Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers and Chemical Engineering 28 (2004) 381–402
Computers & Chemical Engineering
www.elsevier.com/locate/compchemeng

Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes

Derya B. Özyurt[1], Ralph W. Pike*

Department of Chemical Engineering, Louisiana State University, Baton Rouge, LA 70803, USA

Received 16 November 2002; received in revised form 28 July 2003; accepted 30 July 2003

Abstract

On-line optimization provides a means for maintaining a process near its optimum operating conditions by providing set points to the process's distributed control system (DCS). To achieve a plant-model matching for optimization, process measurements are necessary. However, a preprocessing of these measurements is required since they usually contain random and—less frequently—gross errors. These errors should be eliminated and the measurements should satisfy process constraints before any evaluation on the process. In this paper, the importance and effectiveness of simultaneous procedures for data reconciliation and gross error detection is established. These procedures depending on the results from robust statistics reduce the effect of the gross errors. They provide comparable results to those from methods such as modified iterative measurement test method (MIMT) without requiring an iterative procedure. In addition to deriving new robust methods, novel gross error detection criteria are described and their performance is tested. The comparative results of the introduced methods are given for five literature and more importantly, two industrial cases. Methods based on the Cauchy distribution and Hampel's redescending M-estimator give promising results for data reconciliation and gross error detection with less computation.

Keywords: Data reconciliation; Gross error detection; Efficiency; Actual plant data; Robust estimation

1. Introduction

Real time on-line optimization provides set points to the process's distributed control system (DCS) and therefore maintains the process near its optimum operating conditions. This optimization requires an accurate process model and reconciled process data. The process model is a set of inequality and equality constraints and describes the fundamental relationship of process units, such as material and energy balances, rate equations and equilibrium relations. Reconciled process data is used to specify the current status of the plant model and for estimation of the model parameters for plant-model matching.

Data reconciliation adjusts process measurements with random errors by having them satisfy material and energy balance constraints and is a way to improve the quality of the measurements taken from a process via DCS or any other means of data collection.

The elimination of the less frequent gross errors is achieved by gross error detection. Therefore, simultaneous data reconciliation and gross error detection have emerged as a key part of on-line optimization.

Since the first proposed solution to the steady-state data reconciliation problem (Kuehn & Davidson, 1961) a vast body of chemical engineering literature has been developed describing many other approaches. Besides the solution of the linear and nonlinear problem using matrix projection (Crowe, 1986; Crowe, Garcia Campos, & Hyrmak, 1983), a solution of the nonlinear data reconciliation problem via successive linearization is described (Knepper & Gorman, 1980; Veverka & Madron, 1997). Liebman and Edgar (1988) demonstrated that using nonlinear programming instead of successive linearization remarkably improved reconciliation results. Tjoa and Biegler (1991) showed that using nonlinear programming along with a method based on a contaminated Normal (Gaussian) objective function instead of the least squares objective function, any gross error present in the measurements could be replaced with reconciled values,

* Corresponding author. Tel.: +1-225-578-3428; fax: +1-225-578-1476.
E-mail address: pike@lsu.edu (R.W. Pike).
[1] Current address: Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

0098-1354/$ – see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2003.07.001

and an iterative procedure was not required. By establishing an analogy between maximum likelihood rectification (MLR) and robust regression, Johnston and Kramer (1995) reported the feasibility and better performance of the robust estimators as the objective function in the data reconciliation problem especially when the data contain gross errors. Subsequently, different types of robust estimators and their performance in data reconciliation were reported (Albuquerque & Biegler, 1996; Arora & Biegler, 2001). These studies have shown the potential of robust statistics developed by Huber (1981), which attempts accurate estimation of statistical parameters in the presence of gross errors.

However, the simultaneous approach for data reconciliation along with the gross error detection using the results from robust statistics has been employed to solve problems of small size (6–12 equality constraints) except for (Chen, Pike, & Hertwig, 1998) and (Jordache, Ternet, & Brown, 2001). Moreover, the derivation of some other robust objective functions and their comparative performance have not been studied. In this paper, these issues are addressed along with the derivation and performance evaluation of different gross error detection criteria with and without dependence on the objective function used. Most importantly, the presented methods are employed on two industrial plants; the petroleum refinery alkylation process and the contact process for sulfuric acid production.

2. General formulation

A data reconciliation problem begins with the acquisition of the process data measurements. To assess the performance of the process evaluation or control,

$$\boldsymbol{x}^{\mathrm{T}} = [x_1, x_2, x_3, \ldots, x_n]$$

is a set of system variables for which sensors are available to measure their state.

The result of a measurement session (data from the DCS) can be collected in a set of measurement vectors as follows

$$\boldsymbol{y}_i^{\mathrm{T}} = [y_{i,1}, y_{i,2}, \ldots, y_{i,l_i}] \quad \text{for } i = 1, 2, 3, \ldots, n$$

where l_i is the number of sets of measurements taken during steady-state plant operation to estimate the system variable x_i. l_i is equal to one if we are interested in the snapshot of the process and is greater than one if our concern is a smoothed average within a time window of interest.

If there were no gross errors in the system, the difference of the measured values and the system state would have a distribution around the mean zero, i.e.

$$y_{i,1} - x_i, y_{i,2} - x_i, y_{i,3} - x_i, \ldots, y_{i,l_i} - x_i$$

is a sample from a distribution with mean zero. Also, the unknown variance of this distribution can be estimated by using plant's historical data.

The states of the system variables are determined by using the constraints that describe the process. Therefore, using a proper objective function in an NLP, estimates of the x_i's can be obtained which are expected to minimize these differences.

The formulation for the data reconciliation problem with the generalized least squares method has its root in the general regression model. Let us define a single measurement of the ith measured variable at the jth steady state as $y_{i,j}$. The kth fixed regressor (explanatory or independent) variable that we believe to explain the variation between each steady state is called $z_{k,j}$. Then a linear regression problem with fixed regressors using generalized least squares estimation is posed as:

$$\min \sum_{j=1}^{J} \frac{(y_{i,j} - \theta_0 - \theta_1 z_{1,j} \cdots - \theta_k z_{k,j})^2}{\sigma_j^2} \tag{1}$$

for which the regression model is stated as:

$$y_{i,j} = \theta_0 + \theta_1 z_{1,j} \cdots + \theta_k z_{k,j} + \varepsilon_j,$$
$$E(\varepsilon_j) = 0, \qquad \mathrm{Var}(\varepsilon_j) = \sigma_j^2 \quad \forall j \tag{2}$$

An estimate for the location of the steady state can be calculated using a special case of the linear regression problem described above, where $k = 0$ and the sum is over the steady-state points l.

$$\min \sum_{l=1}^{l_i} \frac{(y_{i,l} - \theta_0)^2}{\sigma_{i,l}^2} \tag{3}$$

The corresponding regression model is

$$y_{i,l} = \theta_0 + \varepsilon_{i,l}, \qquad E(\varepsilon_{i,l}) = 0, \qquad \mathrm{Var}(\varepsilon_{i,l}) = \sigma_{i,l}^2 \quad \forall l \tag{4}$$

The minimization problem (3) can also be written as

$$\min \sum_{l=1}^{l_i} \frac{(y_{i,l} - x_{i,l})^2}{\sigma_{i,l}^2} \quad \text{such that } x_{i,1} = x_{i,2} = \cdots = x_{i,l_i} = \theta_0 \tag{5}$$

Formulation (5) is equivalent to (6)

$$\min \sum_{l=1}^{l_i} \frac{(y_{i,l} - x_{i,l})^2}{\sigma_{i,l}^2} \quad \text{such that } Ax_i = 0,$$
$$A((l_i - 1) \times l_i) = \begin{bmatrix} 1 & -1 & 0 & . & 0 \\ 0 & . & . & . & 0 \\ 0 & . & . & . & 0 \\ 0 & . & 1 & -1 & 0 \\ 0 & . & . & 1 & -1 \end{bmatrix} \tag{6}$$

If the measured values are standardized with their true values and standard deviations, they are pragmatically assumed

to be random variables from the same distribution (univariate) with zero mean and unit deviation. Then similar to (6), but with a general matrix A for the linear case, additional constraints, and $l_i = 1$, data reconciliation problem can be stated as:

$$\begin{aligned} &\min \sum_{i=1}^{n} \frac{(y_{i,1} - x_{i,1})^2}{\sigma_{i,1}^2} \quad \text{such that} \\ &Ax = 0, \\ &A \text{ is the process matrix} \\ &Lb \le x \le Ub \end{aligned} \tag{7}$$

Formulation (7) can be further generalized to include the unmeasured variables (u) and nonlinear process model constraints (f, g), which is frequently used in the data reconciliation literature.

$$\begin{aligned} &\min (y - x) Q^{-1} (y - x) \quad \text{such that} \\ &g(x, u) \ge 0 \\ &f(x, u) = 0 \\ &Lb_x \le x \le Ub_x \\ &Lb_u \le u \le Ub_u \\ &\text{where } Q = \text{diag}[\sigma_{1,1}^2, \sigma_{2,1}^2, \ldots, \sigma_{n,1}^2] \end{aligned} \tag{8}$$

An optimum x_i (called x_i^*) to the problem (8) is expected to result in the differences

$$y_1 - x_1^*, y_2 - x_2^*, y_3 - x_3^*, \ldots, y_n - x_n^*$$

from a distribution with zero mean.

A fundamental method to determine whether the measurements are from a distribution with zero mean is applying hypothesis testing with H_0 being "μ is 0" and H_1 being "μ is not equal to 0", where μ denotes the mean. The test statistic for this procedure is

$$t = \frac{\hat{E}(y_i - x_i^*) - 0}{\sqrt{\hat{V}(yi - x_i^*)}} \tag{9}$$

where $\hat{E}$ is an estimate for the expected value of $(y_i - x_i^*)$ and $\hat{V}$ is an estimate for the variance. This test statistic in Eq. (9) is the basis of classical gross error detection procedures. If a particular probability distribution function can be assumed for t, larger t values will describe less likely instances and provide proof for the truth of the hypothesis H_1, i.e. the existence of a gross error (outlier).

3. Definition and comparison of different objective functions for data reconciliation

Different objective functions besides the weighted least squares (WLS) in (8) can be used for data reconciliation. The WLS objective function assumes measurement errors from a distribution with zero mean and known variance. For any possible deviation from this assumption, another objective function, which does not require this assumption, can be a better candidate. This is especially the case when the measurements contain some gross errors. A gross error in a measured variable causes "smearing", contaminating the estimates for other measured variables. Increasing the breakdown point of the data reconciliation method used can reduce "smearing". The breakdown point for location estimators is defined as "the smallest fraction of free contamination that can carry the estimated value beyond all bounds" (Hampel, 1985) and is close to zero for weighted least squares method. In other words, even a single measurement with gross error is enough to invalidate the basis of WLS method causing "smearing".

Objective functions with better breakdown points can be found from Normal-like distribution functions with heavy tails or combining two distributions to account for the contamination caused by the outliers (gross errors), e.g. contaminated Normal distribution. Similar to Gauss's development of the Normal distribution function for residuals into the weighted least squares objective function (Deutsch, 1965), maximum likelihood functions can be utilized to derive these objective functions.

A maximum likelihood function is formed from the probability distribution function of the measured variable x_i, by maximizing the product of individual probability values for each measured variable.

$$\max P = \max \prod_i P_i \tag{10}$$

For the Normal distribution, the product in (10) becomes

$$\begin{aligned} &\max \prod_i P_i = \max \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(y_i - x_i)^2}{2\sigma_i^2}\right) \\ &\text{or} \quad \min -\sum_i \ln \exp\left(-\frac{(y_i - x_i)^2}{2\sigma_i^2}\right) + \sum_i \ln(\sqrt{2\pi}\sigma_i) \\ &\text{or} \quad \min \sum_i \frac{(y_i - x_i)^2}{2\sigma_i^2} \end{aligned} \tag{11}$$

which is equivalent to the weighted least squares objective function. Similarly, for contaminated Normal distribution function, this product becomes

$$\begin{aligned} \max \prod_i P_i = \max \prod_i &(1 - p_i) \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(y_i - x_i)^2}{2\sigma_i^2}\right) \\ &+ p_i \frac{1}{\sqrt{2\pi} b_i \sigma_i} \exp\left(-\frac{(y_i - x_i)^2}{2 b_i^2 \sigma_i^2}\right) \end{aligned} \tag{12}$$

or

$$\begin{aligned} \min - \sum_i \ln &(1 - p_i) \exp\left(-\frac{(y_i - x_i)^2}{2\sigma_i^2}\right) \\ &+ \frac{p_i}{b_i} \exp\left(-\frac{(y_i - x_i)^2}{2 b_i^2 \sigma_i^2}\right) + \sum_i \ln(\sqrt{2\pi}\sigma_i) \end{aligned}$$

where p_i is the probability and $b_i^2\sigma_i^2$ the variance of the contamination by a gross error.

