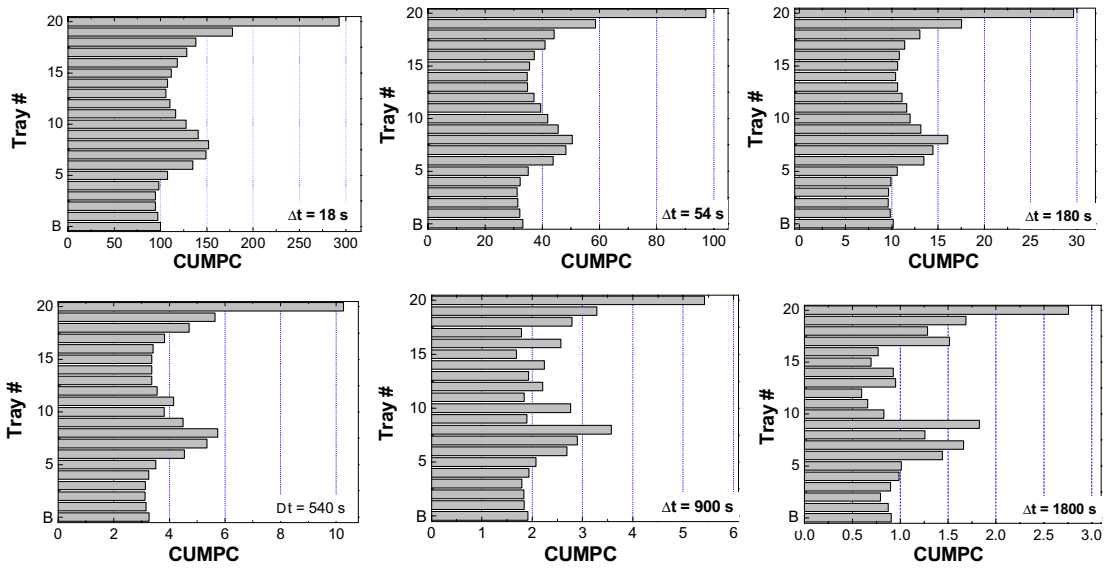

Fig. 10. Effect of the sampling interval Δt on the cumulative sensitivity index CUMPC (PCA analysis of $\widehat{\mathbf{K}}^{\text{inv}}$).

matrix for the benchmark batch distillation column are clearly different from the results obtained previously when the direct gain is employed. Thus, a different characterization of the column sensitivity affects the information content of the sensitivity matrix.

The location of the three most informative measurement points varies during the operation. In general, the optimal locations correspond to the top trays at the beginning of the operation, shift down to the column during the first part of the batch process, and then reverse this trend reaching to the top trays again. It is interesting to note that in this case the reboiler and the bottom column trays are never considered as sensitive measurement points, despite the fact that one of the primary variables is the composition of the bottom product.

As shown in Fig. 10, the sensitivity index CUMPC ($\Delta t = 18$ s) indicates that the overall optimal sensor locations correspond to trays # 7, 8, 9, 19, 20, when five measurements points are allocated. This result however is affected by the sampling interval used to collect the data. Also, because the value of the cumulative sensitivity index CUMPC_i for the ith location decreases with increasing Δt, the variation of the sampling period alters the relative importance of the available temperature measurements. As a result, the characterization of the optimal temperature set changes at different values of Δt. This is a major disadvantage, and it suggests that this sensitivity gain formulation is not appropriate.

6. Development of a composition estimator using alternative sensor configurations

In order to assess which measurement selection strategy is the most effective one among the ones considered so far, composition soft sensors have been developed using temperature measurements from the optimal sensor configurations identified considering different approaches of sensitivity analysis. In particular, linear PLS, nonlinear PLS, and ANN models have been evaluated. For the PLS estimators, three latent variables were retained in the regression models, this number having been determined using cross-validation. As shown by Zamprogna et al. [28], this approach inherently rejects the effect of noise in the temperature measurements, because random noise is typically associated with the higher-order latent variables, and is therefore eliminated when the original data are projected onto a lower dimensional space. Therefore, measurement noise will not be considered in the following example, because it has a negligible effect on the accuracy of composition estimations. Note however that noise may impact (even markedly) the performance of the ANN estimator.

To allow for a wider comparison, several alternative randomly chosen configurations were also considered. The results obtained for four of these supplementary configurations (AC_1, AC_3, AC_4, and AC_5) have been reported, as a representation of the estimation performances that are typically achieved when no specific measurement selection strategy is adopted. All the configurations considered are collected in Table 2.

