ELSEVIER

Chemical Engineering Journal 84 (2001) 309–314

Chemical Engineering Journal

www.elsevier.com/locate/cej

Reconciliation of censored measurements in chemical processes: an alternative approach

V.G. Dovì*, C. Solisio

DICheP, "G.B. Bonino" Genoa University, Via Opera Pia 15, Genova 16145, Italy

Received 28 April 1998; accepted 13 November 2000

Abstract

The importance of considering the censoring of measured data in the reconciliation of process flow rates has been shown in a previous paper [Chem. Eng. Sci. 52 (17) (1997) 3047]. The purpose of the present paper is to introduce a new technique for carrying out the actual reconciliation procedure and compare its significance and performance with those of previous methods. A numerical example shows how nontrivial differences are to be expected. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Data reconciliation; Detection limits; Censored data

1. Introduction

The measurement of concentrations and flowrates close to, or even below, the detection limits of many industrial on-line samplers is becoming more and more frequent. This is due to more and more stringent constraints on the emission of pollutants in industrial effluents and to the requirement of high purity products, which results in both toxic substances and impurities being measured at extremely low concentrations.

Traditional techniques, such as those developed by Vaclavek [2], Vaclavek and Loucka [3], Mah et al. [4], Romagnoli and Stephanopoulos [5], Crowe et al. [6], Crowe [7] and recently by Sanchez and Romagnoli [8] are not suitable for the reconciliation of process measurements when some data may be below the detection limits, because they do not consider the presence of constraints.

On the other hand, the introduction of positivity bounds on the rectified data, as proposed by Narasimhan and Harikumar [9], would overlook the functional form of the distribution function of the measurements close to the detection limits.

Similarly the maximum entropy approach recently proposed by Crowe [10] would neglect the statistical information available.

Measurements subject to detection limits are said to be censored. The following error distribution function has been previously proposed for them [1]:

$$\mathbf{p}(\varepsilon) = \mathbf{p}(\xi' - \hat{\xi}) = \begin{cases} \text{if } \xi' = 0 & \begin{cases} 1/\mathbf{T}_\xi & 0 \le \varepsilon \le \mathbf{T}_\xi \\ 0 & \varepsilon \ge \mathbf{T}_\xi \end{cases} \\ \text{if } \xi' \ge \mathbf{T}_\xi & \mathbf{N}(0, \sigma_\xi^2) \end{cases} \tag{1}$$

where ξ' is the experimental observation, $\hat{\xi}$ the unknown exact value, $\mathbf{T}_\xi$ the detection limit and σ_ξ^2 the variance of the error distribution when the measurement is above $\mathbf{T}_\xi$. In other words, a uniform distribution between zero and $\mathbf{T}_\xi$ was assumed if a zero concentration value had been observed and a normal distribution otherwise.

In this paper we propose an alternative distribution function based on a slightly modified assumption, i.e.

$$\mathbf{p}(\varepsilon) = \mathbf{p}(\xi' - \hat{\xi}) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma_\xi \mathbf{T}_\xi} \displaystyle\int_0^{\mathbf{T}_\xi} \mathrm{e}^{(\eta-\hat{\xi})^2/2\sigma_\xi^2}\, \mathrm{d}\eta & \text{if } \xi' = 0 \\ \mathbf{N}(0, \sigma_\xi^2) & \text{if } \xi' \ge \mathbf{T}_\xi \end{cases} \tag{2}$$

Apparently, the error distribution function assumed in (2) is physically more correct than that in (1), because it is the measurement ξ' that is subject to censoredness, not the unknown true value $\hat{\xi}$, which is not necessarily below the threshold $\mathbf{T}_\xi$ if $\xi' = 0$, as implied by the error distribution function (1).

On the other hand, assuming a Gaussian distribution for the experimental error between the censored measurement and the unknown exact value (and integrating over all the

[*] Corresponding author. Tel./fax: +39-010-3532921.
E-mail address: dovi@istic.unige.it (V.G. Dovì).