For Logistic distribution, function (10) becomes

$$\max_i P_i = \max_i \frac{1}{\sigma_i}\frac{\exp((y_i - x_i)/\sigma_i)}{(1+\exp((y_i - x_i)/\sigma_i))^2} \quad \text{or}$$
$$\min_i \; 2\ln\left(1+\exp\left(\frac{(y_i - x_i)}{\sigma_i}\right)\right) - \left(\frac{(y_i - x_i)}{\sigma_i}\right) + \ln\sigma_i \tag{13}$$

and finally, for the Cauchy (Lorentz) distribution, the maximum likelihood objective function becomes

$$\max_i P_i = \max_i \frac{1}{\pi\sigma_i(1+(y_i - x_i)^2/\sigma_i^2)} \quad \text{or}$$
$$\min_i \; \ln(\pi\sigma_i) + \ln\left(1+\frac{(y_i - x_i)^2}{\sigma_i^2}\right) \tag{14}$$

The generalized maximum likelihood objective function, proposed by Huber (Huber, 1981) has the form

$$\min_i \; \rho\left(\frac{y_i - x_i}{\sigma_i}\right) \tag{15}$$

i.e. any reasonable monotone function, ρ, of $\varepsilon_i = (y_i - x_i)/\sigma_i$, the standard error, can be used for the data reconciliation formulation, provided that the gross errors have a reduced effect on the estimation of measured process variables. Therefore, the three maximum likelihood objective functions for contaminated Normal, Cauchy and Logistic distributions—with the proper tuning parameters—can be used without assuming the underlying measurement error probabilities. In addition, Fair function, "Lorentzian" function and Hampel's redescending M-estimator are three other robust generalized maximum likelihood objective functions that can be employed in the data reconciliation formulation. Among these three functions, Fair function was constructed using a combination of ordinary-least squares for small residuals and least-absolute residual (LAR) for large residuals (Fair, 1974), whereas "Lorentzian" function was introduced by Johnston and Kramer (1995) and Hampel's redescending M-estimator by Hampel (Andrews et al., 1972).

In classical estimation literature, a location (mean) estimation of a sample from a univariate distribution is calculated by Eq. (3). A robust estimate of the location can be obtained by changing the objective function in Eq. (3) with a robust function, such as a generalized maximum likelihood estimator (M-estimator). The efficiency of these estimators is defined (up to a common factor) as the inverse of the variance of the final estimate under the ideal model distribution, which is traditionally chosen as the Normal distribution (Hampel, 1985). If a rejection rule is imposed, the efficiency is calculated using the variance in the location estimates calculated after outlier values are eliminated from the sample. If a smaller critical value is used for rejection, the power of the rejection rule (similar to gross error detection) increases; however, the variance of the estimate, if there are actually no outliers in the sample, increases. This loss of efficiency is called "insurance premium" and can be used to "tune" the estimators with parameters. This tuning by efficiency values is necessary if one desires to compare the performance of different ρ functions and eventually the rejection rules designed on them.

The consequence of this tuning requirement is that in data reconciliation and gross error detection the performance of two different ρ functions can be compared properly only for (nearly) equal efficiency cases. This means that, for instance, Fair function with 95% efficiency can be compared with 95% efficient Hampel's redescending M-estimator.

The ρ functions that we studied are as follows:

WLS

$$\tfrac{1}{2}\varepsilon_i^2 \tag{16}$$

Contaminated Normal

$$-\ln\left((1-p_{CN})\exp\left(-\frac{\varepsilon_i^2}{2}\right) + \frac{p_{CN}}{b_{CN}}\exp\left(-\frac{\varepsilon_i^2}{2b_{CN}^2}\right)\right) \tag{17}$$

Cauchy

$$c_C^2\ln\left(1+\frac{\varepsilon_i^2}{c_C^2}\right) \tag{18}$$

Logistic

$$2\ln\left(1+\exp\left(\frac{\varepsilon_i}{c_{Lo}}\right)\right) - \frac{\varepsilon_i}{c_{Lo}} \tag{19}$$

"Lorentzian"

$$-\frac{1}{1+(\varepsilon_i^2/2c_L^2)} \tag{20}$$

Fair

$$2c_F^2\left[\frac{|\varepsilon_i|}{c_F} - \ln\left(1+\frac{|\varepsilon_i|}{c_F}\right)\right] \tag{21}$$

Hampel's redescending M-estimator

$$\begin{cases} \tfrac{1}{2}\varepsilon_i^2, & 0 \le |\varepsilon_i| \le a_H \\ a_H|\varepsilon_i| - \tfrac{1}{2}a_H^2, & a_H < |\varepsilon_i| \le b_H \\ a_H b_H - \dfrac{a_H^2}{2} + (c_H - b_H)\dfrac{a^2}{2}\left[1 - \left(\dfrac{c_H - |\varepsilon_i|}{c_H - b_H}\right)^2\right], & b_H < |\varepsilon_i| \le c_H \\ a_H b_H - \tfrac{1}{2}a_H^2 + (c_H - b_H)\tfrac{1}{2}a^2, & c_H < |\varepsilon_i| \end{cases} \tag{22}$$

Table 1
Tuning constants for different ρ functions with efficiency values 95.5%

ρ function	Tuning constants
Contaminated Normal	$b_{CN} = 10,\ p_{CN} = 0.235$
Cauchy	$c_C = 2.3849$
Logistic	$c_{Lo} = 0.602$
"Lorentzian"	$c_L = 2.6$
Fair	$c_F = 1.3998$
Hampel	$a_H = 1.35,\ b_H = 2.7,\ c_H = 5.4$

To compare the data reconciliation and gross error detection performance of these ρ functions, they were first standardized by properly tuning their parameters. Some functions have their tuning constants given as a function of asymptotic efficiency such as the Fair and Cauchy functions. However, these asymptotic variances "give only crude indications for the actual variances" for finite sample size (Hampel, 2002). Therefore, approximate finite sample variances and consecutively relative efficiencies were calculated by simulation and Monte Carlo studies (Hampel, 1985; Andrews et al., 1972). We performed a similar study for the above ρ functions with a sample size of 28 and 2000 simulation runs that resulted in the following tuning constant values (efficiency values are approximately 95.5%) given in Table 1.

Fig. 1 depicts individual standardized ρ functions in the objective function, showing that Fair and Logistic functions cases result in a convex objective function. The convexity of the objective function guarantees the global optimality of the nonlinear data reconciliation problem for a process, which can be described by only linear constraints.

Methods to measure the robustness of an estimator involve the use of the influence function, IF (Hampel, Ronchetti, Rousseeuw, & Stahel, 1986), which is defined for a sample x, an estimator T over an assumed distribution function F and a perturbed distribution function F_t as follows:

$$\mathrm{IF}(x, T, F) = \lim_{t\to 0} \frac{T(F_t) - T(F)}{t} = \frac{\partial}{\partial t}[T(F_t)]|_{t=0} \qquad (23)$$

The heuristic interpretation of this influence function is that "it describes the effect of an infinitesimal contamination at the point x on the estimate" (Hampel et al., 1986). Since the influence function is proportional to the derivative of the maximum likelihood function, the weight given to any gross error in the measurements while calculating the estimates can be seen in Fig. 2 (see Appendix A for details).

The influence function for WLS is proportional to the measurement error (derivative of Eq. (12)) justifying the low breakdown point and unbounded effect of large errors. The effect of larger errors is reduced for the ρ function of the Cauchy distribution, "Lorentzian" function and Hampel's redescending M-estimator, shown by gradually decreasing influence functions in the region of greater than 3.0 of the standard error. Therefore, these three ρ functions are called *redescending ρ functions*. Fair function and the ρ function of the Logistic distribution have a bounded influence by the large errors since their influence function increases slowly with respect to the measurement errors approaching a constant value for large errors. The influence of small measurement errors on the ρ function of the contaminated Normal distribution is the same as on the WLS; however, the influence decreases for larger errors and becomes proportional to very large errors after passing through a minimum (at standard error 4.7 in Fig. 2).

Collectively, methods with influence functions which remain bounded as the standard error increases, should be insensitive to gross errors when data reconciliation is conducted with them.

4. Obtaining different gross error detection criteria

"Statistically, a gross error is an error whose occurrence as realization of a random variable is highly unlikely" (Veverka & Madron, 1997). Therefore, the hypothesis testing approach to detect these unlikely occurrences works very well, provided that the measurement errors come from a known probability distribution. In other words, an advantage of having an underlying distribution function for the measurement errors (including the gross errors) is that the rejection of the gross errors can be performed using confidence level or α values. A measurement value which probably occurs less than $(\alpha \times 100)\%$ of the time can be detected as a gross error. This way the measurements with higher errors can be eliminated with a certainty of $(1 - \alpha)$. The value beyond which the measurements are considered as gross errors is called a *cut point*. The cut points for four distribution functions are given in Table 2.

Rejection of the gross errors by employing the hypothesis testing approach can give misleading results if the a priori assumption about the measurement error distribution is violated. Moreover, rejection criteria for the cases without a priori probability distribution functions can not be defined systematically.

Alternative definitions of possible rejection criteria have been proposed in the literature. For instance, the rejection criterion proposed by Farris and Law (1979) for the contaminated Normal distribution function case is equivalent to defining the cut point (x^c) as

$$\max\{P(\text{the measurement is larger than } x^c \text{ and is an outlier}) - P(\text{the measurement is larger than } x^c \text{ and is not an outlier})\}$$

Table 2
Cut points for four distribution functions at $\alpha = 0.03$

Probability distribution function	(mean, variance)	Cut points for $\alpha = 0.03$
Normal	(0, 1)	±2.16
Contaminated Normal	0.235 (0, 100) + 0.765 (0, 1)	±15.2
Logistic	(0, 1)	±4.2
Cauchy	(0, 1)	±21.0

