AIChE

A Multivariate Statistical Process Control Procedure for BIAS Identification in Steady-State Processes

Mabel C. Sánchez, Carlos R. Alvarez, and Adriana Brandolin
Planta Piloto de Ingeniería Química (CONICET - UNS), Camino La Carrindanga Km 7, 8000 Bahía Blanca, Argentina

DOI 10.1002/aic.11547
Published online June 10, 2008 in Wiley InterScience (www.interscience.wiley.com).

In this article, a multivariate statistical process control (MSPC) strategy, devoted to bias identification and estimation for processes operating under steady-state conditions, is presented. The technique makes use of the D *statistic to detect the presence of biases. Besides, it uses a new decomposition of this statistic to identify the faulty sensors. The strategy is based only on historical process data. Neither process modeling nor assumptions about the probability distribution of measurement errors are required. In contrast to methods based on fundamental models, both redundant and nonredundant measurements can be examined to identify the presence of biases. The performance of the proposed technique is evaluated using data-reconciliation benchmarks. Results indicate that the technique succeeds in identifying single and multiple biases and fulfills three paramount issues to practical implementation in commercial software: robustness, uncertainty, and efficiency.*
AIChE J, 54: 2082–2088, 2008
Keywords: data reconciliation, statistical analysis

Introduction

Basic and high-level plant activities, such as monitoring, regulatory and supervisory control, real-time optimization, planning, and scheduling, provide valuable results only if a reliable knowledge of current plant-state is at hand. Because measurements are subject to random and gross errors, a great effort has been made during the last four decades to reduce their detrimental effects on the estimation of process variables. Nowadays, it is a common practice in process industries to optimally adjust measurements in such a way that the corrected values are consistent with mass and energy balances. But, to obtain accurate estimates, some action should be taken to reduce the influence of gross errors, such as measurement biases. They can be caused by many sources, for instance, poorly calibrated or malfunctioning instruments.

Several model-based approaches have been proposed for bias detection and identification. They compare the actual operation of the plant with that predicted by a mathematical model by means of statistical hypothesis tests. The most widely used tests are the Global Test, the Measurement Test, the Nodal Test, the Generalized Likelihood Ratio Test, the Bonferroni Test, and the Principal Component Test, among others. Three types of strategies have been developed for identifying multiple gross errors: (a) serial elimination identifies one gross error at a time and eliminates the corresponding measurement; it goes on until no gross error is detected; (b) serial compensation isolates one gross error and evaluates its size and then compensates the measurement and continues until no gross error is found; (c) simultaneous or collective compensation proposes the identification and estimation of all gross errors simultaneously.[1,2] Furthermore, the equivalence theory of gross errors[3] establishes that two sets of gross errors are equivalent when they have the same effect in data reconciliation. Therefore, when a set of biases is identified, there exists an equal possibility that the true location of gross

Correspondence concerning this article should be addressed to M. C. Sánchez at msanchez@plapiqui.edu.ar

errors is in one of its equivalent sets. This issue introduces uncertainty in the identification task.

To avoid biased estimations of process variables, other strategies incorporate the nonideality of the data distribution in the formulation of the data-reconciliation problem. Thus, random and gross errors are removed simultaneously based on their probability distribution. This is usually accomplished by combining nonlinear programming and the maximum-likelihood principle, after the error distribution has been suitable characterized.[4–6]

In Statistical Process Control, the Hoteling statistic (D) is widely used to reliably detect the out-of-control status, but offers no assistance as identification tool. To perform this task, different techniques are presented that evaluate the contribution of each variable to the value of the statistic when it exceeds the critical one. Some strategies work in the space defined by the original variables and others make use of latent variables. Regarding this type of methods, Jackson[7] proposed the decomposition of the statistic into the sum of principal components. If they represent a meaningful grouping of variables, the identification of out-of-control signals is readily apparent.[8] However, in many examples, it is difficult to associate a meaning to a principal component. Hence, Miller et al.[9] and MacGregor et al.[10] proposed to evaluate the contribution of each process variable to the scores that are outside of their confidence limits, and Pullen et al.[11] presented an approach to calculate the variable contributions to the D statistic instead of to the scores. Furthermore, Westerhuis et al.[12] extended the theory of contribution plots to latent-variable models with correlated scores. The authors introduced control limits for variable contributions that help in finding the measurements, which present a different behavior with respect to that contained in the reference data set.