1385-8947/01/$ – see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S1385-8947(00)00380-6

Nomenclature

A	incidence matrix
C	concentrations (Kmol/m^3)
p	probability distribution
Q	flowrates (m^3/s)
T	detection limit (Kmol/m^3)
u	unmeasured variable
Y	transpose of the matrix that spans the null space matrix A_3
x	measured variable

Greek symbols

α	elements of the incidence matrix
ε	experimental error
η	dummy variable
ξ	censored variables
σ	standard deviation

Superscripts

$'$	experimental value
$\wedge$	exact value
$+$	pseudoinverse operator
$\wedge *$	estimated value

possible values of the censored measurement) requires a homoscedastic (i.e. constant variance) distribution even for values of ξ' (were they detectable) very close to zero. If the latter assumption is not reasonable, then the distribution (1) should be preferred in spite of the drawbacks discussed above.

The aim of this article is to compare the procedures and the results obtained using the distribution (2) rather than distribution (1) or neglecting data censoredness altogether. From case to case the reconciliation problem will have to be solved using the most convenient assumption. To keep the notation simple, we shall consider, as in the previous article, only the basic case of linear constraints. Linear constraints are based on the assumption that total flow rates are measured in any stream in which the concentration of the component subject to detection limits is measured. The product of the total flow rate and the concentration is then used as the measurement to be adjusted.

In fact, if two random variables are distributed normally, their product is also approximately Gaussian. First we shall demonstrate this in a simple case, following the proof provided by Madron et al. [11]. To this purpose consider the stochastic variable $x = Q \cdot c$, where Q and c are measurements subject to errors, whose probability distributions are normal. The unknown exact values are given by $\hat{x} = \hat{Q} \cdot \hat{c}$, where

$$\hat{x} = x + \varepsilon_x, \quad \hat{Q} = Q + \varepsilon_V, \quad \hat{c} = c + \varepsilon_c$$

It follows that

$$x + \varepsilon_x = (Q + \varepsilon_V) \cdot (c + \varepsilon_c)$$

and consequently

$$\varepsilon_x = c\varepsilon_V + Q\varepsilon_c + \varepsilon_c\varepsilon_V$$

If $c\varepsilon_V + Q\varepsilon_c \gg \varepsilon_c\varepsilon_V$, then ε_x is a linear combination of independent random variables with normal distributions and consequently it also has a Gaussian distribution.

To determine the variance–covariance matrix in the general case,

$$x_k = g_k(\mathbf{t})$$

where x_k is an arbitrary function of the measured variables $\mathbf{t}$, we can use the approximate relation suggested by Box [12]:

$$\varepsilon_{x,k} = \sum \frac{\partial g_k}{\partial t_j} \varepsilon_{t,j}$$

and consequently$V_x = B\, V_t\, B^{\mathrm{T}}$, where B is the matrix whose element b_{kj} is given by $b_{kj} = \partial b_k / \partial t_j$.

For clarity we have not taken chemical reactions into consideration, so that the reconciliation of measured flow rates can be carried out separately for each component. However, this does not impair the generality of the method. In fact, if present, stoichiometric variables can be eliminated using the matrix projection technique used in the next paragraph to eliminate unmeasured variables, as described in [6]. Similarly, the presence of nonlinear constraints can be dealt with using the method developed by Dovì and Del Borghi [13].

In particular, we shall focus in the sequel on a single component subject to detection limits. Again, no loss of generality is implied by this choice. In fact, if all the components in a stream are to be balanced, the adjusted total flow rate can be made to equal the sum of the adjusted component flow rates using the procedure described by Crowe et al. [6].

2. Problem statement

Following the notation used in [1], let us describe the process structure by an incidence matrix A whose rows correspond to units (or nodes) and columns to streams. The elements of A are defined as follows:

$$\alpha_{kj} = \begin{cases} 1 & \text{if stream } j \text{ is an input to unit } k \\ -1 & \text{if stream } j \text{ is an output from unit } k \\ 0 & \text{otherwise} \end{cases}$$

Since we have limited our analysis to one nonreacting component, the material balance for the plant can be written as

$$A\mathbf{y} = 0 \tag{3}$$

where $\mathbf{y}$ is the vector of the flow rates of the component considered.

The vector $\mathbf{y}$ is partitioned into the subsets $\mathbf{x}$, ξ and $\mathbf{u}$, $\{\mathbf{y}\} = \{\mathbf{x}|\xi|\mathbf{u}\}$, where $\mathbf{x}$ are measured streams with values above the detection limit, ξ the censored stream data and $\mathbf{u}$ the unmeasured streams.

