Gross Error Detection and Data Reconciliation in Steam-Metering Systems

Several new algorithms for the detection of gross errors in process data are presented and applied to an industrial steam-metering system by means of computer simulation. A number of algorithms that have appeared in the literature are also applied to the steam-metering system, and the performances of the various algorithms are compared.

R. W. Serth and W. A. Heenan
Department of Chemical and
Natural Gas Engineering
Texas A&I University
Kingsville, TX 78363

SCOPE

Steam-metering systems of industrial chemical processes constitute an important potential area of application for error detection and data reconciliation techniques. These systems comprise mass-flow networks characterized by moderate size, relatively complex topography, and a range of flow rates that typically covers two orders of magnitude or more. As such, they represent basic, but nontrivial, applications for error detection algorithms.

In this paper, a number of techniques for gross error detection and data reconciliation are applied to a typical industrial steam-metering system by means of computer simulation. The methods tested include the measurement test (MT) method (Mah and Tamhane, 1982) the method of pseudonodes (MP) (Mah et al., 1976), several modifications of these methods, and a new method, the screened combinatorial (SC) method. The results represent the first direct comparison of the performances of different gross error detection algorithms on a system of practical interest.

CONCLUSIONS AND SIGNIFICANCE

Based on the results of this study, the SC and modified iterative measurement test (MIMT) algorithms constitute effective and reliable methods for gross error detection and data reconciliation in steam-metering systems. Both algorithms detected approximately 80% of the gross errors in the data and achieved average total error reductions of over 60%. However, the MIMT method is computationally more efficient, requiring much less computing time, on the average, than the SC method. While both algorithms are applicable to linear systems in general, the MIMT method should be particularly attractive for problems involving very large systems since efficient matrix techniques are available for performing the iterative calculations required in the algorithm (Ripps, 1965; Romagnoli and Stephanopoulos, 1981; Wang and Stephanopoulos, 1983).

The MT algorithm performed very poorly in this application, for reasons partly related to the system structure and the range of parameters employed. These results are in accord with the recent study of the measurement test by Iordache et al. (1985). The MP algorithm also performed very poorly, and the results demonstrate that error cancellation in the aggregation process can be a serious drawback to this method. Finally, the success of the SC method in this study suggests that a useful approach to gross error detection may be through the combination of different algorithms so as to exploit the strengths of each. Such an approach has previously been advocated by Crowe et al. (1983).

Introduction

Accurate and reliable energy accounting in chemical plants is important for process monitoring as well as for decision-making regarding the implementation and effectiveness of energy conservation measures. Most energy accounting is based on steam-metering systems in various plant operating units. Measured steam flow rates are subject to random measurement errors and also to systematic errors that result from instrument bias or failure. As a result of these errors, the measurements will generally be inconsistent with material balance requirements. Furthermore, the data may present a confusing or misleading picture of energy use patterns in the plant. Of particular concern are large (gross) errors, which may greatly distort the energy use pattern. Such errors will usually be systematic errors resulting from instrument malfunction.

What is needed, then, is a method that will identify faulty steam meters by detecting gross errors in a set of data, and that furthermore will generate an adjusted data set that satisfies the material balance requirements. This problem is a special case of the general problem of error detection and process data reconciliation, a topic that has received considerable attention in the recent literature. Reconciliation of process data subject to iinear material balance constraints and containing only random errors can be achieved by means of a constrained least-squares procedure. This technique was first employed by Kuehn and Davidson (1961) and has since been utilized by numerous workers in the field. However, as pointed out by Ripps (1965), the presence of gross errors in the data can vitiate the least-squares procedure. Hence, it is necessary to identify and eliminate measurements containing gross errors before proceeding with data reconciliation.

A number of methods for gross error detection have been developed, most of which involve the use of statistical tests based on the assumption that the random errors in the data are normally distributed. In one of the simplest methods, the set of residuals from the least-squares procedure is tested for outliers. Any measurement for which the residual fails the test is considered to contain a gross error. This method has been advocated by several investigators, including Mah and Tamhane (1982), Crowe et al. (1983), and Stephenson and Shewchuck (1984). A related method, in which the largest residual is used to identify a single gross error in a data set containing only one gross error, was developed by Almasy and Sztano (1975). A chi-square test to detect the presence of gross error in a data set was established by Reilly and Carpani (1963) and was used by Madron et al. (1977), Crowe et al. (1983), and Romagnoli and Stephanopoulous (1980, 1981). The latter authors incorporated the test into a sequential analysis of the material balance equations that identifies the measurements containing the gross errors. This method was subsequently employed by Wang and Stephanopoulous (1983). Ripps (1965) used a gross error identification scheme in which each measurement is in turn deleted from the data set. The deletion that minimizes the least-squares objective function is identified with the gross error. It was shown by Crowe et al. (1983) that this serial elimination procedure is equivalent to minimizing a chi-square statistic. This method was extended to data sets containing more than one gross error by Nogita (1972). However, his iterative method is based on a test statistic that permits cancellation of errors, as shown by Mah et al. (1976). The latter authors also developed a gross error identification scheme based on a graph-theoretical analysis of material balance networks. Their method utilizes a statistical test for the imbalance in individual material balance equations and combinations of balance equations.

In this paper, a number of algorithms for gross error detection, some of which are new, are applied to a steam-metering system, and their respective performances are compared by means of computer simulation.

Problem Definition

The steam-metering system for a methanol synthesis unit of a large chemical plant was selected as a typical process steam system. Many industrial steam systems are very similar in both size and structure to this system. A network representation of the system consisting of 12 nodes and 28 streams is shown in Figure 1. The environmental node (node 12), which represents the overall material balance for the process, has been omitted from the diagram for the sake of clarity. Where necessary, streams have been aggregated so that each pair of nodes is connected by a single stream. The correct values of the steam flow rates are listed in Table 1. The flow rates are time-averaged values (over an 8 h shift), and represent actual operating data except that the values have been adjusted to balance the system. These values were used without further scaling in the computer simulation experiments described below. Numerical calculations show that the results presented in this paper are insensitive to scaling over a range of scale factors from 10^{-3} to 10^3.