Another approach for calculating variable contributions to the D statistic is carried out in the original-variable space. Mason et al.[13,14] presented a strategy to decompose the D statistic into the contributions of each variable, which is intended to identify the fault source. Because of the combinatorial nature of the formulation, a great number of decompositions are obtained, increasing the complexity of the identification procedure. A straightforward method to decompose the D statistic as a unique sum of each variable contribution was recently developed by Alvarez et al.,[15] which is called OSS (original-space strategy). It provides a clear understanding of positive and negative variable contributions and estimates a bound for the negative ones.

In this work, a strategy for bias detection, identification and estimation devoted to process operating under steady-state conditions is presented. It is based on the multivariate statistical process control (MSPC) technique developed by Alvarez et al.[15] It is wholly developed using historical process data; thus, no fundamental or empirical modeling and no assumptions about the probability distribution of measurement errors are required. In contrast to the techniques based on fundamental models, both redundant and nonredundant measurements can be examined to identify the presence of biases.

The rest of this work is organized as follows. In the next section, the decomposition of the D statistic in terms of the contributions of each variable is briefly reviewed. The proposed strategy for bias identification and estimation is presented in A Strategy for Bias Identification and Estimation section. Then, the procedure applied to evaluate the performance of the strategy is described. In Examples and Results section, results for some benchmarks extracted from data-reconciliation literature are included. Conclusions are presented in the last section.

Decompostion of the *D* Statistic

Let us consider $\mathbf{z}$ represents an N dimensional vector of measurements made on a process operating under steady-state conditions at time t. Given a reference population containing I vectors of observations, the population parameters can be estimated using the sample mean vector ($\bar{\mathbf{z}}$) and the sample covariance matrix ($\mathbf{S}$).

A multivariate control chart for the process is based on the D statistic, which has the form

$$D = (\mathbf{z} - \bar{\mathbf{z}})^{\mathbf{T}} \mathbf{S}^{-1} (\mathbf{z} - \bar{\mathbf{z}}). \tag{1}$$

As it is shown in Eq. 1, the statistic has a quadratic form, and its value is always equal or greater than zero considering that the sample covariance matrix is positive semidefinite.

The D statistic can be formulated as

$$D = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{i,j} x_i x_j, \tag{2}$$

where $a_{i,j}$ are the elements of the inverse of the covariance matrix ($\mathbf{A} = \mathbf{S}^{-1}$) and

$$x_k = z_k - \bar{z}_k, \tag{3}$$

for $k = 1, \ldots, N$.

Alvarez et al.[15] reformulated the D statistic in terms of the variables x_k as follows

$$D = \sum_{k=1}^{N} a_{k,k} \left(x_k^2 - x_k^* x_k \right), \tag{4}$$

where x_k^* represents the value of x_k for which D is minimum given the ($N - 1$) values of the remaining variables and is calculated using the following formula

$$x_k^* = - \frac{\sum_{\substack{j=1 \\ j \neq k}}^{N} a_{k,j} x_j}{a_{k,k}}. \tag{5}$$

An important advance over previous works is that the D statistic can be evaluated as a unique sum of the contributions of each variable, c_k^D ($k = 1, \ldots, N$)

$$D = \sum_{k=1}^{N} c_k^D. \tag{6}$$

Each contribution has a quadratic form, and its roots are located at $x_k = 0$ and $x_k = x_k^*$. Equation 4 provides a straightforward decomposition of the D statistic into the contributions of each variable. They are obtained in the space of the original variables using the same formulation for all of them.

A Strategy for BIAS Identification and Estimation

In this work, a MSPC procedure is presented to identify measurement biases and estimate their magnitudes for a steady-state process. Bias identification is based on the decomposition of the D statistic previously presented. First, the strategy is described, and then some issues with regard to its application are discussed.

