

Improved Mixed Integer Optimization Approach for Data Rectification with Gross Error Candidates*

LI Jianlie (李筑列) and RONG Gang (荣冈)**

State Key Laboratory of Industrial Control Technology, Institute of Cyber System and Control, Zhejiang University, Hangzhou 310027, China

Abstract Mixed integer linear programming (MILP) approach for simultaneous gross error detection and data reconciliation has been proved as an efficient way to adjust process data with material, energy, and other balance constrains. But the efficiency will decrease significantly when this method is applied in a large-scale problem because there are too many binary variables involved. In this article, an improved method is proposed in order to generate gross error candidates with reliability factors before data rectification. Candidates are used in the MILP objective function to improve the efficiency and accuracy by reducing the number of binary variables and giving accurate weights for suspected gross errors candidates. Performance of this improved method is compared and discussed by applying the algorithm in a widely used industrial example.

Keywords data rectification, gross error detection, graphic theory, Bayesian method

1 INTRODUCTION

In most chemical plants, measurement error will lead to poor decisions in the regular plant activities such as planning and process optimization. After the idea of data reconciliation was brought in 1961 [1], the problem of improving the accuracy of process data to let them satisfy the material, component, and energy balance with less adjustment and shorter time has been studied for decades. Besides the widely used solution of linear and nonlinear problem using matrix projection [2, 3], many other methods such as mixed integer linear programming (MILP) method [4], principal component analysis (PCA) method [5], measurement test and node test (MTNT) method [6–8], and redundancy analysis method [9] have been developed to solve the steady-state data rectification problem. More information about the history of data rectification methods can be obtained from a perspective written by Crowe [10].

MILP framework of simultaneous data reconciliation and gross error detection is prompted to remove random errors of plant data, as well as identify and compensate gross errors for the final solution. MILP method defines an associated binary variable for each measurement to indicate the existence of gross error and add penalty in the objective function to activate the binary variables. However, the efficiency will decrease significantly when the method is applied to large-scale problems, as there will be too many binary variables in the calculation. Fortunately, we found that before performing the mixed integer linear optimization, the number of binary variables can be reduced on the basis of historic data. Moreover, accurate weighting factors in the objective function should also improve the results.

This article describes a method for searching gross error candidates in a directed diagram of process flowsheet and calculating their reliability factors. The gross error candidates are used to decide binary variables in the MILP objective function.

2 BASIC MILP METHOD

For a linear, time invariant, and steady state system such as a network of flow rate, balance equation can be written as:

$$\sum_{i=1}^s a_{ki} F_i = 0 \quad k=1, \dots, n \quad (1)$$

where n is the number of process nodes. s is the total number of streams. F_i is the true flow rates of stream i . a_{ki} is the correlative parameter whose value is 1, -1, or 0, depending on whether the stream i enters, leaves, or not correlate to node k .

If we consider measured value $\tilde{F}_i$ consists with true value F_i , random component ε_i with zero mean and a know variance σ_i , and a deterministic bias δ_i then

$$\tilde{F}_i = F_i + \varepsilon_i + \delta_i \quad (2)$$

The MILP method is derived from the standard formulation of data reconciliation:

$$\begin{aligned} \min \Phi &= (\hat{F} - \tilde{F})^T Q (\hat{F} - \tilde{F}) \\ \text{s.t.} \quad A \cdot \hat{F} &= 0 \end{aligned} \quad (3)$$

By substituting the Eq. (2) into Eq. (3) and defining each measurement a binary variable corresponding to the existence of a bias, the new objective function with MILP formulation can be written as:

Received 2008-05-28, accepted 2008-11-06.

* Supported by the National High Technology Research and Development Program of China (2007AA40702 and 2007AA04Z191).