Steam networks represent the simplest type of process network in that only material flows of a single component are involved. Thus, the balance equations are all linear and only flow rate measurements need be considered. In addition, the system size (for an individual process) is generally modest. On the other hand, the network topography is relatively complex and the flow rates vary over a wide range, covering more than two orders of magnitude. Hence, such a system constitutes a nontrivial application for error detection and data reconciliation algorithms. The material balance equations for the steam-metering system can be written in matrix form as

$$\tilde{A}x = 0 \tag{1}$$

where x is the vector of flow rates and the incidence matrix, $\tilde{A}$, is given in Figure 2. We denote by A the incidence matrix for the system excluding node 12, that is, A is obtained from $\tilde{A}$ by deleting the last row. If y represents the vector of measured flow rates and x is the vector of true flow rates, then in the absence of systematic errors

$$y = x + \epsilon \tag{2}$$

where ϵ is the vector of random measurement errors. It is assumed that the ϵ_j are normally distributed with zero means and known covariance matrix, Q. In this work we also assume that the measurements are independent, so that Q is a diagonal matrix. We further assume that the flow rates of all streams are measured. If this is not the case, a problem of lower dimension can be obtained by nodal aggregation (Mah et al., 1976) or other techniques (Romagnoli and Stephanopoulos, 1980). Hence, this assumption entails no loss of generality.

The data reconciliation problem involves finding a set of adjustments to the measured flow rates such that the adjusted values satisfy Eq. 1. Denoting the vector of adjusted flow rates by $\hat{x}$ and the vector of adjustments by a, we have

$$\hat{x} = y + a \tag{3}$$

The constrained least-squares problem can then be stated as

$$\min_{a} a^T Q^{-1} a \quad \text{subject to } A\hat{x} = 0 \tag{4}$$

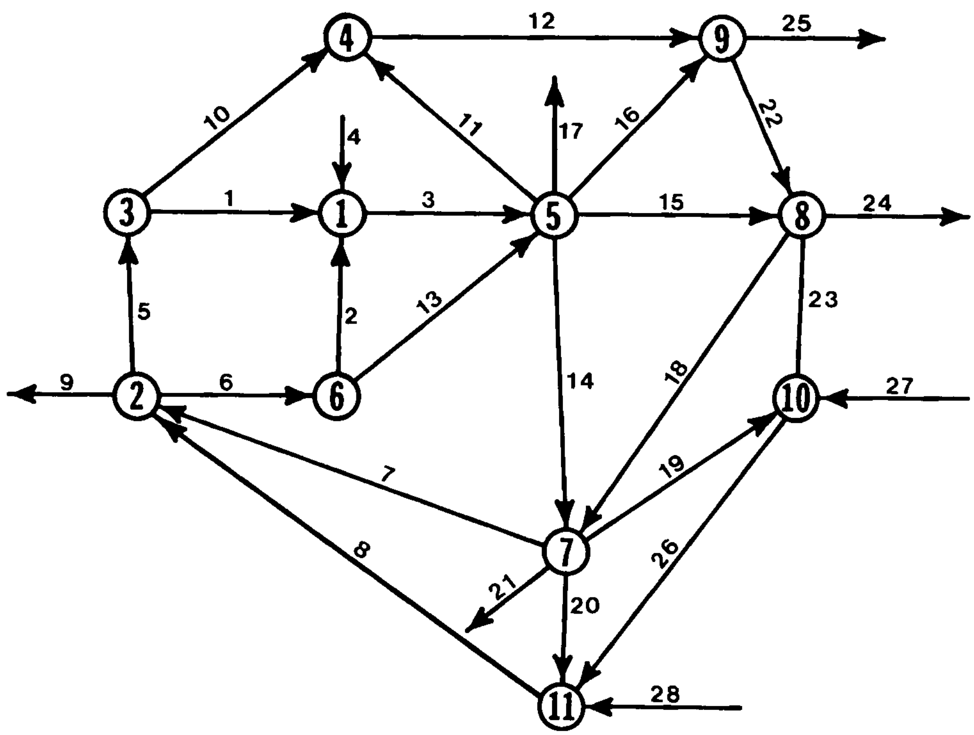

Figure 1. Network representation of process steam system for a methanol synthesis unit.

As noted by Himmelblau (1984), the additional constraints that the adjusted flow rates be nonnegative are generally not explicitly imposed on the least-squares problem. Imposing the inequality constraints would invalidate the simple solution given below, as well as the statistical analysis that is based on that solution. As a result, it is possible for the data reconciliation procedure to generate negative flow rates. This point will be discussed further in the following sections.

The solution $\hat{x}^*$ to the above problem, which can be obtained by the method of Lagrange multipliers (Kuehn and Davidson, 1961), is

$$\hat{x}^* = y - QA^T(AQA^T)^{-1}Ay \tag{5}$$

The uniqueness of this solution follows from the linearity of the normal equations in conjunction with the Kuhn-Tucker Theorem (Hadley, 1964). The vector e of residuals is given by

$$e = y - \hat{x}^* = QA^T(AQA^T)^{-1}Ay \tag{6}$$

For the purpose of the computer simulation experiments, the value of the standard deviation σ_j of the measurement error for stream j was taken as 2.5% of the corresponding true flow rate listed in Table 1. The covariance matrix was then given by

$$Q = \mathrm{Diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_{28}^2) \tag{7}$$

Node/Stream	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	1	1	-1	1																								
2					-1	-1	1	1	-1																			
3	-1				1					-1																		
4										1	-1	-1																
5			1								-1		1	-1	-1	-1	-1											
6		-1				1							-1															
7							-1							1				1	-1	-1	-1							
8															1			-1				1	-1	-1				
9												1				1						-1			-1			
10																			1				1			-1	-1	
11								-1												1						1		1
12				-1					1								1				1			1	1		-1	-1

Figure 2. Incidence matrix for the process steam system of Figure 1.

