IDENTIFICATION OF GROSS ERRORS IN BALANCE MEASUREMENTS

František MADRON

Chemopetrol, Research Institute of Inorganic Chemistry,
400 60 Ústí nad Labem

Received May 10th, 1983

Methods so far proposed for identifying gross error sources in making a one-component mass balance are compared for their effectiveness and the work load involved. Causes of failure of some of the methods are discussed. An approach in which arrangement of normalized adjustments is coupled with successive elimination of measured quantities is recommended as the preferred method.

One of the first steps in testing the performance of chemical plants is to set up a mass balance on the basis of measured data. If redundant measurements are present, statistical methods may be applied to find out whether the measured data are subject to any gross and systematic errors. If the occurrence of such errors is ascertained, other methods may be used to identify measurements suspect on this count, and causes of gross errors may subsequently be determined[1]. Such a procedure is spoken of as identification of gross error sources. In this communication we shall compare methods which have been proposed for identifying gross error sources. In doing so, we shall consider a one-component mass balance, but most conclusions will apply to a multi-component balance as well.

BALANCE PROBLEM

Let us consider J balance nodes, N_j, $j = 1, 2, \ldots J$ and I oriented streams, $H_i = 1, 2, \ldots, I$. The flows in the individual streams will be denoted as x_i. The J-th node usually represents the environment, whose inlet and outlet streams connect the other, so-called technological nodes with the environment. A total of $(J\text{-}1)$ independent balance equations may be written for the individual nodes.

Fig. 1 shows a balance scheme consisting of technological nodes *1′*, *2′*, *3′* and *4′* and streams *1*, *2*, ..., *9*. For clarity, node *5′* representing the environment with outlet stream *1* and inlet streams *6*, *7* and *9* is omitted from the Fig. 1.

A total of four linearly independent balance equations may be written for the balance nodes shown in Fig. 1.

$$x_1 - x_2 - x_3 + x_4 = 0 \qquad (1)$$

$$x_2 - x_5 - x_6 + x_8 = 0 \tag{2}$$

$$-x_4 + x_5 - x_7 = 0 \tag{3}$$

$$x_3 - x_8 - x_9 = 0\,. \tag{4}$$

The fifth equation relating to the environment node is dependent on these four equations. Eqs (*1*) to (*4*) may conveniently be written in the matrix form

$$\boldsymbol{Ax} = \boldsymbol{0}\,, \tag{5}$$

where $\boldsymbol{A}$ is the so-called reduced incidence matrix of the balance scheme diagram. The elements of a $(J\text{-}1) \times I$ reduced incidence matrix are equal to 0, 1, or -1, according to the rule

$A_{ji} = 1$ when stream i enters node j

$A_{ji} = -1$ when stream i leaves node j

$A_{ji} = 0$ in other cases (*i.e.* when stream i is not connected with node j).

Matrix $\boldsymbol{A}$ is called reduced because it does not contain the row relating to the environment node. The reduced incidence matrix for the scheme shown in Fig. 1 is

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & -1 \end{pmatrix}$$

Classification of Streams

According to whether or not a stream is measured directly we distinguish directly measured and directly unmeasured streams. Some directly measured streams are

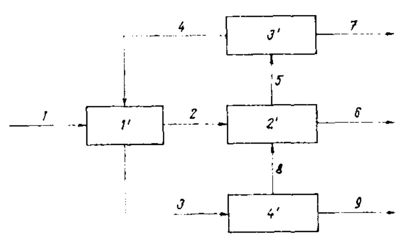

FIG. 1
Balance scheme

redundant in the sense that in the absence of measurement errors they could be unambiguously calculated from values of the remaining directly measured streams. Directly measured streams that are not redundant are called just determined. Directly unmeasured streams are divided into determinable, which can be calculated from directly measured streams, and indeterminable, which cannot be calculated. Determinable unmeasured streams are also called indirectly measured. Some measurements may be incomparably more accurate than others; such streams will be called exactly known.

The above given classification of streams may easily be carried out in the case of a one-component balance. There are two rules[7]:

Rule 1: Let stream H_m linking nodes N_j and N_k be measured. The measurement of stream H_m is redundant just when nodes N_j and N_k are not connected by a path of unmeasured streams (the path is a sequence of streams $H_1, H_2, \ldots, H_n$ such that the nodes connected by stream H_i are linked to streams H_{i-1} and H_{i+1} for $i = 2, 3, \ldots, n - 1$).

Rule 2: A directly unmeasured stream is indeterminable just when it is found on a circuit of directly unmeasured streams (by the circuit we mean a closed path of streams, with the initial node of the path identical with the final node).