** To whom correspondence should be addressed. E-mail: grong@mail.hz.zj.cn

$$\sum_{i=1}^s \frac{1}{\sigma_i} \left| \hat{F}_i - (\tilde{F}_i - \delta_i) \right| + W_i B_i$$

$$\text{s.t. } \mathbf{A} \cdot \hat{\mathbf{F}} = 0 \quad (4)$$

$$\zeta U_i B_i \leq |\delta_i| \leq U_i B_i$$

$$B_i \in \text{binary}$$

Here, the binary variable B_i represents the existence of a bias in i th measurement, and W_i is the weighting factor to change the importance of biased measurement, the second constraint ensures the upper limit of a bias and the activate zone of binary variables by choosing ζU_i as some fraction of the standard deviation of the measurement. In addition, Eq. (4) can be restated to remove the absolute operator to avoid discontinuity problem.

3 GENERATION OF GROSS ERROR CANDIDATES

Because the data reconciliation and gross error detection problems are formulated as MILP frameworks, it is easy to apply other techniques to enhance the performance. Soderstrom suggests using some standard statistical tests for bias as constraints; however, these modifications not offer much benefits over the basic method. And moreover, as an optimization problem, this technique is significantly more computationally intensive. Generate a list of gross error candidates and reduce the binary variables in the objective function is obviously an efficient way to improve the efficiency of the MILP method.

In order to generate gross error candidates with weighting factors and reduce the binary variables in the objective function, an efficient method based on graphic theory and Bayesian method is prompted. This method can generate gross error candidates with prior information and cost less computation time.

Tamhane *et al.* [11, 12] used Bayesian approach to detect gross error in which the prior probabilities of

gross error occurrence are updating in the light of accumulating data. And the probability of gross error occurrence for instrument is:

$$P_i(t) = 1 - \frac{\Gamma(l_i + m_i) \Gamma[m_i + \tau_i(t)]}{\Gamma(l_i) \Gamma[l_i + m_i + \tau_i(t)]} \quad i = 1, 2, \dots, s \quad (5)$$

where l_i is the number of previous failure for instrument, and m_i is the sum of previous lifetimes for instrument i , and $\Gamma(\cdot)$ denotes the gamma function.

Once the prior probabilities are confirmed, the next step is to use these probabilities to generate gross error candidates before using the MILP method. This step can be modified or even removed when there are expert experiences or no accumulated process data. Experts can determine the prior probabilities by checking the quality of process data or starting with same prior probabilities and allowing these values to be updated in the next iteration.

In the view of graphic theory, the flowsheet can be regard as a directed graphic. As illustrated in Fig. 1, spanning tree is the subgraph of a graphic, and this subgraph does not contain any loops. So the streams in the flowsheet can be divided into two sets: branches and chords of the spanning tree. Obviously, the true value of every branch can be obtained through true values of some chords. For example, the branches of spanning tree No.1 in Fig. 1 is stream 2, 3, and 4, so the chords will be stream 1 and 5. So the true value of stream 2 and stream 4 is as same as the true value of stream 1, and true value of stream 3 is identical to the sum of true values of stream 1 and stream 5. In this case, we called these chords as independent variables (IV) and branches as dependent variables (DV).

Usually, there are a large number of spanning trees in a given graph, so the choice of independent variables is not exclusive. In this method, because the independent variables are used as benchmarks in the gross error candidate generation, we choose the chords of graph's maximal spanning tree as independent variables. And the weight of the given flowsheet is defined as prior probabilities calculated by Eq. (5).