Table 1. Correct Values of Flow Rates for the Steam System of Figure 1

Stream No.	Flow Rate 1,000 kg/h
1	0.86
2	1.00
3	111.82
4	109.95
5	53.27
6	112.27
7	2.32
8	164.05
9	0.86
10	52.41
11	14.86
12	67.27
13	111.27
14	91.86
15	60.00
16	23.64
17	32.73
18	16.23
19	7.95
20	10.50
21	87.27
22	5.45
23	2.59
24	46.64
25	85.45
26	81.32
27	70.77
28	72.23

For each simulation run, a measurement vector, y, was constructed as

$$y = x + \epsilon + \delta \tag{8}$$

where δ is the vector of systematic errors. A Gaussian pseudorandom number generator (IBM 360 SSP subroutine GAUSS) was used to generate the ϵ_j. A uniform pseudorandom number generator (IBM 360 SSP subroutine RANDU) was then used to define the number, position, magnitude, and algebraic sign of the nonzero systematic errors. The number of nonzero systematic errors was allowed to vary between one and seven (corresponding to 25% faulty meters). The magnitudes of the nonzero systematic errors were constrained by

$$0.1\,(x_j + \epsilon_j) \leq |\delta_j| \leq x_j + \epsilon_j \tag{9}$$

That is, the magnitudes of the systematic errors were allowed to vary between approximately 10% and approximately 100% of the respective true flow rates. Note that according to Eq. 9, the smallest measured value permitted was zero, corresponding to total meter failure. No restrictions were placed on the positions or algebraic signs of the systematic errors.

Algorithms based on a test of residuals

Three algorithms for gross error detection based on a statistical test for outliers of the least-squares residuals were studied in this work. The statistical test itself has been termed the measurement test by Iordache et al. (1985), who evaluated its characteristics when applied to data sets containing a single gross error.

Algorithm 1. Measurement test (MT) method

The algorithm, as given here, parallels the one outlined by Mah and Tamhane (1982). We note, however, that their equations simplify somewhat in the present application.

Step 1. Apply the least-squares routine to the full system by using Eqs. 5 and 6 to compute $\hat{\boldsymbol{x}}^*$ and $\boldsymbol{e}$.

Step 2. For each stream, j, compute the quantity

$$z_j = e_j/\sqrt{v_{jj}} \tag{10}$$

where

$$V = QA^T(AQA^T)^{-1}AQ \tag{11}$$

Under the null hypothesis that the measured value for stream j contains no systematic error, z_j is a standard normal deviate.

Step 3. Compare each z_j with a critical test value z_c. If $|z_j| > z_c$, denote stream j as a bad stream. The critical value recommended by Mah and Tamhane (1982) is $z_c = z_{1-\beta/2}$, the $1 - \beta/2$ point of the standard normal distribution. Here,

$$\beta = 1 - (1 - \alpha)^{1/n} \tag{12}$$

where n is the number of measurements tested ($n = 28$ in this case), α is the overall probability of a type I error for all n tests, and β is the probability of a type I error for each individual test. For $\alpha = 0.05$, we have $\beta = 0.0018$ and $z_{1-\beta/2} = 3.1165$. Denote by S the set of bad streams found by the above procedure.

Step 4. The measurements $y_j, j \in S$ are considered to contain systematic errors. If S is empty, proceed to step 7. Otherwise, remove the streams contained in S from the system by nodal aggregation. This process yields a system of lower dimension with compressed incidence matrix $\boldsymbol{B}$, measurement vector $\boldsymbol{w}$, and covariance matrix $\boldsymbol{P}$. Note that some good streams may also be removed in the process if they occur in loops with bad streams. Denote by T the set of streams whose measured flow rates are contained in $\boldsymbol{w}$.

Step 5. Compute the least-squares estimates of the flow rates for the streams in T by applying Eq. 5 with $\boldsymbol{A}$, $\boldsymbol{y}$, and $\boldsymbol{Q}$ replaced by $\boldsymbol{B}$, $\boldsymbol{w}$, and $\boldsymbol{P}$, respectively.

Step 6. Compute corrected values for the flow rates of streams in S by solving Eq. 1. In this calculation, the flow rates computed in step 5 are used for the streams in T, and the original measured flow rates are used for streams in the set $R = U - (S \cup T)$, where U is the set of all streams in the system.

Step 7. The vector, $\hat{\boldsymbol{y}}$, of reconciled flow rates is obtained from the values computed in step 6 for the streams in S, together with the least-squares estimates from step 5 for the streams in T and the original measured flow rates for the streams in R. If S is empty, then $\hat{\boldsymbol{y}} = \hat{\boldsymbol{x}}^*$, the vector of least-squares estimates computed in step 1.

As noted by Mah and Tamhane, setting n in Eq. 12 equal to the number of measurements tested provides a conservative test for outliers since the residuals are generally not independent. In the present application, for example, there are only eleven degrees of freedom, corresponding to the eleven independent material balance constraints. It should also be noted that Eq. 12

is not employed by all workers (see, e.g., Crowe et al., 1983). The algorithm will generate the same results whether or not Eq. 12 is used, but the results will be obtained at different nominal significance levels. Since the significance level may be considered an adjustable parameter in an error detection algorithm, however, the nominal value assigned to it is of little consequence. The use of Eq. 12 does have some effect on the iterative algorithms presented below, since the number of measurements tested changes at each iteration. However, the effect is quite small in the present application.

Problems may be encountered in step 6 of the above algorithm. For instance, the set S of bad streams found in step 3 may be unobservable (Stanley and Mah, 1981a,b). This will happen when S includes a perturbation cycle, i.e., all the streams in a loop, such as streams 15, 16, and 22 in Figure 1. In this case, a singular matrix will be encountered in step 6. Even if S is observable, the values computed in step 6 may not be feasible. For example, one or more of the computed flow rates may be negative. These points will be discussed further in the following sections.

The main problem with the MT method is that the least-squares procedure tends to spread the gross errors over all the measurements, thereby creating large residuals corresponding to good measurements. When these residuals fail the test for outliers, the corresponding measurements are erroneously identified as containing gross errors. This problem can be ameliorated by applying the measurement test in an iterative fashion. At each iteration, only the measurement corresponding to the largest normalized residual is discarded, and the process terminates when all remaining measurements satisfy the measurement test. This procedure is embodied in the following algorithm.

Algorithm 2. Iterative measurement test (IMT) method

Step 1. Compute the vectors $\hat{x}^*$ and e as in algorithm 1.

Step 2. Same as algorithm 1.

Step 3. Compare each z_j from step 2 with the critical test value $z_c = z_{1-\beta/2}$ as in algorithm 1. If $|z_j| \le z_c$ for all streams j, proceed to step 5. Otherwise, select the stream corresponding to the largest value of $|z_j|$ and add it to the set S of bad streams (initially, S is empty). If two or more streams have the same maximum value of $|z_j|$, select the one with the lowest index, j.