Temperature measurements from each selected configuration are used by the soft sensors to estimate the light and intermediate component mole fraction in the distillate stream and the heavy component mole fraction in the reboiler during the entire duration of the batch. The most effective measurement selection approach is the one that leads to the soft sensor that has the lowest estimation error MSQ for the validation data. The

Table 2
Summary of the optimal sensor locations obtained using different measurement selection approaches

Measurement selection approach	Symbol	Most sensitive locations (tray #)				
Even distribution [23]	ED	B	05	10	15	20
Direct sensitivity analysis	DA≡ED	B	05	10	15	20
SVD sensitivity analysis	SVD≡ED	B	05	10	15	20
PCA sensitivity analysis (optimal location)	OL	B	01	18	19	20
Alternative configuration #1	AC_1	B	01	02	03	20
Alternative configuration #2	AC_2	03	12	15	16	19
Alternative configuration #3	AC_3	01	07	08	18	19
Alternative configuration #4	AC_4	B	02	04	05	12
PCA sensitivity analysis using inverse gain matrix	AC_5	07	08	09	19	20

operating conditions for the validation data are reported in Table 3.

Fig. 11 represents the values of the prediction error MSQ calculated for the validation data for the PLS and ANN soft sensors using input data from all the considered measurement configurations. The most accurate estimation performance is obtained by using the same measurement set for all the soft sensors. The best configuration is OL of Table 2, because it has the minimum value of the total MSQ for both the PLS soft sensors and the ANN soft sensor.

Table 3
Operating conditions for the validation data

Mixture relative volatility, $\alpha_1/\alpha_2/\alpha_3$	9/3/1
Feed composition, $x_{F,1}/x_{F,2}/x_{F,3}$	0.33/0.50/0.17
Feed charge, F	300 mol
Vapor boilup rate, V	70 mol/h
Distillate withdrawal rate, D	40.54 mol/h
Nominal composition setpoint, $x_{P1}^{sp}/x_{P2}^{sp}/x_{P3}^{sp}$	0.95/0.95/0.95

A very large estimation error in the distillate compositions occurs when only temperature measurements in the lower section of the column are used (configuration AC_4). Configuration AC_2 usually leads to models having relatively large values of MSQ. These results suggest that (as is intuitively expected) it is not desirable to exclude the temperature measurements located in the reboiler from the optimal input set (as in configurations AC_2, AC_3 and AC_5). In general, reduced estimation performance results when using input data from configuration AC_3, which includes temperature measurements located in the central section of the column. This observation confirms that these locations are poorly representative of the output variables, as noted in Section 5. Poor estimation performance is observed also when using configuration AC_5. This fact suggests that