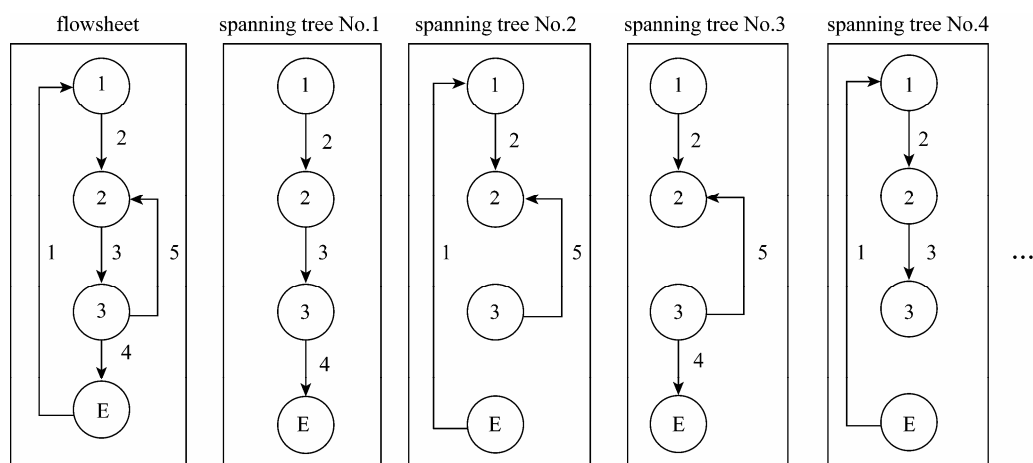


Figure 1 Spanning trees for a simple flowsheet

Because the maximal spanning tree is the spanning tree with the largest weight, the sum of the weights of all chords is minimal, which means these independent variables have lowest probabilities to have gross error. These independent variables will be used as benchmarks in the gross error generation. If a measured value does not contain any gross error, $\tilde{F}_i$ will only consist with true value F_i and random component ε_i with zero mean and a known variance σ_i^2 . So the sum of measured values is

$$\sum_{i=1}^s \tilde{F}_i = \sum_{i=1}^s F_i + \varepsilon$$

$$\varepsilon \sim N\left(0, \sum_{i=1}^s \sigma_i^2\right) \quad (6)$$

Because there are random component ε_i in the $\tilde{F}_i$, only the measured value of dependent variables can be estimated from measured value of independent variables. If we still choose chords of spanning tree No.1 in Fig. 1 as independent variables, the estimated value of stream 2 and stream 4 is the measured value of stream 1 and the estimated value of stream 3 is the sum of measured values of stream 1 and stream 5. So the relationship between estimated value $\bar{F}_{DV}$ and measured value $\tilde{F}_{DV}$ can be calculated as:

$$\bar{F}_{DV} - \tilde{F}_{DV} = \sum_{IV \in R} F_{IV} - F_{DV} + \sum_{IV \in R} \varepsilon_{IV} - \varepsilon_{DV}$$

$$= \sum_{IV \in R} \varepsilon_{IV} - \varepsilon_{DV} \quad (7)$$

where R is the set of independent variables used to estimate the dependent variable. And $\varepsilon_i \sim N(0, \sigma_i^2)$ leads to

$$Z = \sum_{IV \in R} \varepsilon_{IV} - \varepsilon_{DV} \sim N\left(0, \sum_{IV \in R} \sigma_{IV}^2 + \sigma_{DV}^2\right).$$

Statistic Z is established to indicate the measurement bias for dependent variables that has an upper threshold limit of Z_α for a level of significance α

$$|Z| \leq Z_\alpha. \quad (8)$$

There is a remaining problem that independent variables have certain probabilities to include gross error and cause the type I error of gross error candidates in the dependent variables. In order to solve this problem, if an independent variable contains gross error, we assume that all the dependent variables corresponding to this independent variable will be suspected. The probability of gross error occurrence between this independent variable and all the corresponding dependent variables will be compared to decide the gross error candidate. Assume that the probability of gross error occurrence is 0.2 for streams 2, 3, 4, and 0.1 for stream 1 in Fig. 1. If streams 2, 3, and 4, which are corresponding to stream 1, are suspected with gross errors, stream 1 will be added into gross error candidate list because the probability of all streams 2, 3, and 4 that contain gross errors will be 0.23, which is smaller than the probability of gross

error occurrence for stream 1. Once there are candidates in the independent variables, all dependent variables not in the candidate list will be removed from graph and new iteration will be performed in the subgraph till there are no new gross error candidates in the subgraph. Details of the gross error candidate generation method are described as follows:

Step 1 Weight all measured streams with their probability of gross error occurrence calculated by Eq. (5).