Step 4. Remove the streams contained in S from the system by nodal aggregation to obtain a system of lower dimension with compressed incidence matrix $\boldsymbol{B}$, measurement vector $\boldsymbol{w}$, and covariance matrix $\boldsymbol{P}$. Return to step 1 with $\boldsymbol{A}$, $\boldsymbol{y}$, and $\boldsymbol{Q}$ replaced by $\boldsymbol{B}$, $\boldsymbol{w}$, and $\boldsymbol{P}$, respectively.

Step 5. The measurements y_j, $j \in S$ are considered to contain systematic errors. If S is empty, proceed to step 6. Otherwise, compute corrected values for the flow rates of streams contained in S by solving Eq. 1. In this calculation, the least-squares estimates of the flow rates computed in step 1 on the final iteration are used for the streams in T and the original measured flow rates are used for the streams in R, where T and R have the same meaning as in algorithm 1.

Step 6. The vector $\hat{y}$ of reconciled flow rates is obtained from the values computed in step 5 for the streams contained in S, together with the least-squares estimates for the streams contained in T and the original measured values for the streams contained in R. If S is empty, then $\hat{y} = \hat{x}^*$, the vector of least-squares estimates computed in step 1.

It should be noted that the above algorithm is not the same as the serial elimination technique introduced by Ripps (1965). In the latter method, each suspect measurement is deleted in turn, and the least-squares calculation repeated each time. If more than one gross error is present, the entire procedure must be repeated with combinations of two, three, etc., streams until a combination is found that, when deleted, results in the remaining data satisfying the measurement test. Alternatively, the deletion that minimizes a chi-square statistic may be used to identify the bad measurements at each stage of the calculation. Neither procedure will, in general, identify a unique set of bad measurements when the data contain more than one gross error.

In the IMT method, only the measurement corresponding to the largest normalized residual is deleted at each stage, and it is automatically identified as containing a gross error. The least-squares calculation is thus made only once at each stage of the procedure, and the set S of bad streams is uniquely determined. In addition, S is always observable, so corrected flow rates for the bad streams can always be computed in step 5 of the algorithm.

The IMT method retains one drawback of the MT method, namely, that the set of reconciled flow rates may contain negative values or absurdly large values. This situation generally indicates a failure of the algorithm to correctly identify all of the gross errors in the data. In order to eliminate this problem, the algorithm was modified in such a way that unreasonable values for the flow rates are not permitted. This result is accomplished by allowing a stream to be added to S only if its inclusion results in reconciled flow rates that lie within specified bounds.

Algorithm 3. Modified iterative measurement test (MIMT) method

Step 0. Set the iteration counter k to one.

Step 1. Same as step 1 of algorithm 2. In addition, set $\tilde{y}_o = \hat{x}^*$.

Step 2. Same as step 2 of algorithm 2.

Step 3. Select the stream corresponding to the largest value of $|z_j| > z_c$ and add it to the set S (which is initially empty) as in step 3 of algorithm 2. If $|z_j| \le z_c$ for all streams j, proceed to step 8.

Step 4. Same as step 4 of algorithm 2.

Step 5. Compute new values of $\hat{x}^*$ and e using Eqs. 5 and 6 with $\boldsymbol{A}$, $\boldsymbol{y}$ and $\boldsymbol{Q}$ replaced by $\boldsymbol{B}$, $\boldsymbol{w}$, and $\boldsymbol{P}$, respectively.

Step 6. Compute corrected values for the flow rates of the streams in S by solving Eq. 1. In this calculation, the least-squares estimates of the flow rates computed in step 5 are used for the streams in T and the original measured flow rates are used for the streams in R, where the sets T and R are defined as in algorithm 1. Then construct the vector $\tilde{y}_k$ of corrected flow rates as in step 6 of algorithm 2.

Step 7. Determine whether the following inequalities are satisfied:

$$xl_j \le \tilde{y}_{kj} \le xu_j \quad [j \in (U - R)] \tag{13}$$

where $\boldsymbol{xl}$ and $\boldsymbol{xu}$ are vectors of specified lower and upper bounds on the flow rates. If all inequalities are satisfied, increment k and return to step 2 with $\boldsymbol{A}$ and $\boldsymbol{Q}$ replaced by $\boldsymbol{B}$ and $\boldsymbol{P}$, respectively. Otherwise, delete the last entry in S, replace it with the stream corresponding to the next largest value of $|z_j| > z_c$, and return to step 4. If $|z_j| \le z_c$ for all remaining streams, delete the last entry in S and proceed to step 8.

Step 8. The measurements $y_j, j \in S$ are considered to contain systematic errors. The vector $\hat{y}$ of reconciled flow rates is given by $\hat{y} = \tilde{y}_{k-1}$.

The lower and upper bounds on the flow rates in step 7 can generally be set at zero and full-scale meter readings, respectively. However, knowledge of process conditions will usually permit sharper bounds to be specified. For the purpose of the computer simulation experiments, bounds were conservatively specified as zero for the lower bounds and four times the correct flow rates listed in Table 1 for the upper bounds.

Algorithms Based on a Nodal Imbalance Test

Four algorithms for gross error detection based on a statistical test of nodal imbalances were studied in this work. The first two methods are based on the work of Mah et al. (1976), while the last two are new.

Algorithm 4. Method of pseudonodes (MP)

In this algorithm, the nodal imbalance test is applied to each node and also to aggregates of two or more nodes, which are called pseudonodes. The principal assumption underlying the method is that errors in two or more measurements do not cancel.

Step 1. Compute a vector r of nodal imbalances and a vector z of test values by

$$r = \tilde{A}y \tag{14}$$

$$z_i = r_i / \sqrt{g_{ii}} \tag{15}$$

where

$$G = \tilde{A}Q\tilde{A}^T \tag{16}$$

Under the null hypothesis that no systematic errors are present in the measurements, the z_i are standard normal deviates.

Step 2. Compare each z_i with a critical test value $z_c = z_{1-\alpha/2}$. If $|z_i| \le z_c$, denote node i as a good node and denote all streams adjacent to node i as good streams.

We note that for a test at the 95% significance level, $\alpha = 0.05$ and therefore $z_c = 1.96$.

Step 3. If no bad nodes were detected in step 2, proceed to step 5. Otherwise, repeat steps 1 and 2 (with the appropriate changes in the matrices and vectors) for pseudonodes containing 2, 3, . . . , m nodes. In the present application, m was set equal to four since no improvement was observed when aggregates of more than four nodes were used.