Strategy description

The strategy consists of two stages. The first one, which is executed offline, provides the in-control parameters of the process and the critical value of the statistic. They are obtained from a reference population.

During the second stage, the D statistic is calculated online at each sample time. If it exceeds the threshold value, the presence of one or multiple biases is detected. Then, the evaluation of the contributions of each variable to the statistic allows identifying the faulty sensors. Bias magnitudes are approximately estimated by calculating the difference between the measurement vector and the reference-population mean vector for the positions related with the faulty sensors. Each stage is composed of the following steps.

Off-Line Stage.

1. The reference population, constituted by I vectors of observations, is built. Each sample is obtained as an average of the measurements corresponding to a moving window of length H.
2. The population parameters $\bar{\mathbf{z}}$ and $\mathbf{S}$ are evaluated.
3. The D statistic is calculated for each member of the reference population. After that, the probability density function of D is estimated using a kernel smoothing technique. Given the significance level of the test, α, the $(1 - \alpha)$ percentile of the distribution is selected as the critical value of D (D_{crit}).

On-Line Stage.

1. Detection: at each sample time, a new observation $\mathbf{z}_{\mathrm{new}}$ is obtained as the average of the measurements included in a moving window of length H. The D statistic is calculated for $\mathbf{z}_{\mathrm{new}}$ using Eq. 1. If it is greater than D_{crit}, the presence of biases is detected. The initial value of D_{crit} is used until enough data are collected in the second stage. Then, that value should be adjusted to avoid overestimation of the fault signal rate α, as it is indicated by Chou et al.[16]
2. Identification: variable contributions to the statistic value are calculated using Eq. 4. Then, they are compared with a single threshold value τ, which is defined as follows:

$$\tau = cont + \beta\sigma_{\mathrm{cont}}. \qquad (7)$$

In this equation, $cont$ is the mean of the variable-contribution vector, and σ_{cont} is the mean of the standard deviations of variable contributions in the reference population. Because the average number of Type I errors of the procedure (AVTI) is set equal to 0.1 under the null hypothesis, the parameter β is calculated by simulation to satisfy this condition. In this way, the number of false alarms is bounded for each value of the parameter I.

These contributions that exceed the threshold value are associated with faulty sensors.

3. Estimation: bias magnitudes are approximated by the difference between the faulty measurements and their sample means.

Application issues

In this subsection, particular issues with regard to the application of the proposed technique are pointed out:

a. The generation of the reference population can be easily accomplished. The increasing automation of process industries provides the required information for data-driven techniques as the proposed one. Furthermore, the estimation of in-control parameters for the reference population can be updated online by incorporating new unbiased measurements. This enhances the performance of the strategy for bias identification and estimation, as it is shown in Examples and Results section.

b. The formulation of steady-state models (rigorous, empirical, or hybrid) for complex nonlinear processes remains a difficult and time-consuming task. As an alternative, the problem of identification of faulty sensors can be dealt with data-driven procedures, after performing the calibration of the instruments installed in the process.

c. If a steady-state model of the process is available, it usually involves redundant and nonredundant measurements. The last ones cannot be examined to identify the presence of biases, except using sensor voting or temporal redundancy. In contrast, the new technique is suitable to identify biases associated with both types of measurement.

d. As the Equivalency Theory of Gross Errors clearly states, there exists uncertainty about the results of bias identification strategies based on first-principle models. The true set of gross errors may be located on a set, which is equivalent to the predicted one. This problem does not arise when the proposed strategy is used.

e. No assumptions are required about the probability density function of measurement errors.

f. Parameter I has a major influence on the performance of the strategy. For a given process, the increment of I enhances the identification and estimation capabilities of the technique. Furthermore, I should be increased for problems of large size to achieve good performance measures, as it will be shown in Examples and Results section.

g. The technique provides the report of faulty sensors that should be inspected by the maintenance sector. During the interval between the fault identification and the sensor reparation, the value of the faulty measurement can be replaced by its sample mean and used as input for other software packages. If a model of the process is available and the faulty sensor corresponds to a redundant measurement, its value can be estimated in terms of reconciled measurements.