Fig. 2*a* shows a choice of directly measured streams. Here, stream H_4, for instance, is indeterminable because it occurs on a circuit of unmeasured streams H_2, H_4 and H_5.

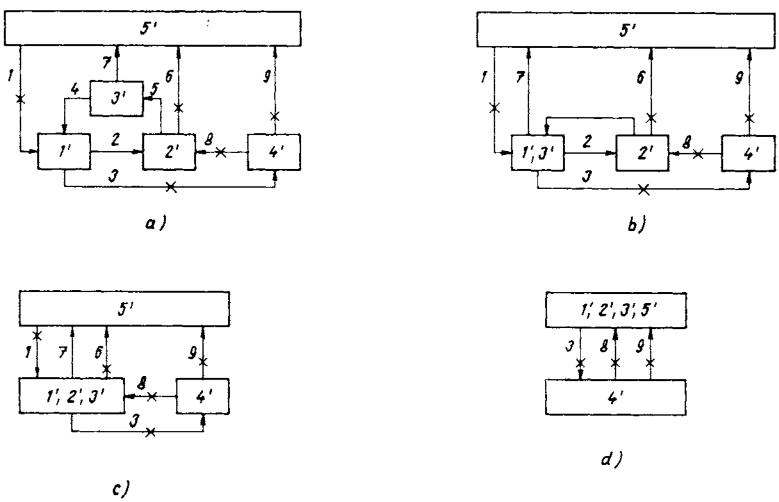

FIG. 2

Balance scheme reduction. —×— directly measured stream; — — directly unmeasured stream

Similarly, stream H_1 is just determined because its nodes N_1 and N_3 are connected through path H_4 and H_7.

The set of redundant measurements may be found as follows. Combining nodes connected through directly unmeasured streams, we finally obtain a simplified balance scheme containing directly measured streams only. All redundant direct measurements can be found in this way. Directly measured streams eliminated in combining the nodes are just determined measured streams.

A balance scheme simplified in this way is said to be reduced, and the process of balance scheme simplification is called reduction[2] (it should be pointed out that these terms have nothing in common with the concept of reduced incidence matrix of balance scheme).

Balance scheme reduction for the given choice of directly measured streams is represented in Figs 2*b*, *c* and *d*, where streams *3*, *8* and *9* are seen to be redundant.

Mathematical Model

Denoting the vectors of directly measured flows, directly unmeasured flows and exactly known flows by $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$, respectively, we can write Eq. (*5*) in the form

$$\boldsymbol{A}_1\boldsymbol{x}_1 + \boldsymbol{A}_2\boldsymbol{x}_2 + \boldsymbol{A}_3\boldsymbol{x}_3 = \boldsymbol{0}\,, \tag{6}$$

where $\boldsymbol{A}_1$, $\boldsymbol{A}_2$ and $\boldsymbol{A}_3$ are the incidence matrices of directly measured, directly unmeasured and exactly known streams, respectively. As the product $\boldsymbol{A}_3\boldsymbol{x}_3$ is a known constant, we can replace it in Eq. (*6*) by a constant vector $\boldsymbol{a}$.

$$\boldsymbol{A}_1\boldsymbol{x}_1 + \boldsymbol{A}_2\boldsymbol{x}_2 + \boldsymbol{a} = \boldsymbol{0}\,. \tag{7}$$

We shall now confine ourselves to the case, where no indeterminable quantities are present and, in addition, at least some of the directly measured quantities are redundant. Where this case arises may be seen from the classification rules given in the preceding section. The conditions for the validity of these assumptions may, in addition, be expressed in terms of the ranks of matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$.

$$h(\boldsymbol{A}_1, \boldsymbol{A}_2) = J - 1\,; \quad h(\boldsymbol{A}_2) = I_2\,; \quad I_1 > J - 1 - I_2 > 0 \tag{8}$$

Eq. (*7*) holds true for actual values of vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The actual value of vector $\boldsymbol{x}_1$ is related to result of direct measurement $\boldsymbol{x}_1^+$ by

$$\boldsymbol{x}_1^+ = \boldsymbol{x}_1 + \boldsymbol{e}\,, \tag{9}$$

where $\boldsymbol{e}$ is the vector of measurement errors. Information on unknown errors of measurement is varied. In some cases, it may be assumed that the elements of vector $\mathbf{e}$ are realizations of random quantities with zero mean and known variance,

$$E(e_i) = 0\,, \quad V(e_i) = \sigma_i^2\,. \tag{10}$$

Sometimes, maximum errors of measurements are known (*e.g.* as accuracy classes of the measurement devices). Denoting by e_i^m the maximum measurement error in i-th quantity, we may define the standard deviation of this error as

$$\sigma_i = k e_i^m \quad k \in \langle 1/2;\, 1/3 \rangle\,. \tag{11}$$

The value $k = 1/3$ is chosen for reliable measurements where errors greater than e_i^m may practically be ruled out. In ordinary cases, $k = 1/2$ may be recommended.