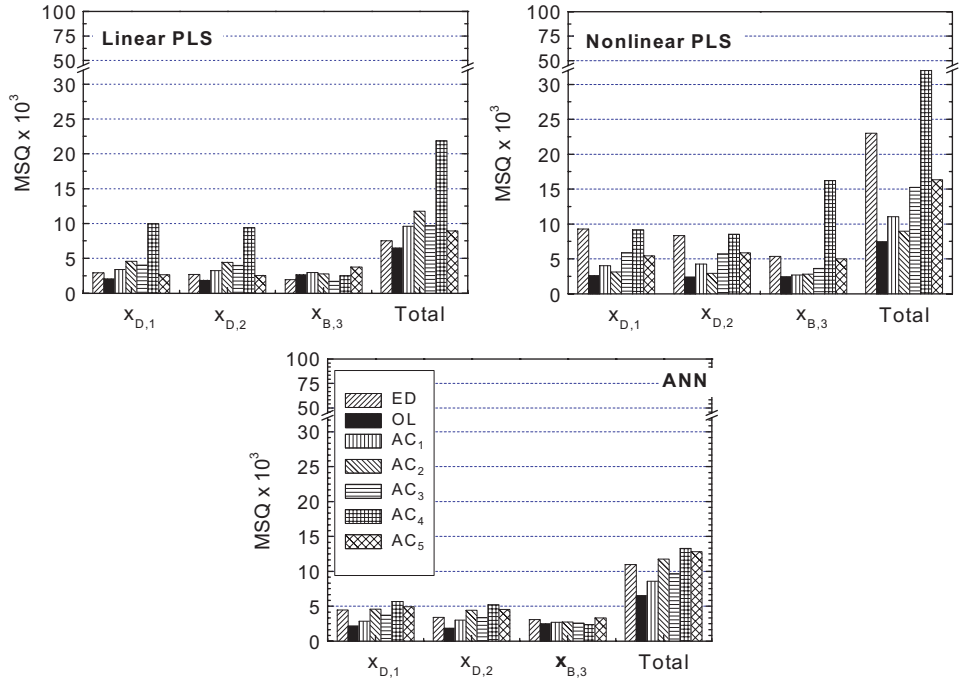

Fig. 11. Validation error MSQ for a linear PLS, a nonlinear PLS, and an ANN soft sensors when using five temperature measurements as model inputs.

that the indications obtained by performing the PCA sensitivity analysis for the inverse gain matrix are in this case misleading.

Fig. 11 also shows that all the soft sensors using configuration OL provide the overall most accurate composition estimation, and are almost equally accurate. The linear PLS soft sensor can be considered to be the most suitable one for this case study, because it shows the lowest MSQ error and has a simpler structure compared to the nonlinear PLS or ANN estimators.

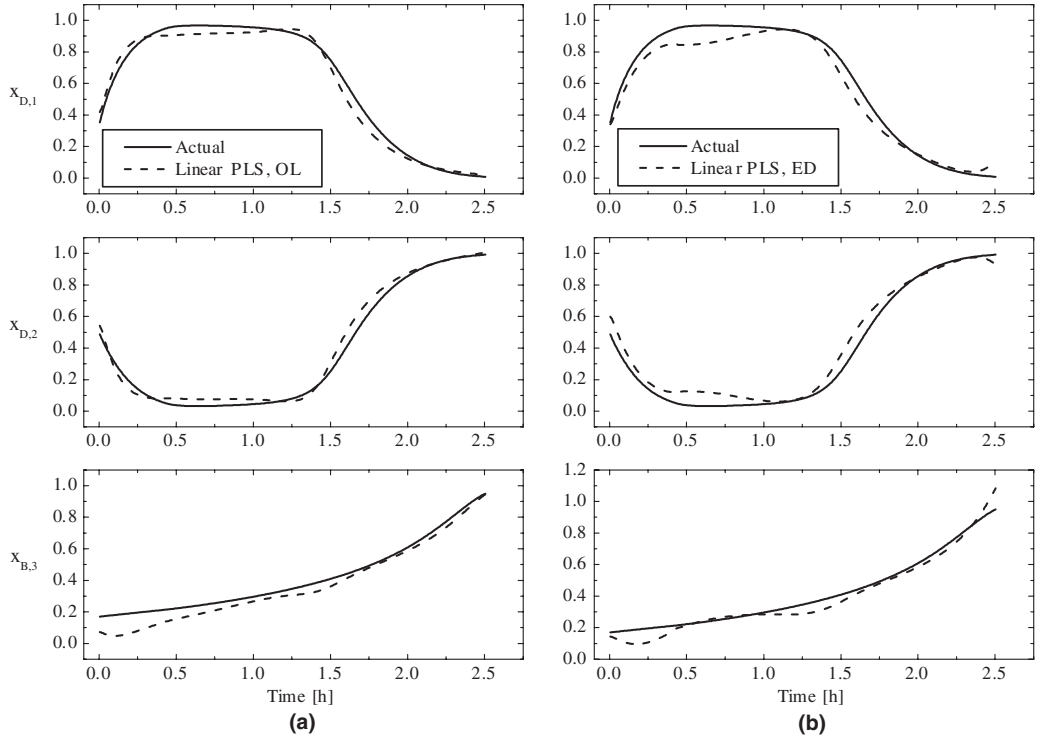

Fig. 12. Validation data: comparison between the product compositions and their estimates provided by a linear PLS soft sensor using temperature measurements from configuration OL (a) and configuration ED (b).