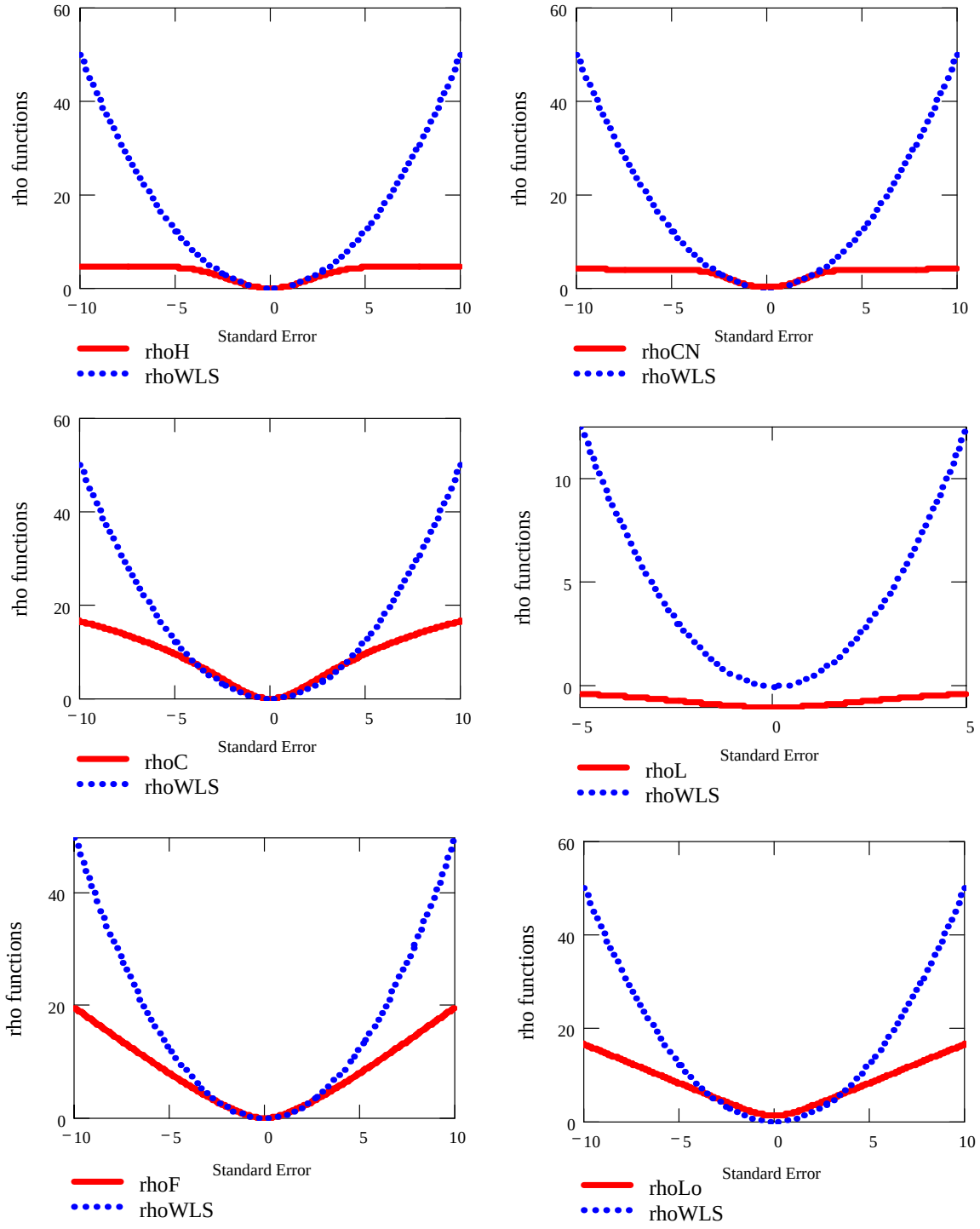

Fig. 1. Individual ρ functions in the objective function compared with weighted least squares objective (rhoL: "Lorentzian", rhoF: Fair function, rhoCN: contaminated Normal distribution, rhoC: Cauchy distribution, rhoLo: Logistic distribution, rhoH: Hampel's redescending M-estimator).

This cut point falls on the descending part of the influence function (Fig. 3). By examining the first and second derivatives of the influence function, additional cut points can be defined systematically, for instance, the minimum, maximum and inflection points of the influence functions can become possible candidates. Choosing a smaller cut point (critical value) can improve the gross error detection but will also increase the false detection and the variance of the estimates under the ideal condition, i.e. if there are no gross errors in the measurements.

For the ρ functions such as Fair function, and the ρ function for the Logistic distribution, the cut points can not be found using this procedure, because their influence functions do not have single maximum, minimum of inflection points.

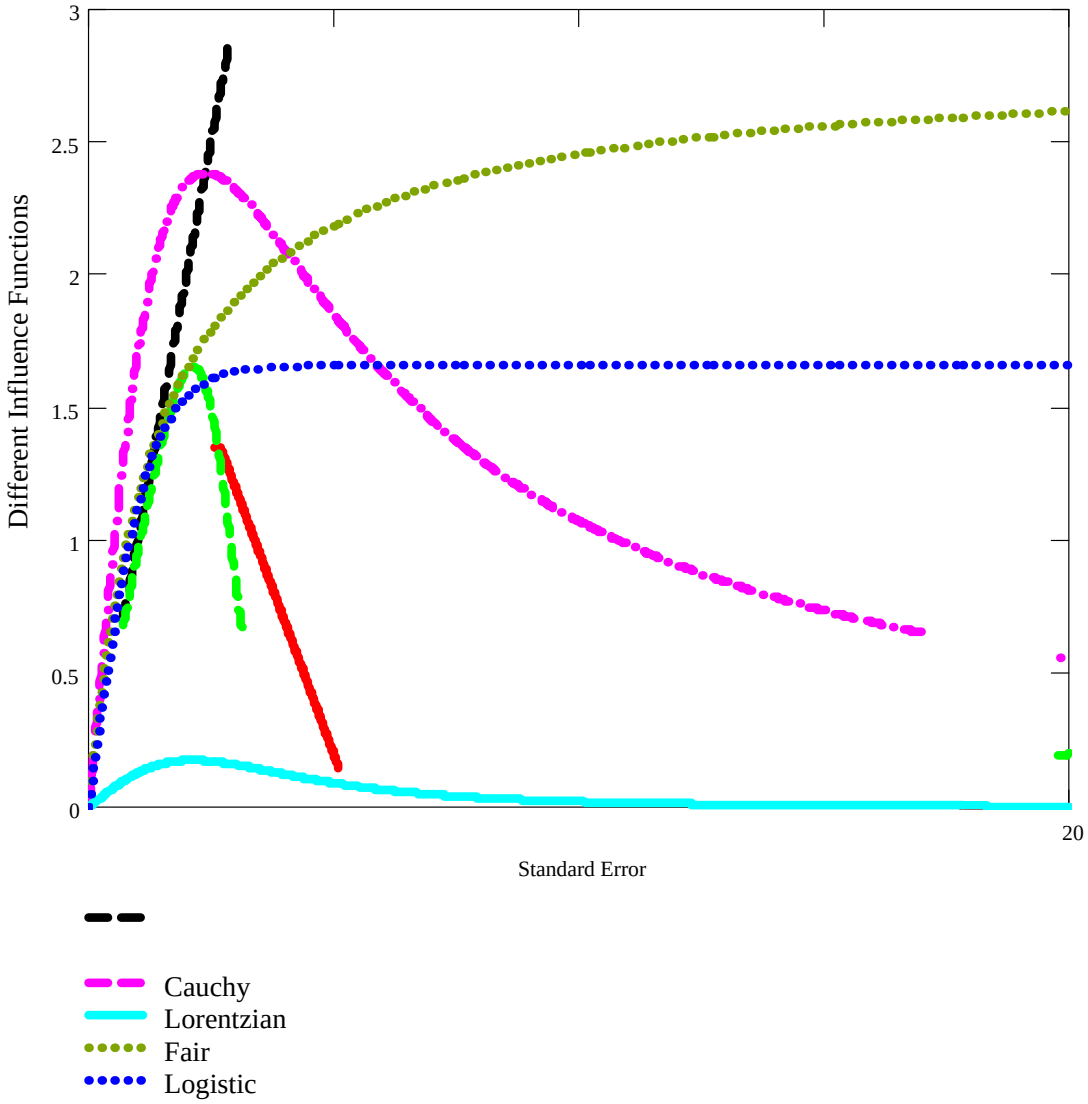

Fig. 2. Relative influence of the errors in the measurement to the objective function (WLS: weighted least squares, CN: contaminated Normal).

However, provided that the functions have same efficiencies, the cut points for redescending ρ functions prove as reasonable cut point candidates for these "non-redescending" ρ functions.

An alternative rejection rule, which is based on the robust median and median deviation, has the code name X84 (Hampel et al., 1986). This rule rejects the measurements, for which the residuals after the data reconciliation are more than 5.2 median deviations away from the median of the residuals. The median deviation (median absolute deviation) is the median of the absolute residuals from the median. This rule does not have a predetermined and ρ-function-dependent cut point. Therefore, it can be used with any ρ function described above.

In the following section, the performance of these different gross error detection criteria is evaluated with numerical experiments.

5. Examples

Numerical experiments for data reconciliation and gross error detection reported in literature have been applied to relatively small plant simulation problems, and there are few cases where results for industrial examples are given (Chen et al., 1998; Jordache et al., 2001; Sanchez, Sentoni, Schbib, Tonelli, & Romagnoli, 1996; Weiss, Romagnoli, & Islam, 1996). In this study, both small simulation and actual plant examples are solved, and their results are compared. We have compared the performance of weighted least squares, the modified iterative measurement test method (MIMT) and ρ functions for the contaminated Normal (CN), Cauchy (C) and Logistic (Lo) distributions, along with "Lorentzian" (L), Fair function (F) and Hampel's redescending M-estimator (H).

For each method, three different gross error detection criteria are tested except for the MIMT method. The summary

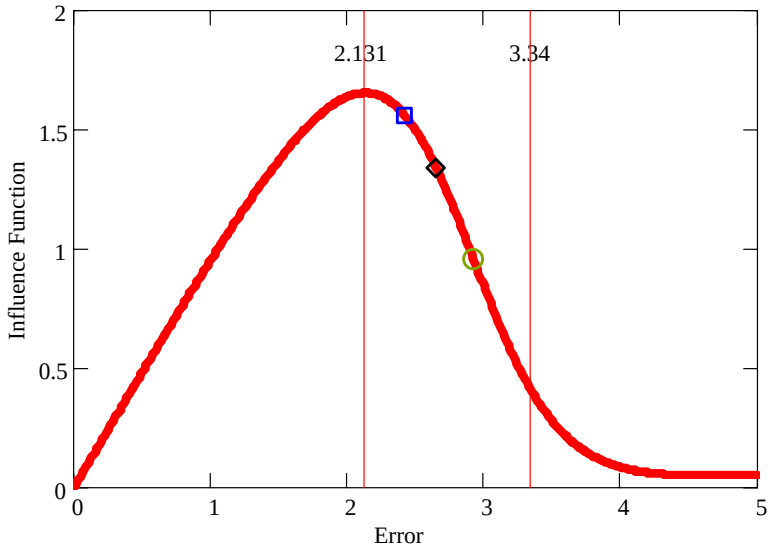

Fig. 3. Influence function for the ρ function of contaminated Normal distribution and five different cut points for gross error detection (first marker: maximum of the influence function (2.131), (□): inflection point of the first derivative of the influence function (2.42), (◇) Farris–Law criteria (2.65), (○) inflection point of the influence function (2.92); second marker: another inflection point of the first derivative of the influence function (3.34)).

of each criterion for different methods is given in Table 3. Data reconciliation and gross error detection using MIMT are performed as described in Kim, Kang, Park, and Edgar (1997).

Performance measures to evaluate different gross error detection criteria employed are the overall power (OP):

$$\text{OP} = \frac{\text{Number of gross errors correctly identified}}{\text{Number of gross errors simulated}} \tag{24}$$

and average number of Type I errors (AVTI) (Narasimhan & Jordache, 2000).

$$\text{AVTI} = \frac{\text{Number of gross errors wrongly identified}}{\text{Number of simulation trials made}} \tag{25}$$

Average and median total error reductions (TER) (Serth, Valero, & Heenan, 1987) are used to compare the data validation performance of different methods, where x_i^t is the true value for the ith measured variable.

$$\text{TER} = \frac{\sqrt{\sum_{i=1}^{n}((y_i - x_i^t)/\sigma_i)^2} - \sqrt{\sum_{i=1}^{n}((x_i^* - x_i^t)/\sigma_i)^2}}{\sqrt{\sum_{i=1}^{n}((y_i - x_i^t)/\sigma_i)^2}} \tag{26}$$

Table 3
Methods and their gross error detection criteria

Method	Gross error detection criteria
WLS	Cut points at $\alpha = 0.05$, $\alpha = 0.025$ and X84
CN	Cut points from influence function (2.131, 3.34) and X84
Cauchy	Cut points from influence function (2.385, 4.131) and X84
Logistic	Same as CN
"Lorentzian"	Cut points from influence function (2.123, 3.658) and X84
Fair	Same as CN
Hampel's M	Cut points from influence function (2.131, 3.34) and X84
MIMT	Cut point at $\alpha = 0.05$

Performance test procedure for the examples with known true values of the measured variables consists, in general, of the following steps.