Let us rewrite Eq. (1) as

$$A_1\mathbf{x} + A_2\xi + A_3\mathbf{u} = 0$$

If the set $\{\mathbf{x}\}$ is assumed uncorrelated to the set $\{\xi\}$, the maximum likelihood assumption gives rise to the following problem:

$$\Phi_1(\hat{\mathbf{x}}, \hat{\xi}) = \max \left\lfloor \mathrm{e}^{-(\mathbf{x}'-\hat{\mathbf{x}})^{\mathrm{T}} V_x^{-1} (\mathbf{x}'-\hat{\mathbf{x}})} \cdot \mathbf{p}(\xi) \right\rfloor$$
$$\text{s.t. } A_1\hat{\mathbf{x}} + A_2\hat{\xi} + A_3\hat{\mathbf{u}} = 0, \quad \underline{\hat{\xi}} \le \underline{\tau_{\underline{\xi}}} \tag{4}$$

where $\mathbf{p}(\hat{\xi})$ is the distribution in the variables $\hat{\xi}$ given by Eq. (2).

Thus, we can rewrite Eq. (4) as

$$\Phi_1(\hat{\mathbf{x}}, \hat{\xi}) = \max \left[\mathrm{e}^{-(\mathbf{x}'-\hat{\mathbf{x}})^{\mathrm{T}} V_x^{-1} (\mathbf{x}'-\hat{\mathbf{x}})} \cdot \prod_j \int_0^{\tau_{\xi_j}} \mathrm{e}^{-((\eta-\hat{\xi}_j)^2/2\sigma_\xi^2)} \, \mathrm{d}\eta \right]$$
$$\text{s.t. } A_1\hat{\mathbf{x}} + A_2\hat{\xi} + A_3\hat{\mathbf{u}} = 0 \tag{5}$$

Let us define now Y^{T} as the matrix that spans the null space of A_3, i.e. $Y^{\mathrm{T}} \cdot A_3 = 0$.

We can rewrite problem (5) as

$$\Phi_1(\hat{\mathbf{x}}, \hat{\xi}) = \max \left[\mathrm{e}^{-(\mathbf{x}'-\hat{\mathbf{x}})^{\mathrm{T}} V_x^{-1} (\mathbf{x}'-\hat{\mathbf{x}})} \cdot \prod_j \int_0^{\tau_{\xi_j}} \mathrm{e}^{-((\eta-\hat{\xi}_j)^2/2\sigma_\xi^2)} \, \mathrm{d}\eta \right]$$
$$\text{s.t. } Y^{\mathrm{T}} A_1\hat{\mathbf{x}} + Y^{\mathrm{T}} A_2\hat{\xi} = 0$$

This is the final formulation of our reconciliation problem.

To numerically solve this problem, we can transform it into an unconstrained optimization problem by solving the linear constraints for a suitable subset of the variables $\{\hat{\mathbf{x}} \oplus \hat{\xi}\}$ and substituting in the objective function. The unconstrained objective function can then be maximized using a BFGS quasi-Newton algorithm. If the presence of local stationary points is to be taken into consideration (this could be the case of complex process structures), the extremely efficient mapped continuation algorithm proposed by Seader et al. [14] to locate all extremum points of a function can be used.

Of course the vector $Y^{\mathrm{T}} A_2 \xi$ does not necessarily include all the ξ, because $Y^{\mathrm{T}} \cdot A_2$ might include columns of zero elements. The jth column of $Y^{\mathrm{T}} \cdot A_2$ is identically zero if ξ_j enters the mass balance equation (3) in a linear combination with components of the $\underline{\mathrm{u}}$ vector. In this case the term

$$\int_0^{\tau_{\xi_j}} \mathrm{e}^{-((\eta-\hat{\xi}_j)^2/2\sigma_\xi^2)} \, \mathrm{d}\eta$$

can be dropped from the objective function and can be estimated from

$$\varphi(\xi_j) = \int_0^{\tau_{\xi_j}} \mathrm{e}^{-((\eta-\hat{\xi}_j)^2/2\sigma_\xi^2)} \, \mathrm{d}\eta = \max$$

which provides the obvious solution

$$\xi_j = \tfrac{1}{2}\tau_{\xi_j}$$

While the procedure described in [1] required a preliminary inspection of the number of linearly independent variables of the matrix $Y^{\mathrm{T}} A_2$ for the determination of the number of variables $\hat{\xi}_{\mathbf{k}}$ which could be estimated independently, this is no longer the case for the method described in this article.