Simulation Procedure

A simulation procedure was applied to evaluate the performance of the proposed strategy for different benchmarks extracted from data-reconciliation literature.

One hundred random configurations of biases are tested in each simulation run. Also one hundred moving windows are examined for a particular configuration of biases. To form a moving window, data are generated by adding noise and

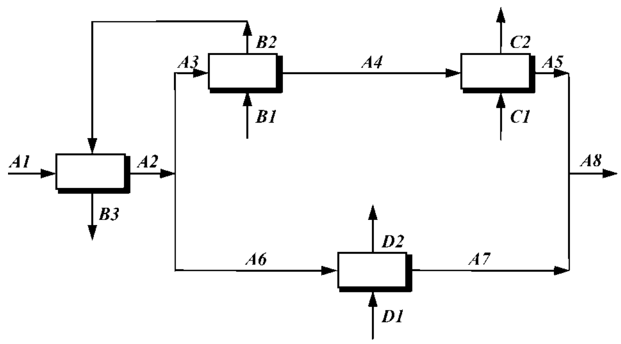

Figure 1. Heat exchanger network.

biases to the true values of the variables. A horizon of data of fixed length $H = 5$ is generated and arranged into a matrix. At each sample interval, the oldest vector of observations is dropped out of the matrix, and the new vector of measurements is appended as the last row of it.[5]

Regarding random errors, three different distributions (Normal, Uniform, and Laplacian) are used to create them. Kernel density estimation is applied to approximate the D statistic distribution, and D_{crit} is calculated assuming $\alpha = 0.05$.

The number, location, and size of gross errors for each configuration are selected randomly. The number of simulated gross errors is in the range [1–0.25 × N]. In contrast to previous works, their sizes are in the narrow range [4–7σ], where σ represents the standard deviation of the measurement errors.

Three well-known measures of identification performance are used: the Average Number of Type I Errors (AVTI), the Overall Power (OP), and the Expected Fraction of Perfect Identification (OPF). They are defined as follows:

$$\text{AVTI} = \frac{\text{No. of gross errors incorrectly identified}}{\text{No. of simulation trials}} \tag{8}$$

$$\text{OP} = \frac{\text{No. of gross errors correctly identified}}{\text{No. of gross errors simulated}} \tag{9}$$

$$\text{OPF} = \frac{\text{No. of trials with perfect identification}}{\text{No. of simulation trials}} \tag{10}$$

Since 100 different configurations of gross errors are simulated in each simulation run, the bias-estimation performance is quantified using relative errors.

A given configuration is associated with a row of the matrix **Er** (100, N). The row has zero and nonzero entries. The first ones correspond to the nonfaulty sensors of the configuration. The nonzero entries are the mean relative errors of the biases estimated for the moving windows with perfect identification.

The mean of each column of **Er**, disregarding its zero entries, represents the mean relative error of bias estimation among all configurations for a given measurement, and it takes part of the vector $\boldsymbol{\delta}_r$ of dimension N. The minimum and maximum values of $\boldsymbol{\delta}_r$ are reported for each simulation run.

Examples and Results

Two data-reconciliation benchmarks are used to analyze the performance of the procedure. One of them is the heat exchanger network (HEN), originally presented by Swartz,[17] and the other one corresponds to a steam metering network (SMN) for a methanol synthesis plant.[18]

Example 1

The HEN shown in Figure 1 is devoted to heat stream A by process streams B, C, and D at various junctions. The system has 16 measured variables and 14 unmeasured ones, which are related by 17 equality constraints (mass and energy balances around the heat exchangers, mixers and dividers). The standard deviations of flow rates and temperatures are 2% and 0.75°C, respectively. A complete description of the example can be found elsewhere.[2]

Variable classification procedures are applied to determine the sets of redundant and nonredundant measurements. They are presented in Table 1, where F means flow rate, T stands for temperature, R represents a redundant measurement, and NR indicates a nonredundant observation. It should be noticed that neither gross error identification nor measurement adjustment is possible for nonredundant measurements, because they do not participate in a set of redundant equations.