1. Using true values such as design data, measurement sets are created for each variable by adding noise from Normal and Cauchy distributions with equal probability, i.e. half of the simulated measurement errors has a Normal probability distribution and the other half are from Cauchy probability distribution. Therefore, the assumption that the random errors have a particular probability distribution, is relaxed.
2. Add gross errors to measurements depending on the percentage of gross error occurrence.
3. Solve data reconciliation problem for each of the eight methods.
4. Using different gross error detection criteria calculate the performance measures for each method, i.e. calculate OP and AVTI.
5. Calculate total error reduction for each method as a performance measure for the data reconciliation if the true values are given, i.e. calculate TER.

All models for the examples and random number generation for the Monte Carlo simulations are implemented in GAMS (Brooke, Kendrick, & Meeraus, 1992). The data reconciliation formulations are solved with the NLP solvers

CONOPT2 and MINOS5. In the first five examples, the piece-wise linear Hampel's redescending M-estimator is modeled as an external function coded in the programming language C and called by GAMS (Kalvelagen, 2002). For the last two problems, these discontinuities are smoothed as described in Arora and Biegler (2001). All calculations for the performance measures and the gross error detection rule X84 are implemented with Perl.

5.1. Examples from literature

The methods presented above are tested first on examples used in various literature articles in the last three decades. Two of these examples (Examples 1 and 2) contain linear and the remaining three (Examples 3-5) nonlinear process models. Except in Example 5, the lower bounds on the variables are set to 50% of the true values and the upper bounds to twice the true values. In Example 5, the lower bounds for all variables are 50% of the true values whereas the upper bounds are set to 150% of the true values.

Example 1 (Ripps, 1965). This example involves a simple chemical reactor with two entering and two leaving mass flows. All four variables are measured in the system, and they are related by three linear mass balance equations (Ripps, 1965; Romagnoli & Sanchez, 2000). For the Monte Carlo study, random measurements are created from Normal and Cauchy distributions as outlined above. Outliers were created in 10% of the measurements randomly by adding or subtracting 10–100% of the true values. With the exceptions of the Hampel's redescending M-estimator and MIMT, all runs were executed independently and with the same initial conditions. For MIMT, all consecutive runs were initiated with the resulting values of the previous run. Hampel's redescending M-estimator converged to an inferior optimal if it was not initialized with the results from Cauchy distribution ρ function or Fair function method.

The results of Monte Carlo study runs for each method are shown in Table 4. The ρ function of the Cauchy distribution shows the best performance with second highest overall power and lowest average number of Type I errors if the first cut point at 2.385 is used. Rule X84 seems to be conservative for this example, and the factor 5.2 can be reduced to improve the results. The comparison of the data reconciliation performance shows that ρ function of the Cauchy distribution is the most effective one among other methods with 77.5% mean and 92.3% median total error reduction. Median TER indicates that 50% of the TER values are above 0.923.

Since the problem has only four measured variables, a special attention to the breakdown point is necessary. The highest breakdown point achievable is 50% (Rousseeuw & Leroy, 1987) which corresponds to two gross errors (outliers) in this case. Therefore "smearing" can occur if the example is solved with two or more gross errors.

Example 2 (Serth and Heenan, 1986). Our second example considers a steam metering system with 28 variables (all measured) and 12 linear equations. The measured values are created using the correct flow rates Serth and Heenan (1986) and 25% of the observations have gross errors ranging from 10 to 100% of the true values.

The results in Table 5 show that modified MIMT has the best performance in data reconciliation with highest mean and median total error reduction. It also possesses the lowest average number of Type I errors in 1000 simulation runs. Cauchy distribution ρ function also performs well, considering that it requires a single NLP solution whereas MIMT requires seven iterations on the average.

Example 3 (Serth et al., 1987). Our first nonlinear example consists of a metallurgical grinding process with 12 equations and 24 variables. Nine mass flow rates and 15 mass fractions are created by using the correct values with addition of measurement errors from Normal and Cauchy distributions. The gross errors generated are on the average 25% of the measured variables and their amount range from 10 to 100% of the true values.

Table 4
Performance of different methods for Example 1

	MIMT	H	WLS	CN	Cauchy	L	Fair	Logistic
Number of runs	650	506	506	506	506	506	506	791
Total GE	253	213	213	213	213	213	213	240
Runs with GE	212	183	183	183	183	183	183	211
OP (GED #1)	0.775	0.817	0.930	0.897	0.894	0.784	0.911	0.896
AVTI (GED #1)	0.363	0.500	0.636	0.397	0.292	0.536	0.504	0.458
OP (GED #2)	–	0.765	0.911	0.859	0.784	0.676	0.751	0.808
AVTI (GED #2)	–	0.462	0.581	0.322	0.213	0.429	0.362	0.322
OP (GED #3)	–	0.512	0.333	0.531	0.596	0.460	0.484	0.512
AVTI (GED #3)	–	0.176	0.146	0.182	0.198	0.213	0.174	0.187
Mean TER	0.736	0.625	0.676	0.751	0.775	0.658	0.696	0.718
Median TER	0.915	0.878	0.836	0.906	0.923	0.892	0.901	0.894

GE: gross errors; OP: overall power; AVTI: average number of Type I errors; TER: total error reduction; GED #i: gross error detection criteria number (i = 1, 2, 3 for first and second cut points and rule X84, respectively).

Table 5
Performance of different methods for Example 2

	MIMT	H	WLS	CN	Cauchy	L	Fair	Logistic
Number of runs	1000	929	1100	1052	1000	1153	1000	1085
Total GE	6955	6456	7579	7415	6914	8053	6908	6824
Runs with GE	1000	929	1100	1052	999	1153	999	1084
OP (GED #1)	0.684	0.705	0.759	0.724	0.720	0.744	0.744	0.712
AVTI (GED #1)	1.364	2.118	7.645	3.371	2.255	4.692	4.193	3.253
OP (GED #2)	–	0.684	0.751	0.705	0.678	0.718	0.704	0.678
AVTI (GED #2)	–	1.826	7.296	3.203	1.500	4.144	2.622	2.038
OP (GED #3)	–	0.700	0.338	0.689	0.702	0.707	0.650	0.670
AVTI (GED #3)	–	2.713	0.882	3.281	2.421	4.846	2.499	2.699
Mean TER	0.558	0.505	0.412	0.455	0.525	0.384	0.494	0.460
Median TER	0.552	0.504	0.385	0.466	0.516	0.400	0.472	0.447

GE: gross errors; OP: overall power; AVTI: average number of Type I errors; TER: total error reduction; GED #*i*: gross error detection criteria number ($i = 1, 2, 3$ for first and second cut points and rule X84, respectively).

Similar to Example 2, modified MIMT outperformed other methods in data reconciliation. Once again, the ρ function of Cauchy distribution shows that comparable, if not superior results can be achieved in a single NLP solution (see Table 6).

Example 4 (Pai and Fisher, 1988). In this example, there are six nonlinear equality constraints, five measured variables—all measurements are redundant—, and three observable unmeasured variables. On the average, 25% of the generated measurements are contaminated with gross errors ranging from 10 to 100% of the exact values reported in Pai and Fisher (1988).

As seen in Table 7, the ρ function of Cauchy distribution results in the highest total error reduction whereas the ρ function for contaminated Normal reaches the highest overall power but with more occurrences of Type I errors.

Example 5 (Swartz, 1989). Another widely used literature example is the nonlinear heat exchanger network problem described by Swartz (1989), and Romagnoli and Sanchez (2000). The system of four heat exchangers is modeled with 17 material and energy balances. The total number of variables in the system is 30, of which 16 are measured and the rest is unmeasured. There are 10 redundant and 6 non-redundant measured variables.

Gross errors are generated only for the redundant measured variables and on the average of 25% of the time. The magnitude of the errors range between 5 and 10 standard deviations for the flow rates and between 5 and 30 standard deviations for the temperature variables.

Most of the methods studied show poor data reconciliation results with close to none average total error reductions (Table 8). The ρ function for contaminated Normal and the "Lorentzian" function prove to be the best options for this case.

5.2. Industrial examples

Not many industrial examples have been investigated for the performance of different data reconciliation and gross error detection methods. The few cases in the open literature study industrial process subsystems such as reactors (Sanchez et al., 1996; Weiss et al., 1996), or utilize simulated plant measurements (Jordache et al., 2001) instead of real time plant data. The first industrial example involving

Table 6
Performance of different methods for Example 3

	MIMT	H	WLS	CN	Cauchy	L	Fair	Logistic
Number of runs	1000	1077	1076	1028	1006	1110	1000	1026
Total GE	5986	6398	6389	6172	5990	6569	5935	5546
Runs with GE	1000	1076	1075	1027	1005	1109	999	1025
OP (GED #1)	0.744	0.776	0.843	0.758	0.774	0.757	0.822	0.799
AVTI (GED #1)	1.744	2.234	8.571	4.488	2.583	3.722	4.986	3.675
OP (GED #2)	–	0.760	0.836	0.742	0.733	0.722	0.779	0.764
AVTI (GED #2)	–	1.964	8.172	4.330	1.809	3.149	3.102	2.362
OP (GED #3)	–	0.736	0.209	0.607	0.703	0.642	0.585	0.667
AVTI (GED #3)	–	1.945	0.391	2.771	1.757	2.478	1.143	1.492
Mean TER	0.622	0.585	0.423	0.450	0.587	0.475	0.552	0.539
Median TER	0.625	0.579	0.400	0.477	0.583	0.511	0.538	0.526

GE: gross errors; OP: overall power; AVTI: average number of Type I errors; TER: total error reduction; GED #*i*: gross error detection criteria number ($i = 1, 2, 3$ for first and second cut points and rule X84, respectively).

Table 7
Performance of different methods for Example 4

	MIMT	H	WLS	CN	Cauchy	L	Fair	Logistic
Number of runs	1000	1058	1000	1032	1000	1000	1000	1018
Total GE	1265	1350	1265	1320	1265	1265	1265	1279
Runs with GE	771	824	771	804	771	771	771	785
OP (GED #1)	0.580	0.601	0.597	0.666	0.614	0.639	0.627	0.611
AVTI (GED #1)	0.225	0.341	0.322	0.411	0.280	0.342	0.351	0.330
OP (GED #2)	–	0.504	0.578	0.588	0.469	0.526	0.515	0.494
AVTI (GED #2)	–	0.278	0.271	0.315	0.136	0.186	0.159	0.161
OP (GED #3)	–	0.442	0.180	0.468	0.386	0.420	0.333	0.335
AVTI (GED #3)	–	0.298	0.160	0.280	0.235	0.250	0.232	0.247
Mean TER	0.538	0.321	0.493	0.369	0.542	0.478	0.526	0.511
Median TER	0.568	0.514	0.538	0.572	0.593	0.586	0.569	0.558

GE: gross errors; OP: overall power; AVTI: average number of Type I errors; TER: total error reduction; GED #*i*: gross error detection criteria number ($i = 1, 2, 3$ for first and second cut points and rule X84, respectively).

real plant data and process model was given in Chen et al. (1998).

In this subsection, the sulfuric acid process from (Chen et al., 1998) and a new alkylation process example are introduced as the first large-scale industrial examples investigated for the performance of different data reconciliation and gross error detection methods with real plant data. For the sulfuric acid process, the real plant design data was also available; therefore, an analysis similar to the literature examples were performed by accepting the plant design data as the true values of the measured variables. For these two industrial examples, the bounds on variables are estimated using available process data, process design data (only for the sulfuric acid plant), process engineer expertise and considering conversion properties for the nonlinear steady-state process simulation.