Step 4. Let S be the set of all streams not denoted as good in the previous steps. The measurements $y_j, j \in S$ are considered to contain systematic errors.

Steps 5–8. The data reconciliation procedure is the same as in steps 4–7 of algorithm 1.

Some of the streams in S can be positively identified as bad streams by applying graph-theoretical rules given by Mah et al. (1976). The remaining streams cannot be so identified, and thus are only potentially bad streams. The above algorithm does not make this distinction, but rather treats all of the streams in S as bad streams. This procedure has two obvious consequences. First, the potentially bad streams that are in fact bad streams, are correctly identified as such, thereby improving algorithm performance. Second, the potentially bad streams that are in fact good streams, are erroneously identified as containing gross errors. However, it was found that the algorithm made very few erroneous identifications. Therefore, in the present application there is little to be gained by additional identification procedures. Furthermore, in order to proceed with data reconciliation it is necessary to treat the potentially bad streams either as good or as bad streams. We believe that it is preferable in this context to err on the side of caution by assuming that potentially bad data are in fact bad.

It will be noted that Eq. 12 is not used to control the type I error in the nodal imbalance test. Although this could be done, the rationale would be questionable since the probability of a type I error in the nodal imbalance test is not equal to the probability of rejecting a good measurement as it is in the measurement test. Equation 12 has not been used in conjunction with the nodal imbalance test by previous investigators.

Algorithm 5. Modified method of pseudonodes (MMP)

In the previous algorithm, it may happen that the set, S, obtained in step 4 is empty, but one or more nodes are bad. This condition is considered to be the result of leaks at the bad nodes (Mah et al., 1976). In practice, however, this condition frequently arises from cancellation of errors during the aggregation process. The latter cause is assumed in the present algorithm.

Step 1. Same as algorithm 4. In addition, set $mk = m$.

Steps 2 and 3. Same as algorithm 4. In addition, let $\tilde{S}_k$ be the set of all streams not denoted as good in stages 1 through k of the aggregation procedure. If at any stage k, $\tilde{S}_k$ is empty, set $mk = k$ and proceed to step 4.

Step 4. If $mk = m$, set $S = \tilde{S}_m$. Otherwise, set $S = \tilde{S}_{mk-1}$.

Steps 5–8. Same as algorithm 4.

Algorithm 6. Combinatorial method

The basic idea behind this method is to identify combinations of gross errors that are consistent with the observed pattern of nodal imbalances. Since the algorithm does not work with aggregates of nodes, it is less susceptible to error cancellation than the previous two methods.

Step 1. Same as algorithm 4.

Step 2. Compare each z_i with a critical test value $z_c = z_{1-\alpha/2}$. If $|z_i| > z_c$, denote node i as a bad node.

Step 3. If the set of bad nodes from the previous step is empty, proceed to step 5. Otherwise, search the set of streams that are adjacent to bad nodes for feasible subsets containing n streams, beginning with $n = 1, 2, \ldots$. A subset is termed feasible if the flow rates of the streams contained in the subset can be adjusted in such a way that the nodal imbalance test is satisfied for each node in the system and the adjusted flow rates fall within specified bounds. The following procedure is used to determine whether a given subset S_n is feasible.

a. Determine whether the streams in S_n intersect all of the bad nodes. If they do not, then S_n is not feasible. If they do, continue.

b. Compute adjusted values $\tilde{y}_j$ for the streams $j \in S_n$ by solving

$$\tilde{A}_n \tilde{y} = 0 \tag{17}$$

where $\tilde{A}_n$ is a submatrix comprising n rows of $\tilde{A}$. In this calculation $\tilde{y}_j = y_j$ for $j \notin S_n$.

c. Determine whether the following inequalities are satisfied:

$$xl_j \leq \tilde{y}_j \leq xu_j, \quad j \in S_n \tag{18}$$

where xl and xu are vectors of specified lower and upper bounds on the flow rates. If any of the inequalities is violated, then S_n is not feasible. If all inequalities are satisfied, continue.

d. Carry out the nodal imbalance test as in steps 1 and 2, but using $\tilde{y}$ instead of y. If $|z_i| > z_c$ for any node i, then S_n is not feasible. Otherwise S_n is feasible. The search for feasible subsets is terminated after the first stage, n, at which at least one feasible subset is found.

Step 4. If more than one feasible subset was found in step 3, employ the following heuristic selection procedure:

a. *Minimum chi-square criterion.* For each subset, S_{nk}, proceed as in steps 4 and 5 of algorithm 1 to compute the vector $\boldsymbol{a}_k^*$ of least-squares adjustments:

$$\boldsymbol{a}_k^* = -\boldsymbol{P}_k\boldsymbol{B}_k^T(\boldsymbol{B}_k\boldsymbol{P}_k\boldsymbol{B}_k^T)^{-1}\boldsymbol{B}_k\boldsymbol{w}_k \tag{19}$$

Then compute the optimum value $f(\boldsymbol{a}_k^*)$ of the least-squares objective function from

$$f(\boldsymbol{a}_k^*) = (\boldsymbol{a}_k^*)^T\boldsymbol{P}_k^{-1}\boldsymbol{a}_k^* \tag{20}$$

Select the subset that yields the smallest value of $f(\boldsymbol{a}_k^*)$.

b. *Minimum adjustment criterion.* If more than one subset yields the same minimum value of $f(\boldsymbol{a}^*)$ in part a, for each such subset S_{nk} calculate

$$C_k = \left(\max_{j \in S_{nk}} \left| \frac{y_j - \tilde{y}_j}{y_j} \right|\right) \sqrt{\sum_{j \in S_{nk}} \left(\frac{y_j - \tilde{y}_j}{y_j}\right)^2} \tag{21}$$

Select the subset that yields the smallest value of C_k.

Denote by S the set of streams contained in the subset selected by the above procedure. The measurements y_j, $j \in S$ are considered to contain systematic errors.

Steps 5–8. The data reconciliation procedure is the same as in steps 4–7 of algorithm 1.

The selection procedure in step 4a is essentially the same as that used by Ripps (1965) except that sets containing more than a single stream are allowed. The optimum value $f(\boldsymbol{a}_k^*)$ of the least-squares objective function is a chi-square statistic with degrees of freedom equal to the rank of the compressed incidence matrix $\boldsymbol{B}$ (Crowe et al., 1983; Wang and Stephanopoulos, 1983). Since the combinatorial search is terminated after the first stage at which a feasible subset is found, all the $f(\boldsymbol{a}_k^*)$ have the same number of degrees of freedom, and can therefore be compared on a consistent basis. Nevertheless, the selection procedure must be considered a heuristic one since the minimum chi-square value does not necessarily correspond to the correct set of bad streams.