Step 2 Finding the maximal spanning tree of the graph (subgraph), classify the streams into independent and dependent variables.

Step 3 Establish statistic Z for each dependent variable according to Eq. (7), put the dependent variable into candidate set if $|Z| \geq Z_\alpha$.

Step 4 Check whether there are any independent variables that all the corresponding dependent variables are in candidate set. If there is such independent variable IV and $P_{IV} \geq \prod_{DV \in R} P_{DV}$, raise the P_{IV}

to 1 and put this independent variable into candidate set C then go to Step 5, else stop and output candidate set C .

Step 5 Merge all nodes without any inner gross error candidates to generate subgraph then go to Step 2.

4 APPLY CANDIDATES IN THE MILP METHOD

Gross error candidates generated in the last section will be introduced into Eq. (4) to improve its calculation efficiency. This step is accomplished by removing the binary variables in Eq. (4) that represent the streams not in the candidates. And the problem can be represented as:

$$\sum_{i \in C} \frac{1}{\sigma_i} \left| \hat{F}_i - (\tilde{F}_i - \delta_i) \right| + W_i B_i + \sum_{j \notin C} \frac{1}{\sigma_j} \left| \hat{F}_j - \tilde{F}_j \right|$$

$$\text{s.t. } \mathbf{A} \cdot \hat{\mathbf{F}} = 0 \quad (9)$$

$$\zeta U_i B_i \leq |\delta_i| \leq U_i B_i$$

$$B_i \in \text{binary}$$

When there are equivalent sets in the detected biases, the updated weighting factors can be used further to output suspect biases with higher probabilities. This strategy can be regarded as the utilization of both spatial and temporal redundancy of the measurement network. In order to understand this method easily, a flow chart that represents the improved method is shown in Fig. 2.

5 SIMULATION STUDY

The system chosen for study was an industrial steam metering process [13]. And this system was also used for comparison in many other gross error detection methods [14]. The system used for simulation is

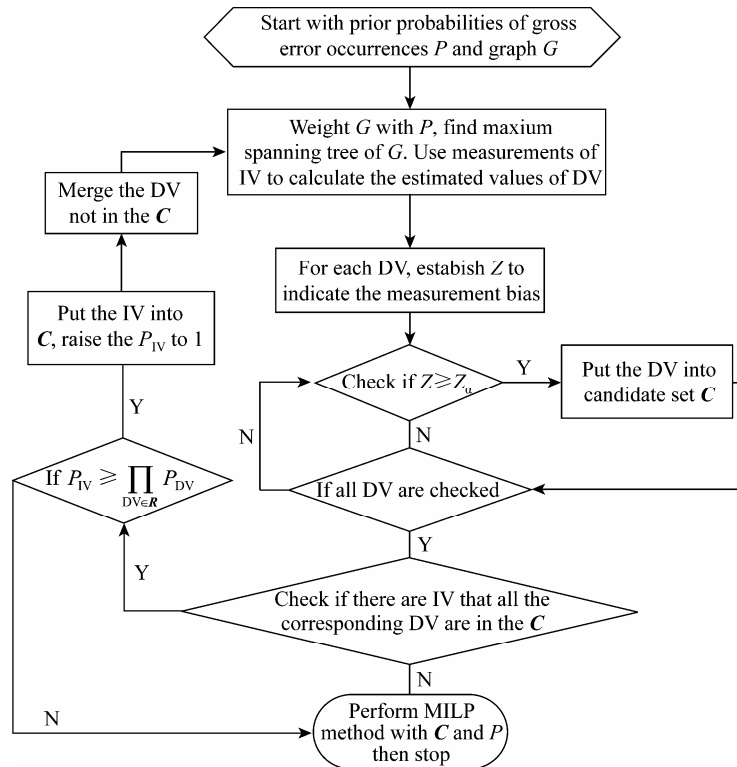


Figure 2 Flowchart of the improved method

illustrated in Fig. 3, and the true flow rates can be found in Table 1.