Example 6 (Sulfuric acid process). The sulfuric acid process modeled, is IMC Agrico contact sulfuric acid plant in Convent, LA, USA. The plant was designed by the Enviro-Chem System Division of Monsanto and began operation in March, 1992. It produces 3200 TPD 93 wt.%

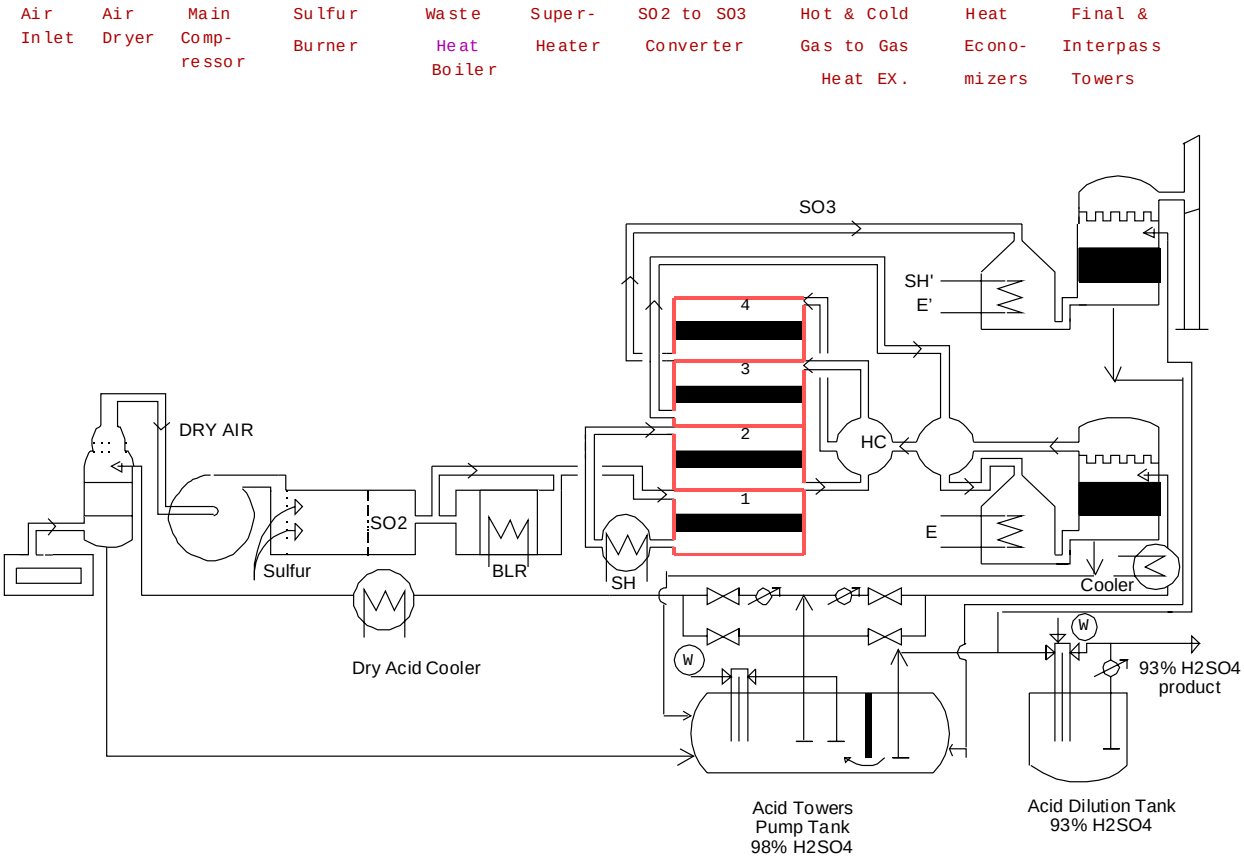

Fig. 4. The contact process for sulfuric acid (Chen, 1998).

sulfuric acid and process steam as a byproduct. This process incorporates many of the types of process units found in chemical plants such as packed bed catalytic reactors, absorption towers and heat exchanger networks, among others. It represents the state-of-art contact sulfuric acid technology.

In the contact process, molten sulfur is combusted with dry air; and the reaction is exothermic and goes to completion in the sulfur furnace. The gas leaving the burner is composed of sulfur dioxide, nitrogen, and unreacted oxygen at approximately 1400 °K. Heat from this gas is recovered in the waste heat boiler as byproduct steam. The gas enters the packed bed catalytic reactor that consists of four beds packed with two different types of vanadium pentoxide catalyst. Here, sulfur trioxide is produced from sulfur dioxide. The reaction is exothermic and approaches equilibrium exiting each bed. Heat is removed to shift the equilibrium, and this heat is used to produce steam. Also, the equilibrium conversion is increased in the fourth catalyst bed by removing SO_3 in the inter-pass absorption tower. In the final absorption tower, SO_3 is removed from the gas with 98 wt.% sulfuric acid. Gases exiting the final absorption tower go to the stack with less than 400 ppm SO_2 as required by regulations for emissions, no more than 4.0 lb of sulfur dioxide per ton of sulfuric acid produced. A flow diagram of the process is given in Fig. 4.

An open form equation-based model was developed from the process flow diagram and process design data. The packed bed catalytic reactor was simulated with a kinetic model developed by Harris and Norman (1972) and Richard (1987). The process model has 43 measured variables, 732 unmeasured variables, 11 parameters and 761 linear and nonlinear equality constraints. The 43 process measurements obtained from the distributed control system included 25 temperature, 11 flowrate, 2 pressure and 5 composition measurements. The standard deviations were determined based on 61 plant data sets from 11 consecutive days. These measured variables and their standard deviations were given

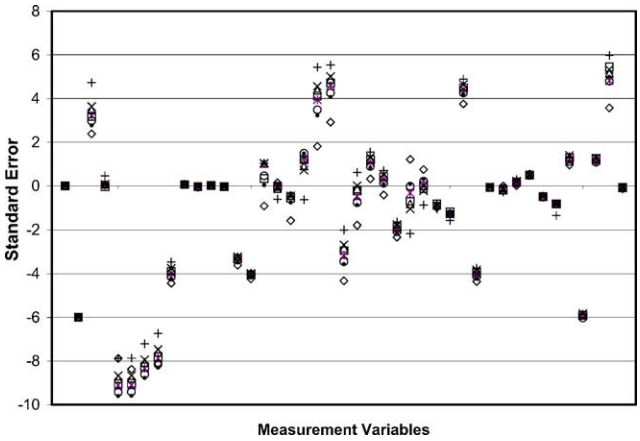

Fig. 5. Standard errors in measurements after reconciliation of the plant data at the first steady state (for Hampel's redescending M-estimator the error in the second measured value is −36 and not shown): (◇) MIMT; (•) Hampel; (+) WLS; (□) CN; (X) Cauchy; (○) Lorentzian; (×) Fair; ( ) Logistic.

(Chen, 1998). Of these 43 measurements, 14 are required to determine the state of the process.

The performance of eight different methods for data reconciliation along with three different gross error detection criteria were evaluated first with using the real plant design data to generate measurement values and gross errors. The procedure applied for the small-scale literature examples was replicated. On the average, 15% of the simulated measurements were contaminated with gross errors ranging between 3 and 30 standard deviations in magnitude.

The results depicted in Table 9 shows that the ρ function of Cauchy distribution has the best data reconciliation performance among the other seven methods. Same ρ function with the second cut point also gives one of the best compromises between overall power and average number of Type I errors. Unlike in the previous small-scale examples, the X84 rule shows promising performance for most of the methods.

Table 8
Performance of different methods for Example 5

	MIMT	H	WLS	CN	Cauchy	L	Fair	Logistic
Number of runs	1000	1018	1000	1016	1000	1000	1000	1072
Total GE	2503	2529	2503	2527	2503	2503	2503	2529
Runs with GE	950	962	950	960	950	950	950	1002
OP (GED #1)	0.251	0.550	0.472	0.605	0.349	0.476	0.371	0.320
AVTI (GED #1)	0.936	0.959	2.105	0.558	0.842	0.627	1.475	1.107
OP (GED #2)	–	0.538	0.450	0.599	0.323	0.457	0.303	0.282
AVTI (GED #2)	–	0.917	1.925	0.531	0.587	0.421	1.010	0.816
OP (GED #3)	–	0.499	0.109	0.591	0.360	0.485	0.294	0.305
AVTI (GED #3)	–	0.879	0.335	0.622	0.966	0.797	1.075	1.091
Mean TER	0.027	−0.222	0.043	0.307	0.150	0.277	0.051	0.017
Median TER	0.086	0.161	0.152	0.378	0.168	0.255	0.136	0.108

GE: gross errors; OP: overall power; AVTI: average number of Type I errors; TER: total error reduction; GED #*i*: gross error detection criteria number ($i = 1, 2, 3$ for first and second cut points and rule X84, respectively).

Table 9
Performance of different methods for Example 6

	MIMT	H	WLS	CN	Cauchy	L	Fair	Logistic
Number of runs	500	500	504	524	514	510	500	509
Total GE	3120	3392	3305	3346	3215	3181	3181	3026
Runs with GE	500	500	504	524	514	510	500	509
OP (GED #1)	0.841	0.863	0.884	0.912	0.889	0.904	0.907	0.896
AVTI (GED #1)	3.064	4.556	6.706	5.179	4.907	5.484	7.802	6.796
OP (GED #2)	–	0.833	0.879	0.897	0.848	0.876	0.866	0.844
AVTI (GED #2)	–	3.168	6.163	3.532	2.580	3.122	4.057	3.220
OP (GED #3)	–	0.819	0.670	0.871	0.827	0.852	0.778	0.784
AVTI (GED #3)	–	2.892	1.730	2.882	2.397	2.698	2.406	2.083
Mean TER	0.721	0.662	0.636	0.708	0.759	0.679	0.665	0.653
Median TER	0.767	0.714	0.661	0.778	0.802	0.779	0.682	0.689

GE: gross errors; OP: overall power; AVTI: average number of Type I errors; TER: total error reduction; GED #*i*: gross error detection criteria number ($i = 1, 2, 3$ for first and second cut points and rule X84, respectively).

Table 10
Number of detected gross errors and total error reductions for the two steady states of the sulfuric acid process

	MIMT	H	WLS	CN	Cauchy	L	Fair	Logistic
Steady state 1								
GED #1	13	16	15	16	16	16	16	16
GED #2	–	12	14	13	9	12	14	14
GED #3	–	7	3	5	5	6	5	4
TER	0.571	0.567	0.467	0.555	0.552	0.566	0.528	0.535
Steady state 2								
GED #1	13	13	13	16	16	16	16	16
GED #2	–	12	13	13	12	12	13	13
GED #3	–	12	8	10	8	13	8	12
TER	0.573	0.569	0.476	0.573	0.569	0.577	0.546	0.555

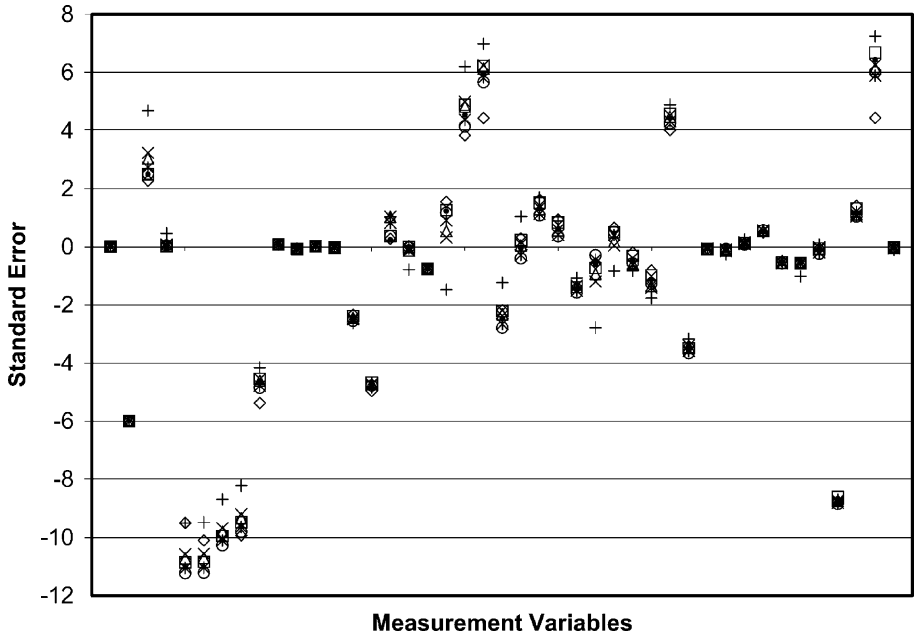

Fig. 6. Standard errors in measurements after reconciliation of the plant data at the second steady state (once again for Hampel's redescending M-estimator the error in the second measured value is −36 and not shown): (◇) MIMT; (•) Hampel; (+) WLS; (□) CN; (⊠) Cauchy; (○) Lorentzian; (×) Fair; ( ) Logistic.