Table 2 presents the performance measures related to bias identification and bias estimation for gross errors located on both redundant and nonredundant observations. Random errors are generated using Normal, Uniform, and Laplacian distributions.

The second column of Table 2 indicates the maximum number of biases that are randomly generated for each simulation. The next two columns present the number of samples of the reference population (I) and the parameter β used to calculate the threshold value τ. The identification-performance measures AVTI, OP, and OPF follow. The minimum and maximum values of vector $\boldsymbol{\delta}_r$ are included in the last column.

When bias-identification techniques are running on-line in process industries and the presence of gross errors is detected, there is only one biased measurement in general. Table 2 shows that OP values are greater than 0.999 when

Table 1. Classification of Measurements for the HEN

Stream	Variable	Classification
A1	F	R
A1	T	NR
A3	F	R
A3	T	R
A4	T	NR
A5	T	R
A6	F	R
A7	T	R
A8	T	R
B1	F	NR
B1	T	NR
C1	F	NR
C1	T	NR
D1	T	NR
D2	F	R
D2	T	R

Table 2. Performance Measures for Example 1

Distribution	Max #B	I	β	AVTI	OP	OPF	Mean (ε_r) (%)
Normal	1	90	9.3	0.0265	1	0.9737	0.40–4.16
	1	120	7.7	0.0176	1	0.9825	0.64–2.77
	1	180	5.5	0.0125	1	0.9875	0.94–3.69
	1	300	4.8	0.0093	1	0.9907	0.61–2.49
	2	90	9.3	0.0175	0.9993	0.9814	0.61–4.07
	2	120	7.7	0.0107	0.9997	0.9888	1.17–3.06
	2	180	5.5	0.0054	0.9992	0.9935	1.13–3.12
	2	300	4.8	0.0053	0.9998	0.9944	0.95–2.35
	4	90	9.3	0.0130	0.9638	0.8995	1.30–6.00
	4	120	7.7	0.0055	0.9796	0.9432	1.29–5.34
	4	180	5.5	0.0020	0.9906	0.9742	1.20–3.41
	4	300	4.8	0.0025	0.9945	0.9836	0.89–2.65
Uniform	1	90	9.8	0.0315	1	0.9695	1.18–4.33
	1	120	7.9	0.0178	1	0.9824	0.80–5.67
	1	180	6.2	0.0120	1	0.9881	0.94–4.32
	1	300	4.9	0.0048	1	0.9952	0.41–4.35
	2	90	9.8	0.0166	0.9996	0.9832	0.98–4.04
	2	120	7.9	0.0153	0.9999	0.9848	1.38–4.75
	2	180	6.2	0.0068	0.9997	0.9927	1.07–4.37
	2	300	4.9	0.0019	1	0.9981	0.96–3.90
	4	90	9.8	0.0172	0.9692	0.9108	1.46–4.66
	4	120	7.9	0.0095	0.9708	0.9178	0.95–4.27
	4	180	6.2	0.0014	0.9861	0.9609	1.12–4.35
	4	300	4.9	0.0014	0.9888	0.9702	1.20–4.43
Laplacian	1	90	10.5	0.0634	0.9996	0.9388	0.82–5.24
	1	120	8.2	0.0528	0.9994	0.9475	0.75–6.82
	1	180	6.5	0.0401	0.9999	0.9603	1.01–5.53
	1	300	5.6	0.0373	0.9997	0.9630	1.02–5.97
	2	90	10.5	0.0548	0.9854	0.9625	1.17–4.78
	2	120	8.2	0.0432	0.9985	0.9562	1.66–5.82
	2	180	6.5	0.0320	0.9976	0.9655	1.43–5.35
	2	300	5.6	0.0270	0.9994	0.9726	1.31–4.75
	4	90	10.5	0.0397	0.9147	0.7763	2.01–12.3
	4	120	8.2	0.0260	0.9394	0.8326	1.45–7.51
	4	180	6.5	0.0156	0.9602	0.8788	1.49–7.07
	4	300	5.6	0.0168	0.9825	0.9408	1.71–5.66

one gross error is simulated for all random-error probability distributions. These figures indicate that the strategy successfully identifies the bias of the faulty sensor, which is of great practical importance. Moreover, simulations are performed reducing in 50% the lowest values of I for all random-error distributions, and the OP values remain above 0.999 for Max#B = 1. When the simulated number of biases increases, a slight diminution of OP is observed for the same values of I. Nevertheless, high values of OP are still achieved for Max#B = 0.25N.