In fact the variables $\hat{\xi}_{\mathbf{k}}$ are now present not only in the constraints, but also in the objective function. This makes it possible to estimate them regardless of the number of linearly independent columns of the matrix $Y^{\mathrm{T}} A_2$.

From a mathematical point of view there are significant differences in both objective function and constraints between the method developed in this article and the one described in [1]. In fact, while the objective function is no longer a quadratic expression, this increased difficulty is offset by the absence of inequality constraints, which are replaced by equality constraints. The latter are generally easier to deal with and, being linear, can be conveniently eliminated with a simultaneous reduction of the independent variables, as discussed above. Thus, even from an algorithmic point of view, there is no clear-cut advantage in using this procedure with respect to the one previously described.

However, nontrivial differences can be detected in applications, as shown in the next section. Therefore we recommend that both procedures be used and the results, if significant differences turn up, examined carefully.

3. A numerical example

A simple process, used to illustrate the previous method and presently being implemented on a pilot plant scale, has been considered for comparison.

A homogeneous catalyst is fed to a reactor whose output undergoes two separation processes. The top product of the first separation unit is recycled to the reactor, whereas the bottom product is fed to the second separation unit. The flowsheet of the process is represented in Fig. 1, where all the streams have been labelled. Furthermore, the variance–covariance matrix is assumed diagonal. In this section the symbol x_j indicates the jth stream, regardless of whether it is measured with a normally distributed error, censored or unmeasured.

Since we are considering only the reconciliation of measurements of the catalyst flow rates, we can use the procedure developed without chemical reactions. Again, if other components were taken into consideration, the more general projection technique described by Crowe et al. [6] can be used.

Using the three constraints the optimization can be carried out with respect to x_1, x_2, x_6. Due to the particular structure of the flowsheet, it is not necessary to actually carry out