For the sulfuric acid plant, two different steady-state operation points were determined from the data obtained on 6 and 10 June 1997. As can be seen in Table 10, except the weighted least squares method, all methods performed similarly for the data reconciliation. However, the total error reduction is 9–38% less than the average of the Monte Carlo study results shown in Table 9. This decrease is expected since there are approximately twice as many gross errors in the plant data. On the average, 6–7 gross errors are created for each Monte Carlo runs, whereas the gross errors detected by the methods requiring single NLP solution vary between 4 and 16 for the first steady state and between 8 and 16 for the second steady state. Considering the overall power results in Table 9, MIMT-GED #1, CN-GED #2 and Cauchy-GED #2 methods can be used to narrow these ranges of number of gross errors, namely 9–13 for the first steady-state and 12–13 for the second steady state. Standard (normalized) errors of the measurements for these two plant operation points (Figs. 5 and 6) also show that at the first steady state more variables fall into the critical region where the distinction of random errors and gross errors is difficult. The cause for the gross error detection of three temperature measurements (measurement numbers 32, 40 and 42) is identified as the instrument measuring errors. Four flow rates (measurement numbers 5, 6, 7 and 8) in the plant data were detected at the same time containing gross errors since all four flow rates were calibrated from the same measurement sources, namely the discharge pressure of compressor and the speed of turbine. Therefore, the measuring error in either/both discharge pressure of the compressor or/and speed of the turbine would cause gross errors in these four flow rates.

Example 7 (Alkylation process). The second large-scale industrial example involves a commercial, sulfuric acid catalyzed alkylation plant at the Motiva Enterprises Refinery in Convent, LO, USA. Motiva alkylation process is a 15,000 BPD STRATCO Effluent Refrigerated Alkylation Plant. The heart of the process is the STRATCO reactor or contactor, which contacts the reactants in a high-velocity propeller stream and removes heat from the exothermic reaction.

In the STRATCO Effluent Refrigerated Alkylation process, light olefins (propylene, butylenes) are reacted with isobutane in the presence of sulfuric acid catalyst to form hydrocarbons, mainly in the iC_7 to iC_8 range, called alkylate.

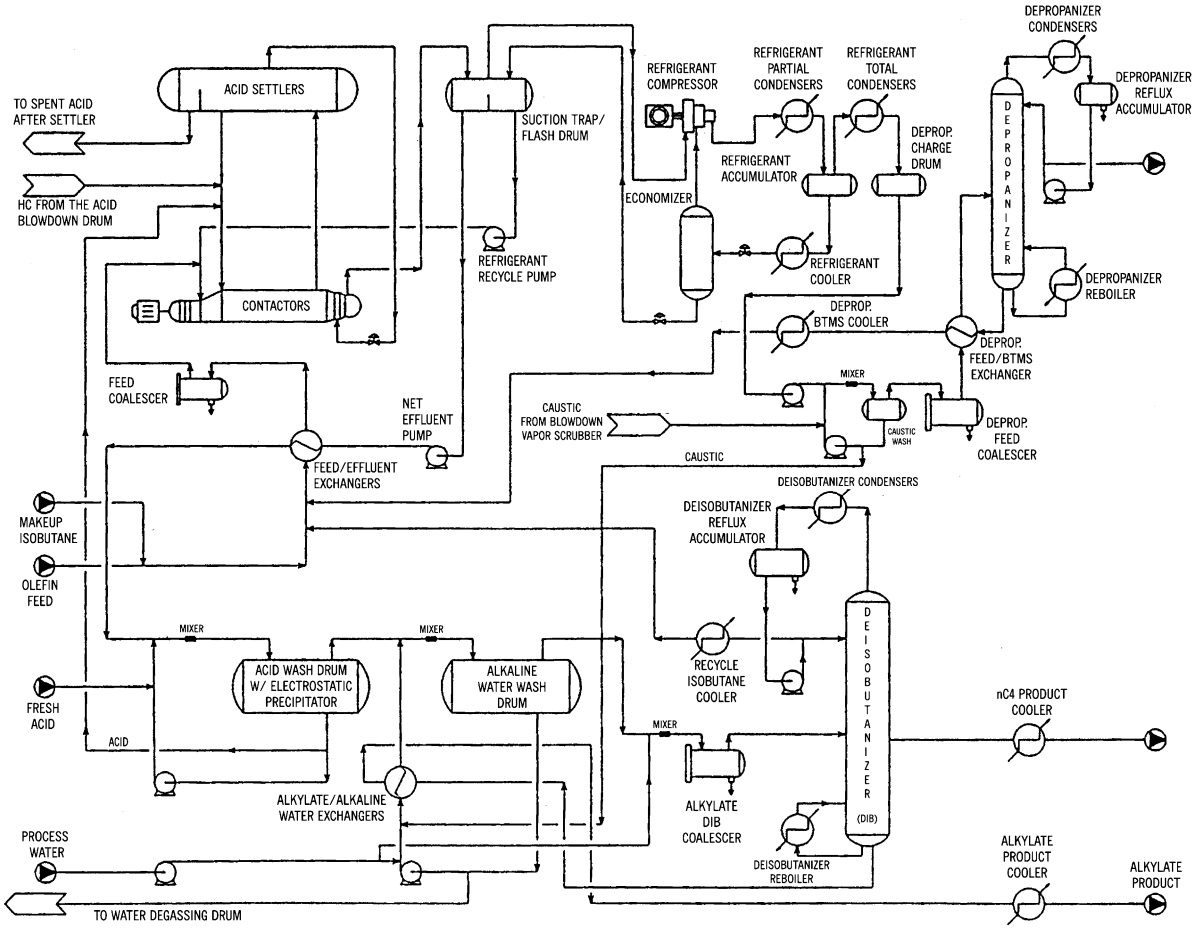

Fig. 7. Sulfuric acid alkylation process (Vichailak, 1995).

Table 11
Number of detected gross errors for the three steady states of the alkylation process

	MIMT	H	WLS	CN	Cauchy	L	Fair	Logistic
Steady state 1								
GED #1	23	27	32	34	34	34	44	39
GED #2	–	23	32	28	24	26	31	31
GED #3	–	27	25	27	23	27	19	31
Steady state 2								
GED #1	28	34	36	39	34	39	42	43
GED #2	–	30	35	34	29	32	33	35
GED #3	–	29	25	30	29	32	26	27
Steady state 3								
GED #1	26	29	33	32	29	31	41	32
GED #2	–	27	33	29	26	28	32	29
GED #3	–	30	23	30	26	30	25	28

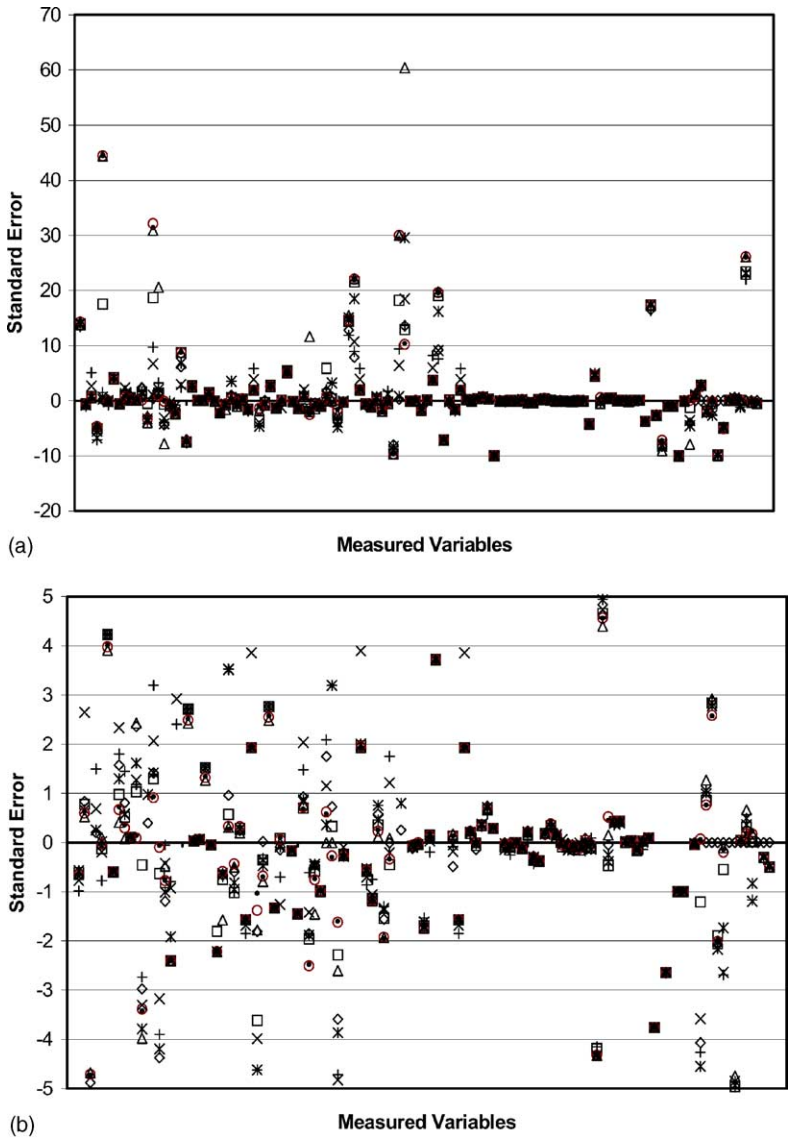

Fig. 8. Standard errors in measurements after reconciliation of the alkylation plant data at the first steady state: (a) all errors; (b) errors between −5 and 5. (◇) MIMT; (•) Hampel; (+) WLS; (□) CN; (X) Cauchy; (○) Lorentzian; (×) Fair; ( ) Logistic.

The alkylate product is a mixture of gasoline boiling range branched hydrocarbons which is blended with the refinery gasoline pool to increase the gasoline octane.

A simplified process flow diagram for a generic sulfuric acid alkylation process is given in Fig. 7. Specifically, Motiva alkylation process consists of five distinct sections, namely reaction, refrigeration, depropanizer, deisobutanizer and saturate deisobutanizer sections. The process has four reactor pairs and four acid settlers. In the reaction section, there are three feed streams: the olefin feed, the isobutane feed and the recycled olefin/isobutane mixture. The olefin feed contains the light olefins that are reacted with isobutane in the alkylation unit's STRATCO stirred reactors. The isobutane stream is in excess to fully react with all of the olefins being charged to the unit.

The alkylation process model developed using process flow diagrams, process data and process systems expertise has 1579 mostly nonlinear equality and 50 inequality constraints. The process model has 112–122 measured variables (122 for the first and second steady states, and 112 for the third steady state investigated in this study), 1512–1522 unmeasured variables and 67 parameters. The process measurements obtained from the distributed control system include 31 temperature, 30 flowrate, four pressure and 47–57 composition measurements. These measured variables, their standard deviations and the details of the model are given in Özyurt, Pike, Hopper, Punuru, and Yaws (2001), and Rich et al. (2001).