Adjustment of Redundant Measurements

If redundant measurements are present, Eq. (7) is mostly not exactly satisfied by the measured values. We then solve the problem of adjustment of redundant measurements, by searching for a vector of adjusted values $\hat{\mathbf{x}}_1$ given by

$$\hat{\mathbf{x}}_1 = \mathbf{x}_1^+ + \mathbf{v}\,,$$

where $\mathbf{v}$ is the vector of adjustments. The requirements are that the adjusted values $\hat{\mathbf{x}}_1$ exactly satisfy Eq. (7) and that the adjustments be minimal in a sense. The most frequent procedure is to minimize the expression

$$Q = \sum_{i=1}^{I_1} v_i^2 \sigma_i^{-2} \tag{12}$$

which may be written in matrix notation as

$$Q = \mathbf{v}^T \mathbf{F}^{-1} \mathbf{v}\,, \tag{13}$$

where $\mathbf{F}$ is the diagonal matrix with variances of measurements σ_i^2 in the diagonal. The solution to this problem (the so-called generalized least squares problem) is obtained by solving the following set of equations[5]

$$\begin{pmatrix} \mathbf{A}_1\mathbf{F}\mathbf{A}_1^T, & \mathbf{A}_2 \\ \mathbf{A}_2^T, & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{k} \\ \hat{\mathbf{x}}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{a} + \mathbf{A}_1\mathbf{x}_1^+ \\ \mathbf{0} \end{pmatrix} = \mathbf{0} \tag{14}$$

$$\mathbf{v} = \mathbf{F}\mathbf{A}^T\mathbf{k} \tag{15}$$

$$\hat{\mathbf{x}}_1 = \mathbf{x}_1^+ + \mathbf{v} \tag{16}$$

where $\boldsymbol{k}$ is the vector of Lagrangian multipliers, and $\hat{\boldsymbol{x}}_2$ is the vector of estimates of directly unmeasured quantities. We first solve Eq. (*14*)

$$\begin{pmatrix} \boldsymbol{k} \\ \hat{\boldsymbol{x}}_2 \end{pmatrix} = - \begin{pmatrix} \boldsymbol{A}_1\boldsymbol{F}\boldsymbol{A}_1^{\mathrm{T}}, & \boldsymbol{A}_2 \\ \boldsymbol{A}_2^{\mathrm{T}}, & \boldsymbol{0} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{a} + \boldsymbol{A}_1\boldsymbol{x}_1^+ \\ \boldsymbol{0} \end{pmatrix} \tag{17}$$

and substitute $\boldsymbol{k}$ into Eq. (*15*).

If unmeasured streams are not present, mathematical model (*7*) simplifies to

$$\boldsymbol{A}_1\boldsymbol{x}_1 + \boldsymbol{a} = \boldsymbol{0}\,. \tag{18}$$

It is useful to realize that this model is arrived at by reduction of balance scheme (matrix $\boldsymbol{A}_1$ is the reduced incidence matrix of reduced balance scheme, and $\boldsymbol{x}_1$ is the vector of all redundant direct measurements). In this case, the adjustment is of the simple form

$$\boldsymbol{v} = -\boldsymbol{F}\boldsymbol{A}_1^{\mathrm{T}}(\boldsymbol{A}_1\boldsymbol{F}\boldsymbol{A}_1^{\mathrm{T}})^{-1}\,(\boldsymbol{a} + \boldsymbol{A}_1\boldsymbol{x}_1^+) \tag{19}$$

$$\hat{\boldsymbol{x}}_1 = \boldsymbol{x}_1^+ + \boldsymbol{v}\,. \tag{20}$$

Since only redundant, directly measured quantities are of importance for identification of gross errors of measurement, it suffices to deal with model (*18*) only. All the information acquired from this model is equivalent to that which would be obtained by solving the general model (*7*). In what follows a mathematical model of the form of Eq. (*18*) will therefore be considered; this implies that either all the streams have been measured or, if unmeasured streams are present, the corresponding reduction of balance scheme has been carired out.