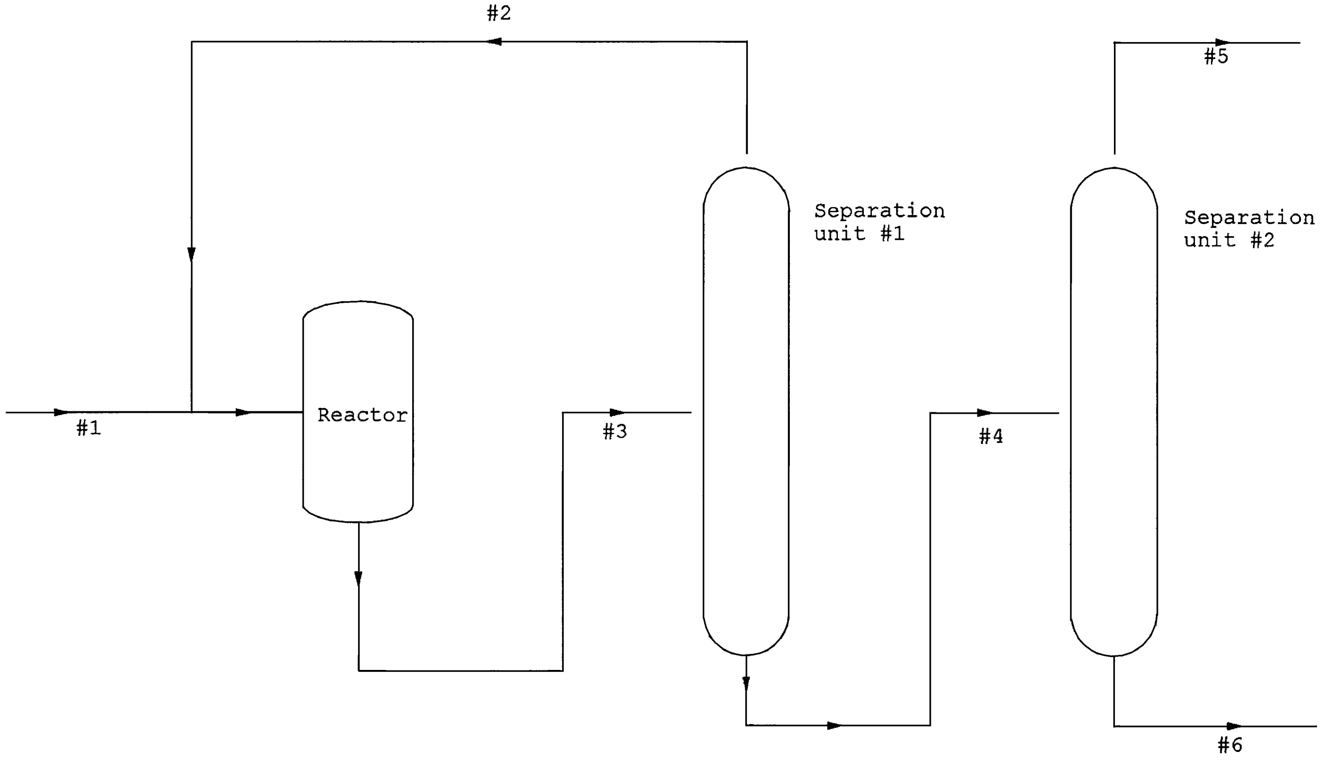

Fig. 1. Flowsheet considered.

the full optimization procedure. In fact both x_2 and x_6 can be calculated separately for every value of x_1, setting the relevant derivatives to zero, i.e.

$$\frac{\partial \ln \Phi_1}{\partial x_2} = 0 \Rightarrow \frac{(x_2' - x_2)^2}{\sigma_2^2} + \frac{[x_3' - (x_1 + x_2)]^2}{\sigma_3^2} = 0$$

$$\frac{\partial \ln \Phi_1}{\partial x_6} = 0 \Rightarrow \frac{\partial}{\partial x_6}\left(\left\{\mathrm{Erf}\left[\frac{x_6}{\sqrt{2}\sigma_6}\right] + \mathrm{Erf}\left[\frac{\tau_{\xi_6} - x_6}{\sqrt{2}\sigma_6}\right]\right\} \cdot \times \left\{\mathrm{Erf}\left[\frac{x_1 - x_6}{\sqrt{2}\sigma_5}\right] + \mathrm{Erf}\left[\frac{\tau_{\xi_6} - (x_1 - x_6)}{\sqrt{2}\sigma_5}\right]\right\}\right) = 0$$

where primed symbols indicate the experimental values.

Thus, by letting x_1 vary over a suitable range, we can locate the overall minimum without convergence difficulties.

Two different cases were considered. In both of them three algorithms have been used for comparison:

1. the censoredness of data is neglected and only the positivity constraint is considered [9];
2. the censoredness of data is taken into account in the expression of the likelihood function according to the procedure described in [1];
3. the censoredness of data is taken into account in the expression of the likelihood function according to the method developed in this work.

3.1. First case

All streams are measured. Only stream No. 6 is below its detection limit. The data are reported in Table 1, whereas the results obtained using various assumptions are reported

Table 1
Experimental data

Stream No.	Measured total flow rate (m^3/s)	Measured concentration (mol/m^3)	Component flow rate (mol/s)	Detection limit	Variance
1	11.00	0.001	0.110	–	0.0001
2	9.90	0.093	0.920	–	0.01
3	21.25	0.048	1.020	–	0.01
4	11.10	0.008	0.089	–	0.0005
5	7.00	0.001	0.070	–	0.0005
6	3.80	0	0	0.03	0.0001

Table 2
Reconciled values using different techniques (one censored measurement)

Stream No.	Rectified flow rates		
	Assumption 1	Assumption 2	Assumption 3
1	0.1020	0.1041	0.0964
2	0.9190	0.9180	0.9218
3	1.0210	1.0220	1.0182
4	0.1020	0.1041	0.0964
5	0.0967	0.0835	0.0589
6	0.0053	0.0206	0.0375

in Table 2. As can be seen, the reconstructed values of x_5 and x_6 depend considerably on the reconciliation method employed. In particular, assumption 1 gives rise to a gross error on x_5, whereas assumption 3 leads to values of x_6 above the detection limit, a possibility that was anticipated in Section 1.

3.2. Second case

Same as case 1, except for the measurement of stream No. 5, which is now censored with a detection limit equal to 0.08, the results are reported in Table 3. Unlike the results obtained using the previous method, we are now able to estimate both x_5 and x_6. However, the rectified value of x_6 corresponds to the presence of a gross error (the reconstructed value is twice the detection limit).