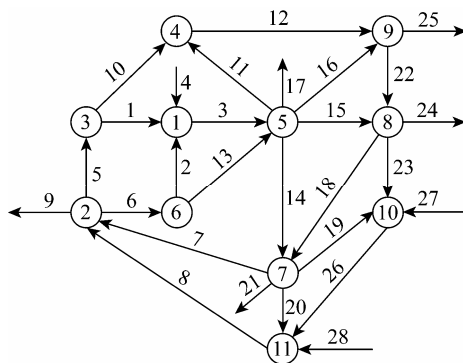


Figure 3 System of steam metering process

Monte Carlo simulations are conducted in this section to evaluate the performance of the proposed method. In order to compare the simulation done by Soderstrom [4], the number of biased variables was fixed at 3, 5, or 7, and the sign of the bias was randomly assigned with equal probability. Because there are prior probabilities of gross error occurrence for each variable, the location of the biased variables was chosen according to the prior probabilities. As same as the simulation study made by Soderstrom [4], the magnitude of the bias was chosen to be between 12.5% and 62.5% of true value of the measured variable in Table 1 and the standard deviation of the

Table 1 True values for the flow rates in steam metering process

Number	True flow rate	Number	True flow rate
1	0.86	15	1.5
2	1	16	0.591
3	111.82	17	0.81825
4	109.95	18	0.40575
5	53.27	19	0.19875
6	112.27	20	0.2625
7	2.32	21	2.1818
8	164.05	22	0.13625
9	0.86	23	0.06475
10	52.41	24	1.166
11	14.86	25	2.1363
12	67.27	26	2.033
13	111.27	27	1.7693
14	91.86	28	1.8058

measurements are chosen to be $\sigma_i = 0.025F_i$ to enable a comparison to the simulation of original MILP method. All simulations in the article were performed using MATLAB and LINGO 9 optimization software.

Average number of type I error (AVTI) and overall power (OP) [15] were applied to evaluate the performance of the improved method. The criteria are

defined as follows:

$$OP = \frac{\text{number of gross errors correctly identified}}{\text{number of gross errors simulated}} \quad (10)$$

$$AVTI = \frac{\text{number of gross errors wrongly identified}}{\text{number of simulation trials made}} \quad (11)$$

Both OP and AVTI are calculated after 100 simulation runs in different conditions.

Several interesting comparison can be made by checking the simulation results presented below. One of these is the influence of choosing more reliable independent variables. In the simulation trials, the number of higher weight bias is set as 0,1/3,2/3,1 times of number of the random selected biases. Higher the weight (prior probability of gross error occurrence) the stream has, less probability that this stream will be chosen as independent variable. Nearly all the OP is increased together with the number of higher weight bias, whereas the AVTI remains at the same level. When all biased streams are weighted with higher value, the OP increases significantly. But it is rare that all the biased streams have higher weights; however, the result seems acceptable when few biased streams are given higher weights. Fig. 4 and Fig. 5 show that good results can be obtained by fewer high weight biased streams.

Table 2 Results for the improved method ($\alpha = 0.05$)

No. (high weight biases)	Percentage (approximately)	No. (biased stream)	OP	AVTI
0	0	3	0.701	0.211
1	33.3%		0.763	0.271
2	66.7%		0.803	0.235
3	100%		0.905	0.255
0	0	5	0.626	0.532
1	33.3%		0.721	0.633
3	66.7%		0.750	0.668
5	100%		0.840	0.771
0	0	7	0.613	0.947
2	33.3%		0.678	0.961
4	66.7%		0.726	1.030
7	100%		0.780	1.160