Parameter I affects the measure OPF, that is, the ability of the strategy of indicating perfectly the set of faulty sensors; if I increases, OPF also increases. Thus, the updating of the reference population online, by incorporating in-control process observations, enhances the identification performance and reduces the number of suspected measurements that should be inspected by the maintenance sector.

Regarding the value of parameter β, it is adjusted to maintain an AVTI equal to 0.1 when the null hypothesis is satisfied. Its value decreases for larger samples, because, in these cases, better estimations of the population parameters are available.

In general, simulation results show that the maximum values of the mean relative errors of bias estimates diminish for large values of I. The same conclusion arises for all probability distributions. The highest mean relative errors are obtained for the Laplacian distribution, which is in agreement with the fact that it has longer thicker tails than the normal distribution.

As it is expected, the same performance of the strategy is achieved when biases are simulated exclusively for nonredundant measurements, because the MSPC approach does not distinguish between the two variable categories.

Example 2

SMN of a methanol synthesis plant involves 28 streams that interconnect 11 units (see Figure 2). The flow rates of all streams are measured, thus all measurements are redundant. Random errors are generated considering that the standard deviations of observations are 2.5% of their true values. Table 3 presents the performance measures for bias identification and estimation.

In this example, the number of measured variables increases in 75% with respect to the previous one. Therefore, the number of samples of the reference population is also increased to achieve high OPFs. Populations made up of 150, 250, 350, and 500 samples are generated assuming Normal, Uniform, and Laplacian random errors.

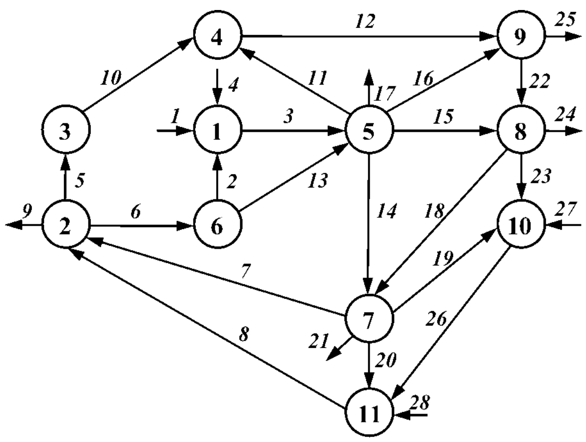

Figure 2. Steam metering network.

The effect of the sample size on OP and OPF is confirmed. Parameter I affects slightly the OP, but its increment enhances the perfect-identification capabilities of the technique for all random-error distributions. As it is expected, the worst identification performance of the strategy occurs when Laplacian random errors are used to generate the set of measurements.

Regarding the bias-estimation capabilities, it can be seen that the performance enhances, in general, with the number of samples until some threshold value is reached. No significant improvement is achieved above it. For this example, the best behavior is obtained for a sample size of 350. Also, measurements generated using Laplacian random errors provide the worst results.