The quantity C defined by Eq. 21 is the product of a Chebyshev, or L_∞, norm and an L_2 norm, both of which are measures of the distance between the measurement vector $\boldsymbol{y}$ and the adjusted measurement vector $\tilde{\boldsymbol{y}}$. The minimum adjustment criterion is based on the results of extensive numerical tests on a small steam network in which various combinations of the L_1, L_2, and L_∞ norms were investigated as selection criteria. The particular combination denoted by C was found to have two advantages. First, the minimum value of C was always found to be unique. Second, it yielded better results overall than other criteria. In fact, it was found to work nearly as well as the minimum chi-square criterion. These two properties make the minimum adjustment criterion well-suited for use as a "tie breaker" in conjunction with the minimum chi-square criterion. Clearly, however, this is strictly a heuristic procedure.

It should be noted that the original unsmoothed data are used in Eq. 17 to compute the adjusted values $\tilde{y}_j$. This is necessary because if the least-squares procedure were applied first to smooth the data, the nodal imbalance test in step 3d would always be satisfied. However, these adjusted values are used only for the purpose of gross error identification. The final reconciled flow rates $\hat{y}_j$, $j \in S$ are computed using the smoothed data just as they are in the other algorithms. As a result, it is possible that the reconciled flow rates will not satisfy the constraints given by Eq. 18. In practice, therefore, data reconciliation is performed for each feasible subset found in step 3. In Step 4a, the value of $f(\boldsymbol{a}_k^*)$ is minimized over those subsets for which all reconciled flow rates lie within their respective lower and upper bounds. This procedure insures that unreasonable flow rates will not be generated by the algorithm.

The combinatorial search in step 3 of the above algorithm is conducted over only those streams that are adjacent to bad nodes. Nevertheless, even a system of moderate size can result in an impractically large combinatorial problem. For example, with the network of Figure 1, if six or more gross errors are present in the measurements it is possible for all twelve nodes to be unbalanced. In that case, the search would have to be conducted over all 28 streams in the system. The following algorithm was designed to alleviate this problem.

Algorithm 7. Screened combinatorial (SC) method

This algorithm is a combination of algorithms 5 and 6. In effect, the MMP algorithm is used as a screening procedure to reduce the size of the combinatorial problem that must be handled in the combinatorial method.

Step 1. Same as in algorithm 5.

Steps 2 and 3. Same as in algorithm 5 except that those nodes for which $|z_i| > z_c$ in step 2 are also denoted as bad nodes.

Step 4. Select the set $\tilde{S}_k$ from step 3 corresponding to the highest stage of aggregation at which a nonempty set was found. Perform a combinatorial search on this set as in step 3 of algorithm 6. If no feasible subsets are found, perform the search on the set $\tilde{S}_{k-1}$ from the previous stage of aggregation in step 3. Continue in this manner until a feasible subset is found or until all sets $\tilde{S}_k$ determined in step 3 have been exhausted. In the latter case, perform one additional combinatorial search using the set of all streams that are adjacent to bad nodes. (In this case, the algorithm reverts to the combinatorial method).

Step 5. Same as step 4 of algorithm 6.

Steps 6–9. The data reconciliation procedure is the same as in steps 4–7 of algorithm 1.

Since the above algorithm may revert to the combinatorial method, it is necessary to incorporate a timing routine to prevent it from becoming bogged down in a large combinatorial calculation. A default option could be included whereby the algorithm would switch to another method in such situations. The MMP algorithm would be a logical choice. For the simulation tests, the bounds in Eq. 18 were specified as in algorithm 3.

Results and Discussion

The performance of each of the algorithms described in the previous sections, with the exception of the combinatorial method, was tested on the steam-metering network of Figure 1 using 100 randomly generated test cases. A measurement vector was constructed for each test case by using random number generators as previously described. For each algorithm, the same seeds were used in the random number generators in order to generate the same set of test cases. The results of the calculations are summarized in Table 2.

The first column of Table 2 lists the percentage of total systematic errors that was detected by each algorithm. The second and third columns give the percentages of systematic errors with absolute values greater than 2,250 and 4,500 kg/h, respectively, that were detected. Errors of 4,500 kg/h or less would normally be considered small in this system since some of the measurements generally contained random errors of this magnitude or greater. Absolute values rather than percentages are used to delineate the small errors, because in practice one is usually more concerned about detecting errors of large magnitude. Due to the wide range of flow rates in the system, a fixed percentage error includes absolute errors spanning more than two orders of magnitude. For example, an error of 10% in the largest stream amounts to 16,400 kg/h, while the same percentage error in the smallest stream amounts to 86 kg/h. Clearly, in most applications one would be more concerned about detecting the former error than the latter.

The fourth column of Table 2 gives the average number of measurements that were erroneously indicated as containing a systematic error. The values in parentheses represent the same data expressed as percentages of the number of systematic errors in the data. The fifth column lists the average percentage reduction in error achieved by each algorithm, where

$$\% \text{ Error Reduction} = \frac{E_1 - E_2}{E_1} \times 100 \tag{22}$$

and

$$E_1 = \sum_{j=1}^{28} |y_j - x_j| \tag{23}$$

$$E_2 = \sum_{j=1}^{28} |\hat{y}_j - x_j| \tag{24}$$

The reduction in absolute error is taken as the criterion of merit because the reduction in percentage error would present a misleading picture of the accomplishments of the algorithms, again due to the wide range of flow rates in the system. It should be noted, however, that since this criterion includes both random and systematic error components, it does tend to understate the achievements of the gross error detection routines per se. Nevertheless, it is the reduction in total error that is of importance in practice. In addition, the individual effects of gross error detection and data reconciliation cannot be strictly separated because, if the gross error detection procedure fails to correctly identify all the gross errors, the data reconciliation procedure may introduce large errors into good measurements in order to balance the system.

It should also be noted that an L_1 norm is used in Eqs. 23 and 24 as a measure of the distance betwen the original (or reconciled) measurement vector and the vector, x, of true flow rates. This is not inconsistent with the use of an L_2 norm in the least-squares procedure, because in the latter instance it is the distance between the measurement vector and the vector $\hat{x}$ of adjusted flow rates that is measured. That is, the two norms are used to measure different quantities for different purposes. In this application, the L_1 norm is generally a more conservative measure of the error reduction. It usually gives error reductions on the order of five to ten percentage points lower than the L_2 norm.