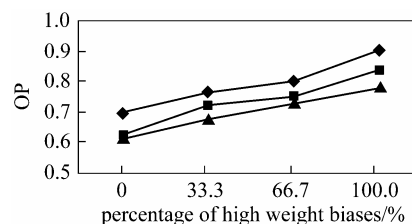


Figure 4 Influence of high weight biases percentage on overall power

◆ 3 biases; ■ 5 biases; ▲ 7 biases

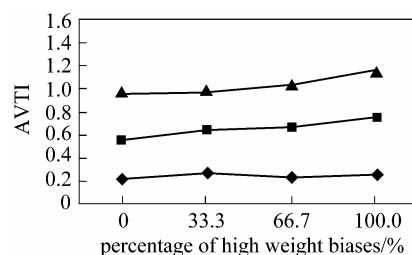


Figure 5 Influence of high weight biases percentage on average type I error

◆ 3 biases; ■ 5 biases; ▲ 7 biases

Table 3 indicates the simulation results for this method and other widely used gross error detection methods. Under most situations, performance measures of improved method (IMILP) are higher than modified iterative measurement test (MIMT) [13], modified simultaneous estimation of gross error (MSEGE), and serial identification with collective compensation (SICC) [17]. The result clearly shows that when historical data are used to evaluate the prior probability of gross error occurrence, the identification of gross errors is more accurate than just using statistical methods. However, when there are no historical data can be used to calculate prior gross error occurrence, the performance of improved method are also better than the original MILP method because the binary variables are significantly reduced.

The number of the dependent variables is equal to the number of branches in the flowsheet, so the relationship between number of dependent variable (N_{DV}); the number of independent variables (N_{IV}); and the total number of the measurements (s) are:

$$N_{DV} = n - 1 \quad (12)$$

$$N_{IV} = s - n + 1. \quad (13)$$

When the independent variable has bigger σ , the

Table 3 Performance comparison ($\alpha = 0.05$)

No. (biased stream)	OP				AVTI			
	IMILP	MT	MSEGE	SICC	IMILP	MT	MSEGE	SICC
1	0.999	0.967	0.973	0.999	0.038	0.103	0.045	0.0712
3	0.853	0.732	0.754	0.698	0.259	0.973	1.012	1.534
5	0.773	0.532	0.636	0.593	0.742	1.304	1.532	1.892
7	0.704	0.554	0.578	0.518	0.801	1.458	1.794	2.057

bias in the streams with small true value might be ignored in the candidate generation. This problem may be improved by adjusting the weights of streams with small σ (increase a little to raise the probability of being independent variable). Once the equivalent set exists because the candidate streams form a loop in the graph representing the flowsheet [16], a solution can be used by updating weights to give gross error candidates with larger probabilities.

Simulation time of different conditions is presented in Table 4. The results show that the solution time of gross error candidate generation (Average CPU time 1) is rather limited.

Table 4 Average CPU time for 3 biased streams test

Percentage of high weight bias	Average CPU time 1/ s	Average CPU time 2/ s
0%	0.183	1.033
33.3%	0.170	0.982
66.7%	0.162	1.114
100%	0.142	1.042

The solution time (Average CPU time 2) is increased due to the increase in the number of biased streams. And the trend is opposite when the percentage of high weight biases is increased.

6 CONCLUSIONS

This article proposes an improved MILP method for simultaneous gross error detection and data reconciliation. The data reconciliation and gross error detection based on the framework of MILP has a great advantage of process system integration in process industry. Because the planning/scheduling are normally solved under the framework of MILP or mixed integer nonlinear programs (MINLP), it will be much more convenient to integrate data rectification into the refinery manufacturing execution system than other methods. Effective gross error candidate generation method is applied before using the original mixed integer linear programming. Performance and comparison show that choosing reliable independent variables can improve the accuracy of measurements within a short period of time. Also, this method was found to have better performance over the original MILP method and other commonly used methods. Reduction of binary variables in the objective function leads to a superior improvement of the calculation efficiency. This improvement makes the MILP method more suitable for the application in large-scale process industries.