Table 3. Performance Measures for Example 2

Distribution	Max #B	I	β	AVTI	OP	OPF	Mean (ε_r) (%)
Normal	1	150	11.4	0.0637	1	0.9379	0.53–4.13
	1	250	8.0	0.0376	1	0.9631	0.12–3.61
	1	350	6.8	0.0300	1	0.9705	0.32–2.71
	1	500	5.9	0.0280	1	0.9726	0.06–2.98
	3	150	11.4	0.0401	0.9999	0.9600	0.72–3.88
	3	250	8.0	0.0172	1	0.9827	1.05–2.92
	3	350	6.8	0.0142	0.9998	0.9855	0.62–3.18
	3	500	5.9	0.0149	1	0.9854	0.94–2.67
	5	150	11.4	0.0263	0.9760	0.9009	1.29–3.89
	5	250	8.0	0.0138	0.9946	0.9699	1.27–4.40
	5	350	6.8	0.0109	0.9979	0.9825	0.59–2.50
	5	500	5.9	0.0075	0.9992	0.9901	0.93–2.78
	7	150	11.4	0.0209	0.9532	0.8175	1.34–4.07
	7	250	8.0	0.0086	0.9831	0.9294	0.94–3.57
	7	350	6.8	0.0066	0.9933	0.9707	0.83–2.46
	7	500	5.9	0.0038	0.9961	0.9814	0.95–2.29
Uniform	1	150	10.4	0.0559	1	0.9454	0.45–4.22
	1	250	7.7	0.0414	1	0.9591	0.36–4.80
	1	350	6.4	0.0275	1	0.9728	0.17–3.12
	1	500	5.5	0.0258	1	0.9744	0.51–5.98
	3	150	10.4	0.0375	0.9998	0.9625	0.33–3.45
	3	250	7.7	0.0184	0.9994	0.9804	0.68–2.80
	3	350	6.4	0.0129	0.9999	0.9872	0.51–3.14
	3	500	5.5	0.0065	0.9999	0.9933	0.78–2.78
	5	150	10.4	0.0263	0.9847	0.9297	1.03–3.13
	5	250	7.7	0.0157	0.9959	0.9725	0.78–3.14
	5	350	6.4	0.0094	0.9988	0.9871	0.75–2.68
	5	500	5.5	0.0040	0.9974	0.9876	0.93–2.73
	7	150	10.4	0.0187	0.9592	0.8371	0.91–4.48
	7	250	7.7	0.0075	0.9807	0.9223	0.94–3.70
	7	350	6.4	0.0065	0.9941	0.9710	1.23–2.50
	7	500	5.5	0.0022	0.9924	0.9641	1.03–2.56
Laplacian	1	150	12.1	0.0867	0.9987	0.9162	0.69–5.74
	1	250	9.0	0.0652	0.9982	0.9352	0.48–6.52
	1	350	7.6	0.0659	0.9995	0.9364	0.80–4.33
	1	500	6.5	0.0716	0.9997	0.9308	0.02–5.24
	3	150	12.1	0.0614	0.9930	0.9269	0.49–4.52
	3	250	9.0	0.0424	0.9973	0.9534	0.97–3.28
	3	350	7.6	0.0365	0.9972	0.9594	0.68–3.81
	3	500	6.5	0.0408	0.9978	0.9561	1.11–4.20
	5	150	12.1	0.0529	0.9609	0.8476	1.61–4.74
	5	250	9.0	0.0321	0.9773	0.9039	1.02–5.37
	5	350	7.6	0.0346	0.9815	0.9129	0.94–4.07
	5	500	6.5	0.0238	0.9853	0.9287	1.41–4.33
	7	150	12.1	0.0412	0.9351	0.7413	1.96–5.66
	7	250	9.0	0.0264	0.9558	0.8372	1.72–5.95
	7	350	7.6	0.0247	0.9707	0.8736	1.86–3.84
	7	500	6.5	0.0150	0.9721	0.8673	1.21–4.62

Conclusions

In this work, a MSPC strategy devoted to the identification and estimation of biases for processes operating under steady-state conditions is presented. The *D* statistic is applied to detect the presence of biases in a set of measurements. Besides, a new formulation of the variable contributions to the inflated statistic allows identifying the faulty sensors. Also, an approximate estimation of bias magnitudes is proposed.

The strategy is a data-driven method. The same procedure can be applied to any steady-state process without knowing the functional relationships among the variables. The linearization errors that arise when techniques developed for linear systems are applied to nonlinear systems are avoided.

Since the process model is not required, the methodology succeeds in identifying biases for the measurements of any kind. Therefore, it overcomes the limitation of other widespread strategies, whose use is restricted to measurements classified as redundant. Furthermore, there is not uncertainty related to the position of biases in the sense explained by the Equivalency Theory of Gross Errors.