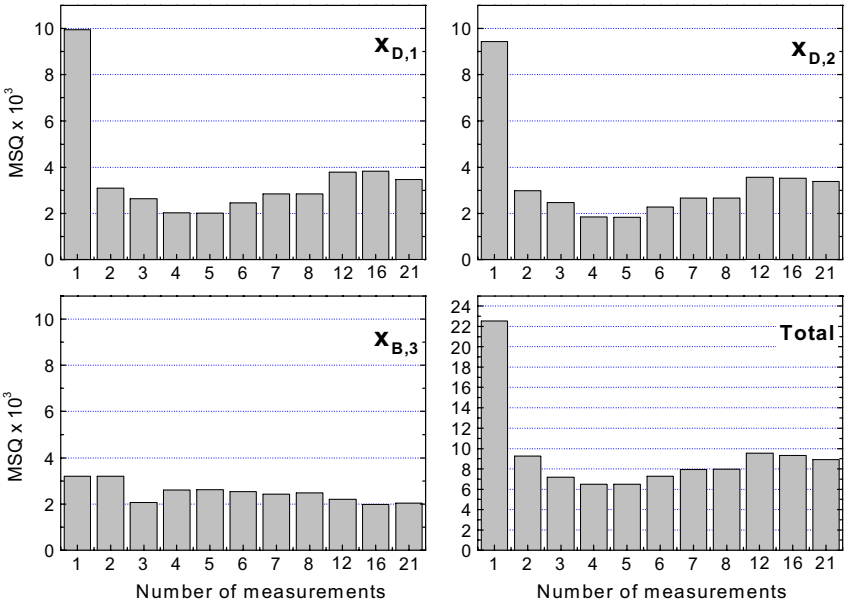

Fig. 13. MSQ validation error for linear PLS soft sensors using configuration OL with different numbers of temperature measurements.

This topic has been further discussed by Zamprogna et al. [28].

Fig. 12 compares the actual value of the product compositions and their estimates calculated by the linear PLS model for the validation data. This figure provides further confirmation that configuration OL provides good estimates of the composition profiles.

In particular, the estimation accuracy is higher than what can be obtained when using the configuration usually suggested (configuration ED) [23], where the temperature sensors are evenly distributed along the column.

6.1. Effect of the number of temperature measurements

The PCA sensitivity analysis carried out in Section 5 suggests a priori that the optimal number of temperature measurements for the regression model is five. In order to verify this result, linear PLS estimators have been developed using configurations having a different number of measurements.

The temperature measurements for each configuration have been selected according to the location ranking suggested by the PCA sensitivity analysis. The estimation error calculated for the linear PLS models obtained are represented in the form of bar plots in Fig. 13.

For the estimation of $x_{D,1}$ and $x_{D,2}$ in Fig. 13, the value of MSQ depends quite markedly upon the number of temperatures incorporated in the optimal set. The minimum MSQ is indeed achieved by a soft sensor using the configuration OL with five measurements, as indicated by the PCA sensitivity analysis (a soft sensor using four input measurements provides almost the same accuracy, however). The number of measurements included in configuration OL affects only marginally the accuracy of estimation $x_{B,3}$; the MSQ index for this primary variable shows however a minimum when three temperature measurements are considered.

7. Conclusions

A novel methodology has been proposed in order to identify the most suitable number and locations of temperature measurements to be used as soft sensor inputs for estimating composition profiles in a batch distillation column. The proposed approach is based on: (i) the characterization of the instantaneous sensitivity of each secondary variable to the primary variables, and (ii) on the identification of the most sensitive secondary variables from this sensitivity matrix by exploiting the properties of the PCA transformation.

The simulation results have shown that the proposed approach can effectively help to select the most informative secondary process variables, leading to a soft sensor with good estimation performance. It has also been shown that the length of the sampling interval affects the results obtained from the PCA sensitivity analysis only marginally. Even though the presence of measurement noise can make it more difficult to rank the available temperature measurements and to identify the most sensitive ones through the PCA sensitivity analysis, it was shown that the detrimental effects of measurement noise can be counteracted through appropriate adjustment of the sampling interval.

The proposed methodology can be easily extended to other batch processes, and to distributed parameter systems. In this regard, interesting results are being obtained for the optimal selection of input measurements in tubular reactors, and the results will be reported elsewhere.

Acknowledgements

This research was carried out in the framework of the MIUR-PRIN 2002 project "Operability and controllability of middle-vessel distillation columns" (ref. no. 2002095147_002).