Table 3
Reconciled values using different techniques (two censored measurements)

Stream No.	Rectified flow rates		
	Assumption 1	Assumption 2	Assumption 3
1	0.0935	0.1	0.1061
2	0.9232	0.9200	0.91695
3	1.0168	1.0200	1.02305
4	0.0935	0.1	0.1061
5	0.0779	0.1 ($\hat{x}_5 + \hat{x}_6$)[a]	0.0471
6	0.0156	–	0.0590

[a] The individual values of $\hat{x}_5$ and $\hat{x}_6$ could not be estimated separately.

Table 4
Sets of simulated data

Stream No.	Reconciliation example (mol/s)				
	1°	2°	n°	10°	11°
1	0.1	0.1	0.1	0.1	0.1
2	0.9	0.9	0.9	0.9	0.9
3	1.0	1.0	1.0	1.0	1.0
4	0.1	0.1	0.1	0.1	0.1
5	0.07	0.073	–	0.097	0.099
6	0.03	0.027	–	0.003	0.001

4. Conclusions

We have described, as an alternative to a method developed in a previous paper, a new algorithm for the

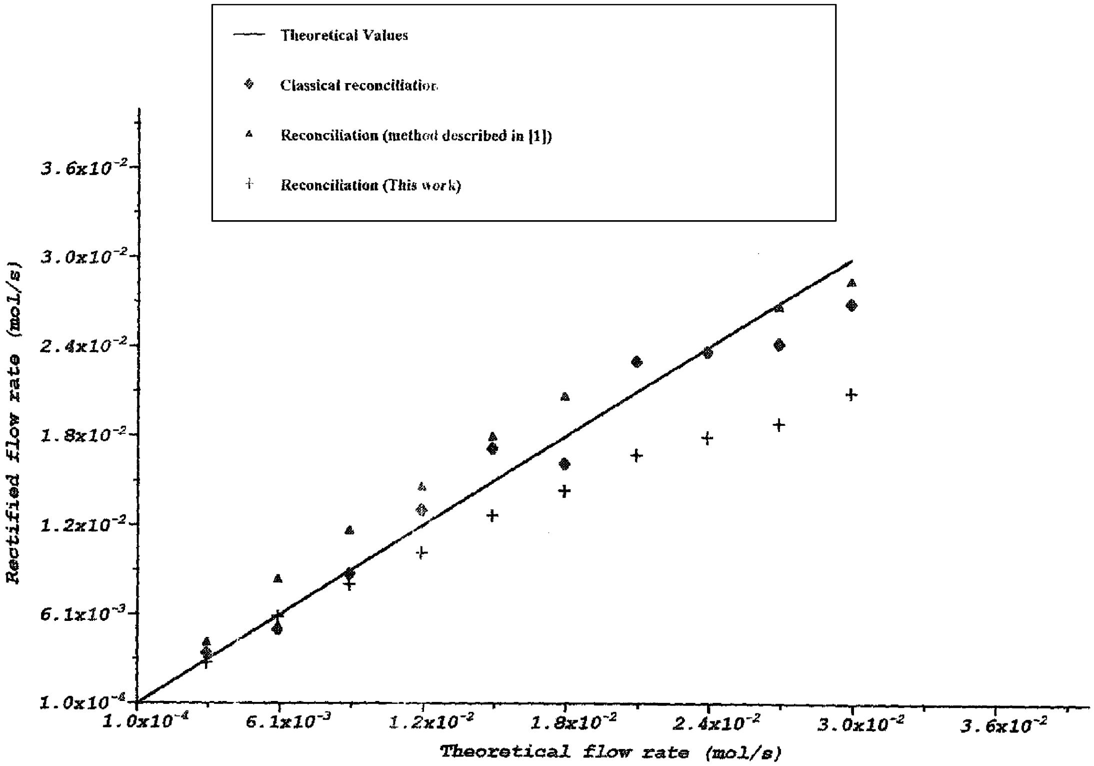

Fig. 2. Rectified flowrates using different techniques.
]

maximization of a generalized likelihood function in reconciliation problems that include censored data.

Nontrivial differences in the rectified values obtained using the two methods have been detected. From case to case the use of either method should be decided according to the general characteristics of the structure of the experiment (homo- or heteroscedasticity).