For the alkylation plant, three different steady-state operation points were determined from the data obtained on

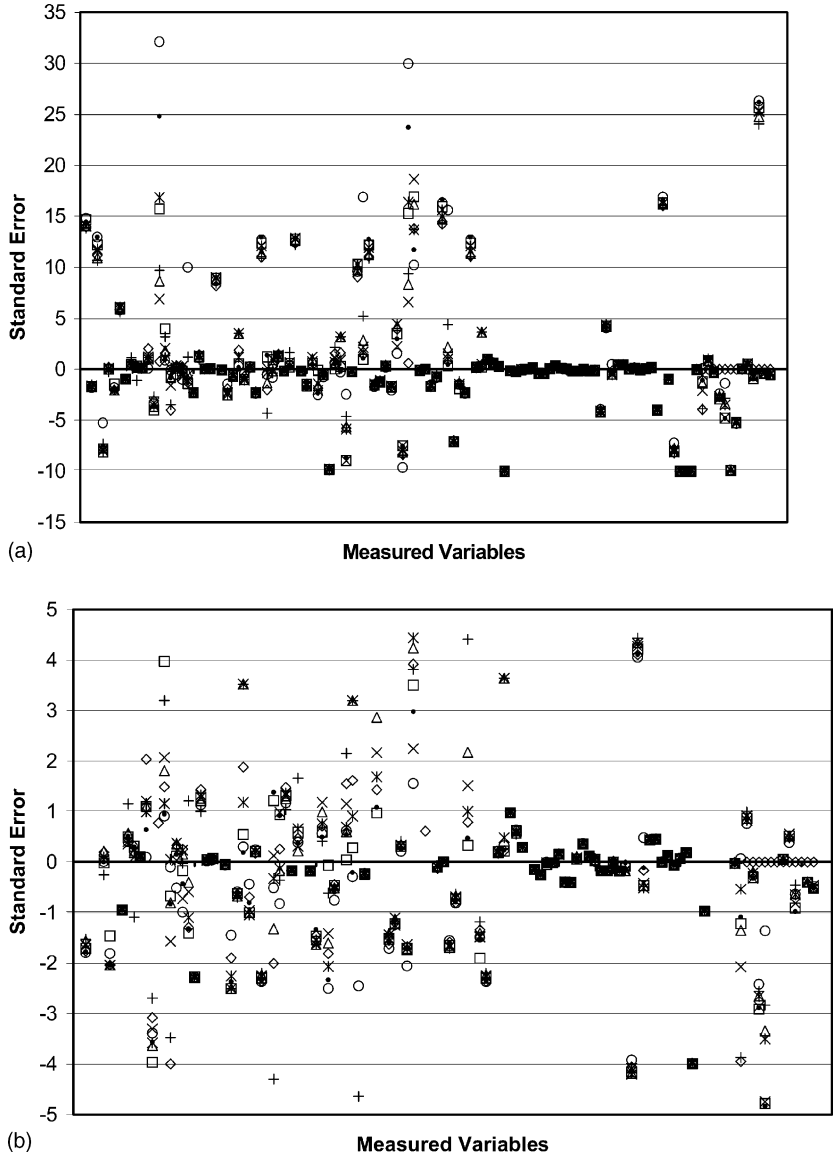

Fig. 9. Standard errors in measurements after reconciliation of the alkylation plant data at the second steady state: (a) all errors; (b) errors between –5 and 5. (◇) MIMT; (●) Hampel; (+) WLS; (□) CN; (✕) Cauchy; (○) Lorentzian; (×) Fair; ( ) Logistic.

8–9 November and 6–7 December 1998. The gross errors detected by the methods requiring single NLP solution vary between 19 and 44 for the first, between 25 and 43 for the second and between 23 and 41 for the third steady state (Table 11). MIMT-GED #1 and GED #2 for all other methods suggest that the second steady state has the most gross errors followed by the third steady-state operation point. Considering MIMT-GED #1, H-GED #2 and Cauchy-GED #2, the range for detected gross errors is 23–24, 28–30 and 26–27 for the first, second and third steady states, respectively. Since the true values of the measured variables are not available publicly, total error reduction, overall power and average Type I error values can not be calculated for the Motiva alkylation process. However, the measurements of the variables detected for containing gross errors were examined by the plant process engineer and found to be close to the limits or outside of the possible measurement ranges.

Standard (normalized) errors of the measurements for the three plant operation points are given in Figs. 8–10.

5.3. Discussion

In the previous two subsections, five small- and two large-scale examples were investigated, considering different data reconciliation and gross error detection methods. The systematic approach presented is novel in the following aspects compared to the previous studies:

a. The robust methods were tuned for the same efficiency values, which is required for a reasonable comparison between different methods.

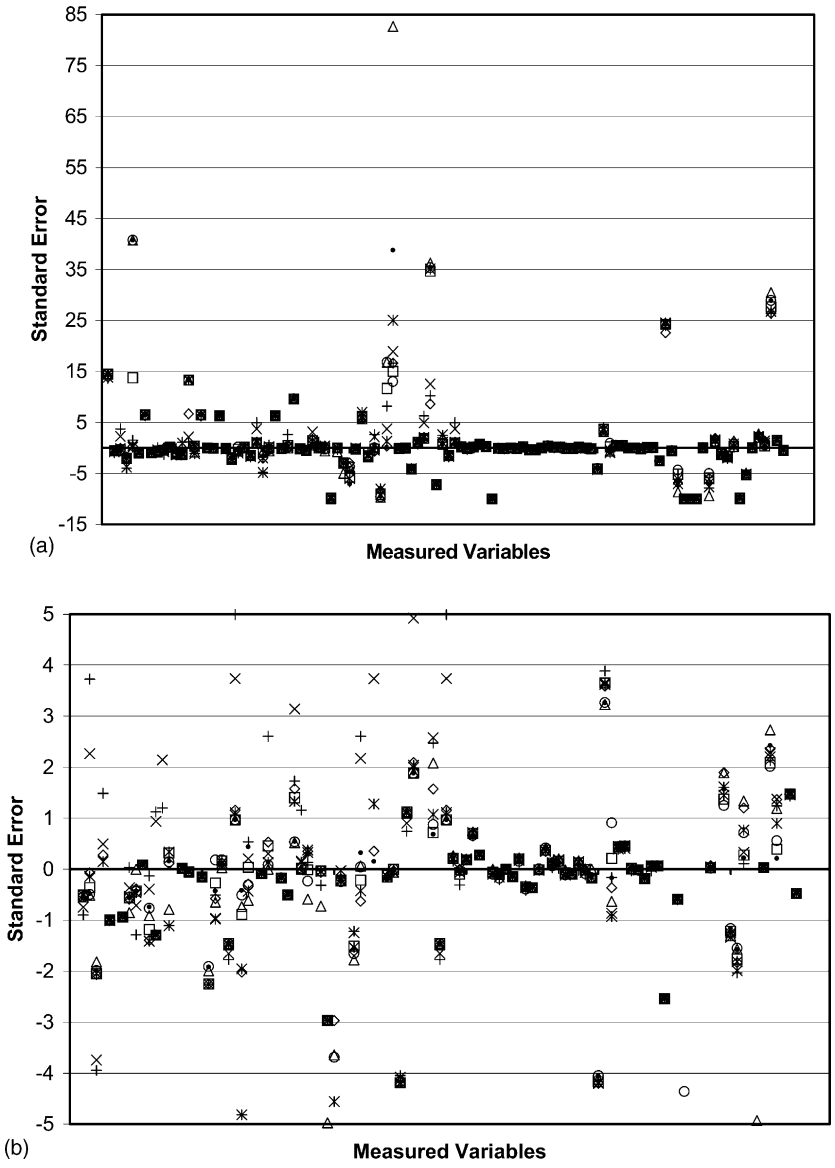

Fig. 10. Standard errors in measurements after reconciliation of the alkylation plant data at the third steady state: (a) all errors; (b) errors between –5 and 5. (◇) MIMT; (•) Hampel; (+) WLS; (□) CN; (⨯) Cauchy; (○) Lorentzian; (×) Fair; ( ) Logistic.

b. The tuning of the ρ functions for data reconciliation and the determination of gross error detection criteria were done separately.
c. The assumption that the random errors are from a Normal distribution was relaxed by generating a mixture of Normal and Cauchy random variables for simulated measurements.
d. In all examples, most of the eight methods were tested with the same simulated measurements and gross errors.
e. Two large-scale industrial systems were investigated with real plant data.

6. Conclusions

Data reconciliation is an important step in real time on-line optimization of a plant. It adjusts the process data to satisfy the constraints of the system model and provides estimates for unmeasured variables and process parameters, which are used in the consecutive economic optimization step. In this study, the focus has been on the simultaneous data reconciliation and gross error detection strategies to improve this initial step in on-line optimization. To this end, six different methods derived from robust statistics have been investigated along with weighted least squares and a modified version of MIMT for nonlinear models. Unlike previous studies, special attention was given to the concept of tuning the ρ functions to obtain the same efficiency at the ideal condition. This proves necessary for a comparative study of different methods. This tuning inevitably affects the relative shape of the influence functions for the ρ functions. Using these individual influence functions, several cut points can be defined as prospective gross error detection criterion. Even for ρ functions which do not have cut points, such as Fair function and ρ function of Logistic distribution, similar gross error detection criteria can be proposed borrowing the cut points of other ρ functions. Moreover, an outlier rejection rule (X84) adopted from univariate robust estimation and depending solely on the observation of the residuals in the data reconciliation solution, is introduced.

The evaluation of the performance of a total of eight methods is undertaken using five small-scale examples from the literature and two cases involving industrial plants with real process data. The Monte Carlo study shows that the robust approaches for the simultaneous data reconciliation and gross error detection of chemical processes can provide similar or better results compared to a sequential method, with a single (two for Hampel's redescending M-estimator) solution of the NLP.

A box-plot of the six observations for the average and median total error reduction values (Fig. 11) reveals that on the average one can expect mean TER values between 0.4 and 0.8 with similar variability among different methods. The median TER values for MIMT and ρ function of the Cauchy distribution are above 0.6 for nearly half of the cases, proving them as good data reconciliation methods.

The overall power of the first and second gross error detection criterion was higher than 0.7 for half of the cases investigated (Fig. 12). Hampel's redescending M-estimator, ρ functions for Cauchy and Logistic distributions along with the modified MIMT method achieved this performance with lower variability in the average number of Type I errors (Fig. 13). In general, the second and third gross error detection criteria reduced the average number of wrong identifications.

The guidelines compiled from the results of six different examples facilitate an intelligent selection of the ρ function and gross error detection criterion by comparing the median and variability of each method. Future work will expand the current analysis to include the determination of the gross

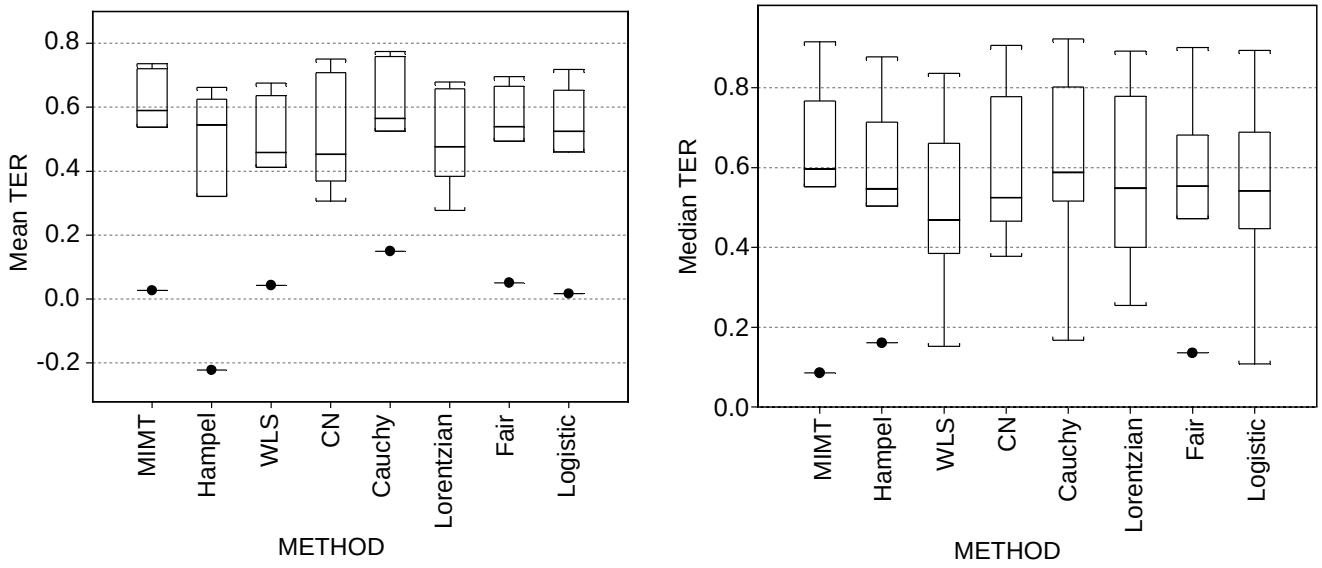

Fig. 11. Box-plots showing mean and median total error reduction values of different methods for the six examples ((•) outlier).