In contrast to other strategies, results are independent of the probability distribution assumed for measurement errors. No assumptions about this distribution are required.

A simulation procedure is carried out to analyze the performance of the new methodology, and the influence of its parameters is discussed. Results show that it works satisfactorily for the examined benchmarks. It should be highlighted that the strategy fulfills three paramount issues related with its practical implementation in commercial software: robustness, uncertainty, and efficiency.

Acknowledgments

The authors thank the financial support of CONICET (National Research Council of Argentina), ANPCyT (National Agency for the Science and Technological Promotion), and UNS (Universidad Nacional del Sur, Bahía Blanca, Argentina).

Literature Cited

1. Narasimhan S, Jordache C. *Data Reconciliation and Gross Error Detection*. Houston: Gulf Publishing Company, 2000.
2. Romagnoli J, Sánchez M. *Data Processing and Reconciliation for Chemical Process Operations*. San Diego: Academic Press, 2000.
3. Bagajewicz M, Jiang Q. Gross error modeling and detection in plant linear dynamic reconciliation. *Comp. Chem. Eng.* 1998;22:1789–1810.
4. Albuquerque JS, Biegler LT. Data reconciliation and gross error detection for dynamic systems. *AIChE J.* 1996;42:2841–2856.
5. Arora N, Biegler LT. Redescending estimators for data reconciliation and parameter estimation. *Comp. Chem. Eng.* 2001;25:1585–1599.
6. Wang D, Romagnoli JA. A framework for robust data reconciliation based on a generalized objective function. *Ind. Eng. Chem. Res.* 2003;42:3075–3084.
7. Jackson JE. *A User's Guide to Principal Components*. New York: Wiley, 1991.
8. Jeffery TC, Young JC. Monitoring a chlorine production unit by multivariate statistical procedures. In: Jeffery T, Ota K, Fenton J, Kawamoto H, editors. *Proceedings of the Symposia of Chlor-Alkali and Chlorate Production and New Mathematical and Computational Methods in Electrochemical Engineering*. Honolulu: Electrochemical Engineering Society, 1993:136–147.
9. Miller P, Swanson R, Heckler C. Contribution plots: a missing link in multivariate quality control. *Appl. Math. Comp. Sci.* 1998;8:775–792.
10. MacGregor JF, Jaeckle C, Kiparissides C, Koutoudi M. Monitoring and diagnosis of process operating performance by multi/block PLS methods with an application to a low density polyethylene production. *AIChE J.* 1994;40:826–838.
11. Pullen GK, Allen NS, Navaratnam S, Edge M, Weddell I, Swart R, Catalina F, de Oliveira SMB, Imbernon RAL, de M. Partiti CS, Rechenberg HR, Nomikos P. Detection and diagnosis of abnormal batch operations based on multi-way principal component analysis. *ISA Trans.* 1996;35:259–266.
12. Westerhuis JA, Gurden SP, Smilde AK. Generalized contribution plots in multivariate statistical process monitoring. *Chemometr. Intell. Lab. Syst.* 2000;51:95–114.
13. Mason RL, Tracy ND, Young JC. Decomposition of T^2 for multivariate control charts interpretation. *J. Qual. Technol.* 1995;27:99–108.
14. Mason RL, Tracy ND, Young JC. A practical approach for interpreting multivariate T^2 control chart signals. *J. Qual. Technol.* 1997;29:396–406.
15. Alvarez R, Brandolin A, Sánchez M. On the variable contributions to the *D*-statistic. *Chemometr. Intell. Lab. Syst.* 2007;88:89–196.
16. Chou Y, Mason R, Young J. The control chart for individual observations from a multivariate non-normal distribution. *Commun. Stat. Theory Meth.* 2001;30:1937–1949.
17. Swartz CLE. Data reconciliation for generalized flowsheet applications. Presented at the National Meeting of the American Chemical Society, Dallas, 1989.
18. Serth R, Heenan W. Gross error detection and data reconciliation in steam-metering systems. *AIChE J.* 1986;32:733–741.

Manuscript received Nov. 5, 2007, and revision received Apr. 10, 2008.