Finally, the last column in Table 2 shows the number of cases for which the reconciled data contained more total error than the original data, i.e., for which $E_2 > E_1$.

The MT algorithm with the recommended value of z_c = 3.1165 detected a very high percentage of systematic errors, but only at the expense of a very large number of erroneous identifi-

Table 2. Algorithm Performance Summary

Algorithm	Systematic Errors Detected %	Systematic Errors >2,250 Detected %	Systematic Errors >4,500 Detected %	Avg. No. Good Streams Found Bad % Bad Streams	Avg. Error Reduction %	No. Cases for Which Error Increased
MT, Z_c = 3.1165, α = 0.05	83	89	92	10.9 (272)	—	—
MT, Z_c = 10.0, $\alpha \simeq 0$	52	60	63	3.5 (87)	—	—
MT, Z_c = 15.0, $\alpha \simeq 0$	38	45	47	1.5 (37)	37	20
IMT, α = 0.05	51	66	69	0.62 (15)	52	12
IMT, α = 0.20	53	68	70	0.76 (19)	51	13
MIMT, α = 0.05	57	73	77	0.38 (9)	61	5
MIMT, α = 0.20	60	76	79	0.48 (12)	61	6
MP, α = 0.05	21	27	29	0.04 (1)	33	22
MMP, α = 0.05	39	49	52	0.66 (16)	43	12
SC, α = 0.05	60	76	80	0.79 (20)	62	1

Table 3. Systematic Errors Detected as Function of Number of Systematic Errors

No. Systematic Errors	No. Cases	Systematic Errors Detected, by Algorithm, %					
		MT	IMT	MIMT	MP	MMP	SC
1	20	75	60	70	30	59	65
2	10	75	70	70	20	59	75
3	10	89	63	63	33	50	67
4	12	85	48	54	25	38	65
5	17	81	49	52	25	38	54
6	19	87	50	60	17	34	62
7	12	86	45	52	14	32	55

Table 4. Average Computing Times

Algorithm	Avg. CPU Time per run, s
MT	1.4
IMT	2.9
MIMT	3.1
MP	4.9
MMP	4.8
SC	41.3

cations, which rendered the results virtually useless. In an attempt to tune the algorithm, we tried larger values of z_c. With $z_c = 10.0$, the number of erroneous identifications was still excessive. With $z_c = 15.0$, the results were more reasonable, but the percentage of systematic errors detected was greatly reduced.

For the runs with $z_c = 3.1165$ and 10.0, data reconciliation was not attempted because most of the sets of bad streams identified by the algorithm were unobservable. With $z_c = 15.0$, most of the sets were observable and those that were not could be rendered observable by removal of a single stream. In those cases, the stream to be removed was selected arbitrarily and its value was set equal to its measured value. When negative flow rates were computed, they were set to zero. These manipulations were performed only for the purpose of computing the final error, E_2, and had no effect on the entries in the first four columns of Table 2.

The poor performance of the MT algorithm on this system is not surprising in light of the results of Iordache et al. (1985). The system topography is relatively complex and includes nodes, such as nodes 5, 7, and 12 in Figure 1, that have a large number of adjacent streams; the measurement standard deviations span a wide range, covering more than two orders of magnitude; and the ratio of systematic errors to standard errors is large, with values of $|\delta_j|/\sigma_j$ as great as 40. All of these factors have been shown by Iordache et al. (1985) to be detrimental to the performance of the measurement test.

The performance of the IMT algorithm with $\alpha = 0.05$ was significantly better than that of the MT algorithm with $z_c = 15.0$ in all categories listed in Table 2. In an effort to increase the percentage of systematic errors detected, the significance level was reduced from 95 to 80% ($\alpha = 0.2$). The result was a small increase in the number of errors detected and a substantial increase in the number of erroneous identifications. The MIMT algorithm outperformed the IMT algorithm in all categories listed in Table 2. The magnitude of the error increase was less than 10% in three of the five (or six) cases in which the error increased.

The algorithms based on the nodal imbalance test were run at the 95% significance level. Runs were also made at the 90% level recommended by Mah et al. (1976), with virtually identical results. The MP algorithm made very few incorrect identifications of systematic errors, but was able to detect only 21% of the errors. The MMP algorithm performed substantially better, and the difference in the percentage of systematic errors detected by the two methods is an indication of the extent to which error cancellation occurred in the aggregation process. The MMP algorithm and the MT algorithm with $z_c = 15.0$ performed about equally well in detecting systematic errors, but the latter method made more than twice as many erroneous identifications of systematic errors. The SC algorithm was far superior to the MP and MMP algorithms in detecting systematic errors and in reducing the total error in the data. However, it made about 20% more erroneous identifications of systematic errors than did the MMP algorithm.

The overall performance of the SC algorithm was comparable to that of the MIMT method. However, the SC method generated only one case of increased error, and that was a case in which a single systematic error was partially cancelled by random errors. The error increase in that case could have been eliminated by specifying somewhat sharper lower bounds on the flow rates. In one case, the method failed to achieve a solution in the allotted time. Hence, the results in the last row of Table 2 are based on 99 cases. However, a default option using the MMP method would have achieved a solution in that case. Inclusion of that result in the data compilation would have had no significant effect on the values listed in Table 2.

Table 3 gives the percentage of systematic errors detected by each algorithm as a function of the number of systematic errors in the data. All results are for a significance level of 95%. Although the number of cases in each category is small, the results appear to be relatively independent of the number of errors. Only the MMP algorithm exhibited a monotonic decrease in the percentage of errors detected with the number of errors in the data.

The average computing time per run for each algorithm on an IBM 4341 computer is given in Table 4. The MIMT algorithm clearly represents the best combination of speed and effectiveness. However, the computing time for the SC method is not excessive if it is to be executed once per day or per shift.