Fig. 12. Box-plots showing overall power values of different methods for the six examples ((•) outlier).

Fig. 13. Box-plots showing average Type I error values of different methods for the six examples.

error detection criteria based on estimation efficiency and the breakdown point analysis for different data reconciliation methods. Moreover, the methodologies presented herein will be implemented in a future version of the Advanced Process Analysis System (APAS), a tool to perform comprehensive and in-depth evaluations of economic, environmental, safety and hazard analysis projects (Telang et al., 1999).

Acknowledgements

The authors would like to thank Professor Frank Hampel for invaluable discussions/references on robust estimation and efficiency, Professor Luis Escobar for discussions on nonlinear multivariate regression and on the derivation of the influence function, and Professor Chris Swartz for providing a copy of his paper. Thomas Hertwig and Michael Rich are also acknowledged for their help as experts in sulfuric acid and alkylation processes, respectively.

Appendix A

For a set of one-dimensional observations $X_1, \ldots, X_n$ which are independent and identically distributed, a maximum likelihood type estimate $T_n = T_n(X_1, \ldots, X_n)$ is defined (Hampel et al., 1986; Huber, 1981; Rey, 1983) by a minimization problem

$$\min_{T_n} \sum_{i=1}^{n} \rho(X_i; T_n) \tag{A.1}$$

As an example, assume that $X_1, \ldots, X_n$ are observations from Normal distribution with the probability density function $\phi(x) = (2\pi)^{-1/2} \exp(-1/2(x-\theta)^2)$, i.e. variance of 1 and mean θ. The arithmetic mean of the observations ($T_n = (1/n) \sum_{i=1}^{n} X_i$) is a well-known estimate for the population mean, and is defined by the following

$$\min_{T_n} \sum_{i=1}^{n} -\ln\left[(2\pi)^{-1/2} \exp\left(-\frac{1}{2}(X_i - T_n)^2\right)\right] = \min_{T_n} \sum_{i=1}^{n} \frac{1}{2}(X_i - T_n)^2 \tag{A.2}$$

or equivalently

$$\sum_{i=1}^{n} (X_i - T_n) = 0 \Leftrightarrow T_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{A.3}$$

Similar to the above example, if the first derivative of $\rho(X_i, T_n)$ exists, T_n can be defined by the implicit equation

$$\sum_{i=1}^{n} \psi(X_i; T_n) = 0 \tag{A.4}$$

where

$$\frac{\partial}{\partial T_n} \rho(X_i; T_n) = \psi(X_i; T_n) \tag{A.5}$$

If F_n is the empirical cumulative distribution function generated by the observations $X_1, \ldots, X_n$, then T_n in (A.4) can also be written as $T(F_n)$, where T is the following functional

$$\int \psi(x; T(F))\, \mathrm{d}F = 0 \tag{A.6}$$

for all distributions F for which the integral is defined (Hampel et al., 1986).

To evaluate the influence function of an M-estimate, replace F with $F_t = (1-t)F + t\Delta_x$ in (A.6) and take the derivative with respect to t at $t = 0$, since the influence function IF is defined as

$$\mathrm{IF}(x, T, F) = \lim_{t \to 0} \frac{T(F_t) - T(F)}{t} = \frac{\partial}{\partial t}[T(F_t)]|_{t=0} \tag{A.7}$$

Here, Δ_x is the probability measure, which puts mass 1 at the point x.

$$\frac{\partial}{\partial t} \int \psi(x, T((1-t)F + t\Delta_x))\, \mathrm{d}[(1-t)F + t\Delta_x]|_{t=0} = 0 \tag{A.8}$$

Changing the order of the integration and differentiation, gives

$$\int \psi(x, T((1-t)F + t\Delta_x))\, \mathrm{d}(\Delta_x - F)|_{t=0} + \int \frac{\partial}{\partial T(F_t)} \psi(x, T((1-t)F + t\Delta_x)) \times |_{t=0} \frac{\partial}{\partial t}[T(F_t)]|_{t=0} \mathrm{d}F_t|_{t=0} = 0 \tag{A.9}$$

Simplifying gives

$$\int \psi(x, T(F))\, \mathrm{d}(\Delta_x - F) + \frac{\partial}{\partial t}[T(F_t)]|_{t=0} \times \int \frac{\partial}{\partial T(F)} \psi(x, T(F))\, \mathrm{d}F = 0 \tag{A.10}$$

Making use of (A.6) and (A.7), gives

$$\mathrm{IF}(x, T, F) = \frac{\psi(x, T(F))}{-\int (\partial/(\partial T(F)))\psi(x, T(F))\, \mathrm{d}F} \tag{A.11}$$

provided that the denominator is nonzero. Therefore, the influence function IF(x, T, F) is proportional to $\psi(x, T(F))$, i.e.

$$\mathrm{IF}(x, T, F) \propto \frac{\partial}{\partial T_n} \rho(X_i, T_n) \tag{A.12}$$

References

Albuquerque, J. S., & Biegler, L. T. (1996). Data reconciliation and gross error detection for dynamic system. *AIChE Journal, 42*(10), 2841–2856.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location: Survey and advances*. Princeton, NJ: Princeton University Press.

Arora, N., & Biegler, L. T. (2001). Redescending estimators for data reconciliation and parameter estimation. *Computers and Chemical Engineering, 25*, 1585–1599.

Brooke, A., Kendrick, D., & Meeraus, A. (1992). *Release 2.25: GAMS: A user's guide*. Danvers, MA: Boyd & Fraser Publishing Co.

Chen, X. (1998). The optimal implementation of on-line optimization for chemical and refinery processes, Ph.D. Dissertation. Baton Rouge, LA 70803: Louisiana State University.

Chen, X., Pike, R. W., Hertwig, T. A., & Hopper, J. R. (1998). Optimal implementation of on-line optimization. *Computers and Chemical Engineering, 22*(Suppl.), S435–S442.

Crowe, C. M. (1986). Reconciliation of process flow rates by matrix projection. Part II. Nonlinear case. *AIChE Journal, 32*(4), 616–623.

Crowe, C. M., Garcia Campos, Y. A., & Hyrmak, A. (1983). Reconciliation of process flow rates by matrix projection. Part I. Linear case. *AIChE Journal, 29*(6), 881–888.

Deutsch, R. (1965). *Estimation theory*. Englewood Cliffs, NJ: Prentice-Hall.

Fair, R. C. (1974). On the robust estimation of econometric models. *Annals of Economic and Social Measurement, 3*, 667–677.

Farris, R. H., & Law, V. J. (1979). An efficient computational technique for generalized application of maximum likelihood to improve correlation of experimental data. *Computers and Chemical Engineering, 3*, 95–104.

Hampel, F. R. (1985). The breakdown points of the mean combined with some rejection rules. *Technometrics, 27*(2), 95–107.

Hampel, F. R. (2002). Personal communication.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics—the approach based on influence functions*. New York: Wiley.

Harris, J. L., & Norman, J. R. (1972). Temperature-dependent kinetic equation for catalytic oxidation of sulfur dioxide. *Industrial Engineering, Chemical Processing and Design Development, 11*(4), 564.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Johnston, L. P. M., & Kramer, M. A. (1995). Maximum likelihood data rectification: Steady-state systems. *AIChE Journal, 41*(11), 2415–2426.

Jordache, C., Ternet, D., & Brown, S. (2001). Efficient gross error elimination methods for rigorous on-line optimization. In R. Gani, & B. Jorgensen (Eds.), *Computer-aided chemical engineering series* (Vol. 9, pp. 675–680). Amsterdam, The Netherlands: Elsevier.

Kalvelagen, E. (2002). External functions in GAMS, http://www.gams.com/docs/extfunc.htm, last accessed in June 2003.

Kim, I.-W., Kang, M. S., Park, S., & Edgar, T. F. (1997). Robust data reconciliation and gross error detection: the modified MIMT using NLP. *Computers and Chemical Engineering, 21*(7), 775–782.

Kuehn, D. R., & Davidson, H. (1961). Computer control. II. Mathematics of control. *Chemical Engineering Progress, 57*, 44.

Knepper, J. C., & Gorman, J. W. (1980). Statistical analysis of constrained data sets. *AIChE Journal, 26*, 260–264.

Liebman, M. J., Edgar, T. F. (1988). Data reconciliation for nonlinear process. In *Proceedings of the paper presented at the AIChE annual meeting*, Washington, DC.

Narasimhan, S., & Jordache, C. (2000). *Data reconciliation and gross error detection: An intelligent use of process data*. Houston, TX: Gulf Publishing Company.

Özyurt, D. B., Pike, R. W., Hopper, J. R., Punuru, J. R., & Yaws, C. L. (2001). *Advanced process analysis system: User's manual and tutorial for the alkylation process*. Baton Rouge, LA: Mineral Processing Institute, Louisiana State University. URL: http://www.mpri.lsu.edu/Manuals/AlkylationManual.pdf.

Pai, D. C. C., & Fisher, G. D. (1988). Application of Broyden's method to reconciliation of nonlinear constrained data. *AIChE Journal, 34*, 873–876.

Rey, J. J. W. (1983). *Introduction to robust and quasi-robust statistical methods*. Berlin, Germany: Springer-Verlag.

Rich, M. K., McGee, D., Hopper, J. R., Yaws, C. L., Özyurt, D. B., & Pike, R. W. (2001). Advanced process analysis for source reduction in the sulfuric acid petroleum alkylation process. Final Technical Report submitted to US Department of Energy. NICE 3 Grant DE-FG48-96R8-10598.

Richard, M. J. (1987). An evaluation of applicability of nonlinear programming algorithms to a typical commercial process flowsheeting simulator, Ph.D. Dissertation. Baton Rouge, LA: Louisiana State University.

Ripps, D. L. (1965). Adjustment of experimental data. *Chemical Engineering Progress Symposium Series, 61*, 8–13.

Romagnoli, J. A., & Sanchez, M. C. (2000). Data processing and reconciliation for chemical process operations. In *Process systems engineering series* (Vol. 2). San Diego, CA: Academic Press.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York, NY: Wiley.

Sanchez, M., Sentoni, G., Schbib, S., Tonelli, S., & Romagnoli, J. (1996). Gross measurements error detection/identification for an industrial ethylene reactor. *Computers and Chemical Engineering, 20*(Suppl.), S1559–S1564.

Swartz, C. L. E. (1989). Data reconciliation for generalized flowsheet applications. In *Proceedings of the paper presented at American Chemical Society National Meeting*, Dallas, TX.

Serth, R. W., & Heenan, W. A. (1986). Gross error detection and data reconciliation in steam-metering systems. *AIChE Journal, 32*, 733–742.

Serth, R. W., Valero, C. M., & Heenan, W. A. (1987). Detection of gross errors in nonlinearly constrained data: A case study. *Chemical Engineering Communications, 51*, 89–104.

Telang, K. S., Pike, R. W., Knopf, F. C., Hopper, J. R., Saleh, J., Waghchoure, S., Hedge, S. C., & Hertwig, T. A. (1999). An advanced process analysis system for improving chemical and refinery processes. *Computers and Chemical Engineering, 23*(Suppl.), S727–S730.

Tjoa, I. B., & Biegler, L. T. (1991). Simultaneous strategies for data reconciliation and gross error detection of nonlinear systems. *Computers and Chemical Engineering, 15*(10), 679–690.

Veverka, V. V., & Madron, F. (1997). Material and energy balancing in the process industries: From microscopic balances to large plants. In *Computer-aided chemical engineering series* (Vol. 7). Amsterdam, The Netherlands: Elsevier.

Vichailak, M. (1995). Pollution prevention by process modification. D.E. Field Study. Beaumont, TX: Lamar University.

Weiss, G. H., Romagnoli, J. A., & Islam, K. A. (1996). Data reconciliation—an industrial case study. *Computers and Chemical Engineering, 20*(12), 1441–1449.