To show this influence we have carried out a test using a sequence of rectifications based on simulated data. In particular, we have considered the same flowsheet as the one in Fig. 1. The theoretical partial flowrates were kept constant for the streams 1–4, whereas stream 6 ranged between 10^{-4} and 3×10^{-2}. The range of stream 5 followed from the mass balances around unit 3. The corresponding values are reported in Table 4.

Gaussian random errors with the variances reported in Table 1 were added to the data concerning flow rates 1–4 and to flow rate 5. A heteroscedastic error with a variance equal to $10^{-6} + 10^{-1}\xi^2$ was added to Stream 6, where ξ is the actual value of the stream.

Ten cases with different random errors were considered for each set of theoretical values. The data were rectified using the same different assumptions 1–3 stated before, with the only difference that the measurement of stream 5 was considered detectable if assumption 1 was used.

The rectified values of stream 6 in the three cases are reported in Fig. 2. For every theoretical value of stream 6, the reconstructed values are the average values obtained in the 10 cases generated using different random errors as explained above.

Regarding the values rectified using procedure 1 (i.e. disregarding the detection limit) as the most reliable ones, it is evident that a better match is obtained using the model developed in this work as the value of ξ moves towards the detection limit (10^{-3}) and consequently the true variance (equal to $10^{-4} + 10^{-1} \times 9 \times 10^{-4}$) becomes $\approx 10^{-4}$, equal to the constant value used in the reconstruction method proposed in this work.

This is the expected confirmation that in case the measurement errors (were they detectable) have a heteroscedastic structure, the reconciliation method developed in [1] has to be preferred unless a detailed knowledge of the heteroscedasticity coefficients are available. Generally, this is hardly ever the case in real industrial practice.

References

[1] V.G. Dovì, A.P. Reverberi, L. Maga, Reconciliation of process measurements when data are subject to detection limits, Chem. Eng. Sci. 52 (17) (1997) 3047–3050.

[2] V. Vaclavek, Studies on system engineering. III. Optimal choice of the balance measurements in complicated chemical systems, Chem. Eng. Sci. 24 (1969) 947–955.

[3] V. Vaclavek, M. Loucka, Selection of measurements necessary to achieve multicomponent mass balances in chemical plants, Chem. Eng. Sci. 31 (1976) 1199–1205.

[4] R.S.H. Mah, G.M. Stanley, D.M. Downing, Reconciliation and rectification of process flow and inventory data, Ind. Eng. Chem. Proc. Des. Dev. 15 (1976) 175–183.

[5] J.A. Romagnoli, G. Stephanopoulos, On the rectification of measurement errors for complex chemical plants, Chem. Eng. Sci. 35 (1980) 1067–1081.

[6] C.M. Crowe, Y.A. Garcia Campos, A. Hrymak, Reconciliation of process flow rates by matrix projection. Part I. Linear case, AIChE J. 29 (1983) 881–888.

[7] C.M. Crowe, Reconciliation of process flow rates by matrix projection. Part II. The nonlinear case, AIChE J. 32 (1986) 616–623.

[8] M. Sanchez, J.A. Romagnoli, Use of orthogonal transformations in data classification–reconciliation, Comput. Chem. Eng. 20 (1996) 483–493.

[9] S. Narasimhan, P. Harikumar, A method to incorporate bounds in data reconciliation and gross error detection. I. The bounded data reconciliation problem, Comput. Chem. Eng. 17 (1993) 1115–1128.

[10] C.M. Crowe, Formulation of linear data reconciliation using information theory, Chem. Eng. Sci. 51 (1996) 3359–3366.

[11] F. Madron, V. Veverka, V. Vanecek, Statistical analysis of material balance of a chemical reactor, AIChE J. 23 (4) (1977) 482–486.

[12] M.J. Box, Improved parameter estimation, Technometrics 12 (1970) 219–232.

[13] V.G. Dovì, A. Del Borghi, Reconciliation of process flow rates when measurements are subject to detection limits. The bilinear case, I & EC 38 (7) (1999) 2861–2866.

[14] J.D. Seader, D. Kuno, W.J. Lin, S.A. Johnson, K. Unsworth, J.W. Wiskin, Mapped continuation methods for computing all solutions to general systems of nonlinear equations, Comput. Chem. Eng. 16 (16) (1990) 71–82.