References

[1] M. Barolo, F. Berto, Composition control in batch distillation: binary and multicomponent mixtures, Ind. Eng. Chem. Res. 37 (1998) 4689–4698.
[2] M. Barolo, A. Pistillo, A. Trotta, Issues in the development of a composition estimator for a middle vessel batch column, in: L.T. Biegler, A. Brambilla, C. Scali (Eds.), Advanced Control of Chemical Processes 2000—IFAC ADCHEM 2000, Elsevier, Oxford, UK, 2000, pp. 923–928.
[3] B.W. Bequette, T.F. Edgar, Non-interacting control system design methods in distillation, Comp. Chem. Eng. 13 (1989) 641–650.
[4] B.W. Bequette, T.F. Edgar, Selection of process measurements in distillation column control to minimize multivariable interactions, AIChE Annual Meeting, San Francisco, USA, 1984.
[5] C. Brosilow, B. Joseph, Techniques of Model Based Control, Prentice Hall, New York, USA, 2002.
[6] I. Chien, B.A. Ogunnaike, Modeling and control of a temperature-based high-purity distillation column, Chem. Eng. Commun. 158 (1997) 71–105.
[7] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[8] C. Georgakis, D.H. Kindt, M. Kasotaki, Extensive variable control structures for binary distillation columns, AIChE Annual Meeting, San Francisco, USA, 1984.
[9] S.S. Haykin, Neural Networks: A Comprehensive Foundation, second ed., Prentice Hall, New York, USA, 1999.
[10] J.E. Jackson, A User's Guide to Principal Components, John Wiley & Sons, New York, USA, 1991.
[11] B. Joseph, C.B. Brosilow, Inferential control of processes. Part I: steady state analysis and design, AIChE J. 24 (1978) 485–492.
[12] R.E. Kalman, A new approach to linear filtering and prediction problems, Trans. ASME, J. Basic Eng. 82 (1960) 35–45.

[13] M. Kano, K. Miyazaki, S. Hasebe, I. Hashimoto, Inferential control system of distillation compositions using dynamic partial least squares regression, J. Process Control 10 (2000) 157–166.
[14] T. Kourti, J.F. MacGregor, Tutorial: Process analysis, monitoring and diagnosis, using multivariate regression methods, Chemom. Intell. Lab. Syst. 28 (1995) 3–21.
[15] H. Leegwater, Industrial experience with double quality control, in: W.L. Luyben (Ed.), Practical Distillation Control, Van Nostrand Reinhold, New York, USA, 1992.
[16] D.C. Luenberger, Observing the state of a system, IEEE Trans. Military Electron. MIL-8 (1964) 74–80.
[17] W.L. Luyben, Multicomponent batch distillation.1. Ternary systems with slop recycle, Ind. Chem. Eng. Res. 27 (1991) 642–657.
[18] T. Mejdell, S. Skogestad, Estimation of distillation compositions from multiple temperature measurements using partial-least-squares regression, Ind. Eng. Chem. Res. 30 (1991) 2543–2555.
[19] C.F. Moore, Selection of controlled and manipulated variables, in: W.L. Luyben (Ed.), Practical Distillation Control, Van Nostrand Reinold, New York, USA, 1992.
[20] R.M. Oisiovici, S.L. Cruz, Sensor location and noise reduction in high-purity batch distillation control loops, Braz. J. Chem. Eng. 17 (2000) 671–683.
[21] R. Osiovici, S.L. Cruz, Inferential control of high-purity multicomponent batch distillation columns using an extended Kalman filter, Ind. Eng. Chem. Res. 40 (2001) 2628–2639.
[22] S.J. Qin, Neural network for intelligent sensors and control—practical issues and some solutions, in: O. Omidvar, D.L. Elliott (Eds.), Neural Systems for Control, Academic Press, New York, USA, 1997.
[23] E. Quintero-Marmol, W.L. Luyben, C. Georgakis, Application of an extended Luenberger observer to the control of multicomponent batch distillation, Ind. Chem. Eng. Res. 30 (1991) 1870–1880.
[24] D.E. Seborg, T.F. Edgar, D.A. Mellichamp, Process Dynamics and Control, second ed., John Wiley & Sons, New York, USA, 2004.
[25] T.L. Tolliver, L.C. McCune, Distillation control design based on steady state simulation, ISA Trans. 17 (1978) 3–10.
[26] E. Zamprogna, M. Barolo, D.E. Seborg, Composition estimations in a middle-vessel batch distillation column using artificial neural networks, Chem. Eng. Res. Des. 79 (2001) 689–696.
[27] E. Zamprogna, Development of virtual sensors for batch distillation monitoring and control using multivariate regression techniques, Ph.D. Dissertation, Department of Chemical Engineering Principles and Practice, University of Padova, Italy, 2001.
[28] E. Zamprogna, M. Barolo, D.E. Seborg, Estimating product composition profiles in batch distillation via partial-least-squares regression, Control Eng. Practice 12 (2004) 917–929.