NOMENCLATURE

A constraint matrix

a_{ki} correlative parameter
 B binary variable
 C set of gross error candidates
 F true flow rate
 $\tilde{F}$ measured value
 l number of previous failure
 m sum of previous lifetimes
 N_{DV} number of dependent variables
 N_{IV} number of independent variables
 n number of process nodes
 R set of independent variables used to estimate the dependent variable
 s number of streams
 W weighting factor
 $\Gamma(\cdot)$ gamma function
 δ_i deterministic bias
 ε_i random error
 σ_i standard deviation

REFERENCES

- 1 Kuehn, D.R., Davidson, H., "Computer control (II) Mathematics of control", *Chemical Engineering Progress*, **57** (6), 44–47 (1961).
- 2 Crowe, C.M., "Reconciliation of process flow rates by matrix projection (II) Nonlinear case", *AIChE J.*, **32** (4), 616–623 (1986).
- 3 Crowe, C.M., Campos, Y.A.G., Hrymak, A., "Reconciliation of process flow rates by matrix projection (I) Linear case", *AIChE J.*, **29** (6), 881–888 (1983).
- 4 Soderstrom, T.A., Himmelblau, D.M., Edgar, T.F., "A mixed integer optimization approach for simultaneous data reconciliation and identification of measurement bias", *Control Engineering Practice*, **9** (8), 869–876 (2001).
- 5 Tong, H., Crowe, C.M., "Detection of gross errors in data reconciliation by principal component analysis", *AIChE J.*, **41** (7), 1712–1722 (1995).
- 6 Wang, F., Jia, X.P., Zheng, S.Q., Yue, J.C., "An improved MT-NT method for gross error detection and data reconciliation", *Comp. Chem. Eng.*, **28** (11), 2189–2192 (2004).
- 7 Yang, Y., Ten, R., Jao, L., "Study of gross error detection and data reconciliation in process industries", *Comp. Chem. Eng.*, **19** (Suppl), S217–S222 (1995).
- 8 Mei, C., Su, H., Chu, J., "An NT-MT combined method for gross error detection and data reconciliation", *Chin. J. Chem. Eng.*, **14** (5), 592–596 (2006).
- 9 Zhang, P., Rong, G., Wang, Y., "A new method of redundancy analysis in data reconciliation and its application", *Comp. Chem. Eng.*, **25** (7/8), 941–949 (2001).
- 10 Crowe, C.M., "Data reconciliation—Progress and challenges", *Journal of Process Control*, **6** (2/3), 89–98 (1996).
- 11 Tamhane, A.C., Iordache, C., Mah, R.S.H., "A Bayesian approach to gross error detection in chemical process data (I) Model development", *Chemometrics and Intelligent Laboratory Systems*, **4** (1), 33–45 (1988).
- 12 Tamhane, A.C., Iordache, C., Mah, R.S.H., "A Bayesian approach to gross error detection in chemical process data (II) Simulation results", *Chemometrics and Intelligent Laboratory Systems*, **4** (2), 131–146 (1988).
- 13 Serth, R.W., Heenan, W. A., "Gross error detection and data reconciliation in steam-metering systems", *AIChE J.*, **32** (5), 733–742 (1986).
- 14 Rollins, D.K., Cheng, Y., Devanathan, S., "Intelligent selection of hypothesis tests to enhance gross error identification", *Comp. Chem. Eng.*, **20** (5), 517–530 (1996).
- 15 Narasimhan, S., Iordache, C., Data Reconciliation & Gross Error Detection, Gulf Publishing Company, Houston (2000).
- 16 Jiang, Q., Bagajewicz, M.J., "On a strategy of serial identification with collective compensation for multiple gross error estimation in linear steady-state reconciliation", *Ind. Eng. Chem. Res.*, **38** (5), 2119–2128 (1999).
- 17 Jiang, Q., Sanchez, M., Bagajewicz, M.J., "On the performance of principal component analysis in multiple gross error identification", *Ind. Eng. Chem. Res.*, **38** (5), 2005–2012 (1999).