In order to determine the effects on the results of changes in system parameters, a sensitivity analysis was performed using the MIMT algorithm with $\alpha = 0.05$. Doubling the error standard deviations, σ_j, resulted in a decrease in the number of gross errors detected (column 3 of Table 2) from 77 to 63%, while the average error reduction decreased from 61 to 52%. The number of erroneous identifications also decreased from 9 to 6%. The deterioration in performance was due to an increased amount of overlap between random and systematic errors, which made the detection of the latter more difficult. Decreasing the lower bound on the magnitude of the systematic errors in Eq. 9 from 10 to 5% had a similar but less pronounced effect. The number of gross errors detected in this case was 75%, the average error reduction was 59%, and the number of erroneous identifications was 10%. Increasing the lower bound in Eq. 9 from 10 to 20% produced a very small improvement in algorithm performance, one that left the performance figures essentially unchanged from the base case. Decreasing the upper bound in Eq. 9 from 100 to 50% resulted in somewhat poorer performance due to shifting of the systematic errors toward the smaller size ranges, thereby making their detection more difficult. The number of gross errors detected was 74%, the average error reduction was 50%, and the number of erroneous identifications was 8%. The relatively large decrease in average error reduction was primarily due to the fact that the systematic errors, being smaller, constituted a smaller fraction of the total error in the data. Finally, it should be noted that specifying sharper bounds on the flow rates in the MIMT and SC algorithms may result in significantly improved performance of these methods. However, this effect was not systematically investigated in this work.

Notation

a = vector of adjustments to measured flow rates
a^* = vector of least-squares adjustments to measured flow rates
A = incidence matrix excluding environmental node
$\tilde{A}$ = incidence matrix including environmental node
B = compressed incidence matrix obtained by deleting bad streams
C = minimum adjustment parameter, Eq. 21
e = vector of residuals
E_1, E_2 = total error in data before and after reconciliation
G = matrix, Eq. 16
i = node index
j = stream index
k = index of feasible subsets; iteration counter
m = maximum number of nodes in a pseudonode
mk = identification variable in MMP method
n = number of measurements tested in Eq. 12; number of streams in subset S_n
P = compressed covariance matrix
Q = covariance matrix
r = vector of nodal imbalances
S = final set of bad streams determined by gross error detection procedure
S_n = stream subset containing n streams
$\tilde{S}_k$ = set of bad streams after k stages of aggregation
T = set of streams remaining in system after nodal aggregation
U = set of streams in original system
V = matrix, Eq. 11
w = compressed flow rate vector
x = vector of true flow rates
$\hat{x}$ = vector of adjusted flow rates
$\hat{x}^*$ = vector of adjusted flow rates from least-squares procedure
xl = vector of specified lower bounds on flow rates
xu = vector of specified upper bounds on flow rates
y = vector of measured flow rates
$\hat{y}$ = vector of final reconciled flow rates
$\tilde{y}$ = vector of adjusted flow rates corresponding to feasible subset S_n
$\tilde{y}_k$ = vector of corrected flow rates at kth iteration of MIMT algorithm
z = test statistic, Eq. 10 or Eq. 15
z_c = critical test value
$z_{1-\alpha/2}, z_{1-\beta/2}$ = $(1 - \alpha/2)$th or $(1 - \beta/2)$th quantile of standard normal distribution

Greek letters

α = probability of type I error
β = probability of type I error in testing each measurement when a group of n measurements is being tested with overall type I error probability
δ = vector of systematic errors
ϵ = vector of random errors
σ_j = standard deviation of measurement error for stream j

Literature Cited

Almasy, G. A., and T. Sztano, "Checking and Correction of Measurements on the Basis of Linear System Model," *Prob. Control Infor. Theory*, **4**, 57 (1975).

Crowe, C. M., Y. A. Garcia Campos, and A. Hrymak, "Reconciliation of Process Flow Rates by Matrix Projection," *AIChE J.*, **29**, 881 (1983).

Himmelblau, D. M., "Material Balance Rectification via Interval Arithmetic," Paper no. 55c, AIChE Meet., Anaheim, CA (May, 1984).

Hadley, G., *Nonlinear and Dynamic Programming*, Addison-Wesley, Reading, MA (1964).

Iordache, C., R. S. H. Mah, and A. C. Tamhane, "Performance Studies of the Measurement Test for Detection of Gross Errors in Process Data," *AIChE J.*, **31**, 1,187 (1985).

Kuehn, D. R., and H. Davidson, "Computer Control. II: Mathematics of Control," *Chem. Eng. Prog.*, **57**, 44 (1961).

Madron, F., V. Veverka, and V. Vanacek, "Statistical Analysis of Material Balance of a Chemical Reactor," *AIChE J.*, **23**, 482 (1977).

Mah, R. S., G. M. Stanley, and D. Downing, "Reconciliation and Rectification of Process Flow and Inventory Data," *IEC Proc. Des. Dev.*, **15**, 175 (1976).

Mah, R. S. H., and A. C. Tamhane, "Detection of Gross Errors in Process Data," *AIChE J.*, **28**, 828 (1982).

Nogita, S., "Statistical Test and Adjustment of Process Data," *IEC Proc. Des. Dev.*, **11**, 197 (1972).

Reilly, P. M., and R. E. Carpani, "Application of Statistical Theory of Adjustment to Material Balances," *Proc. 13th Can. Chem. Eng. Conf.*, Montreal, (Oct., 1963).

Ripps, D. L., "Adjustment of Experimental Data," *Chem. Eng. Prog. Symp. Ser. No. 55*, **61**, 8 (1965).

Romagnoli, J. A., and G. Stephanopoulos, "On the Rectification of Measurement Errors for Complex Chemical Plants," *Chem. Eng. Sci.*, **35**, 1,067 (1980).

———, "Rectification of Process Measurement Data in the Presence of Gross Errors," *Chem. Eng. Sci.*, **36**, 1,849 (1981).

Stanley, G. M., and R. S. H. Mah, "Observability and Redundancy in Process Data Estimation," *Chem. Eng. Sci.*, **36**, 259 (1981a).

———, "Observability and Redundancy Classification in Process Networks. Theorems and Algorithms," *Chem. Eng. Sci.*, **36**, 1,941 (1981b).

Stephenson, G. R., and C. F. Shewchuk, "The Reconciliation of Process Data with Process Simulation," Paper No. 55a, AIChE Meet., Anaheim, CA (May, 1984).

Wang, N. S., and G. Stephanopoulos, "Application of Macroscopic Balances to the Identification of Gross Measurement Errors," *Biotech. and Bioeng.*, **25**, 2,177 (1983).

Manuscript received June 29, 1984, and revision received Aug. 28, 1985.