We shall now present important probability properties of the vectors $\hat{\boldsymbol{x}}_1$ and $\boldsymbol{v}$ and of the quantity Q, assuming that the measurement errors possess the I-variate normal distribution $\mathbf{e} \approx \mathrm{N}_I\,(\boldsymbol{0}, \boldsymbol{F})$. The covariance matrices of the vectors $\boldsymbol{v}$ and $\hat{\boldsymbol{x}}_1$ are[5]

$$\boldsymbol{F}_{\mathrm{v}} = \boldsymbol{F}\boldsymbol{A}_1^{\mathrm{T}}(\boldsymbol{A}_1\boldsymbol{F}\boldsymbol{A}_1^{\mathrm{T}})^{-1}\,\boldsymbol{A}_1\boldsymbol{F} \tag{21}$$

$$\boldsymbol{F}_{\hat{\boldsymbol{x}}_1} = \boldsymbol{F} - \boldsymbol{F}_{\boldsymbol{v}}\,. \tag{22}$$

Q is the realization of a random variable with χ^2 distribution having ν degrees of freedom, where ν is the number of linearly independent equations in set (*18*).

Another important concept is that of residuals, $\boldsymbol{r}$, of Eq. (*18*).

$$\boldsymbol{r} = \boldsymbol{A}_1\boldsymbol{x}_1^+ + \boldsymbol{a} \tag{23}$$

The mean value of the residuals is zero, and their covariance matrix is given by

$$\boldsymbol{F}_r = \boldsymbol{A}_1 \boldsymbol{F} \boldsymbol{A}_1^{\mathrm{T}} . \tag{24}$$

As in the case of $\boldsymbol{x}_1$ and $\boldsymbol{v}$, the distribution of the residuals is normal.

Detection of Gross Errors of Measurement

By a gross error we mean one whose magnitude does not fit into our picture of the magnitude of random error variance. It is customary to consider as gross error one which is greater in absolute value than three times the standard deviation of random error.

The presence of a gross error may be inferred from essentially three types of information – adjustments, residuals, and the value of Q. If no error is present, all the three quantities are zero. If only random errors occur, the quantities should lie, with a given probability, within limits determined by the distribution of the given random quantity.

In detecting the presence of a gross error, a suitable criterion is the value of the quantity Q. The hypothesis that a gross error is absent (hypothesis H_0) will be rejected if

$$Q > \chi^2_{1-\alpha}(\nu) , \tag{25}$$

where $\chi^2_{1-\alpha}(\nu)$ is the $100(1 - \alpha)$ percentile of $\chi^2(\nu)$ distribution. The probability α is the probability of rejecting the hypothesis when it is true (error of the first kind). Conversely, we may, with probability γ, make an error of the second kind by not rejecting hypothesis H_0 when it is false. The value $\beta = 1 - \gamma$ is called the power of a test, and the dependence of β on the magnitude of gross error is the power curve of the test.

Identification of Gross Error Source

The methods of identifying measurements subject to gross errors are based on the fact that a gross error will propagate in a characteristic manner into adjustments of the various measured quantities and into residuals of the individual equations. It should be borne in mind. however, that the propagation is characteristic in a stochastic sense only (the errors cannot be rigorously calculated from the adjustments or residuals). We shall now discuss the three methods most frequently employed in identifying gross error sources. For simplicity, we shall consider the occurrence of a single error. The method may be applied to cases where several gross errors occur simultaneously but their effectiveness falls off rapidly as the number of gross errors increases.

Analysis of Adjustments

The magnitudes of the individual adjustments depend on three factors:

- assumed variances of measurements (the variance is the greater the larger the adjustment)
- the structure of the mathematical model
- particular values of measurement errors.

In order to isolate the last named effect, normalization of adjustments is carried out in such a way as to achieve a compensation of the first two effects. The adjustments are random quantities with a covariance matrix given by Eq. (*21*). By finding the square roots of the diagonal elements of matrix $\mathbf{F}_v$, we obtain standard deviations of the adjustments, $\sigma_{v_i} = (F_{v_{ii}})^{1/2}$. Forming the ratios

$$w_i = v_i/\sigma_{v_i} , \tag{26}$$

we obtain a vector of quantities with standard normal distribution. Experience shows that the quantity w_i is a suitable indicator of the presence of gross error in i-th quantity. If some measurement is subject to gross error, the absolute value of w_i for this quantity ranks among the highest. By simply arranging the w_i's in order of their magnitudes, we obtain the order of the quantities with respect to their being "suspected" of the presence of gross error.

Adjustments may also be tested for the presence of gross error. The quantities w_i have standard normal distribution, and an i-th measured quantity will be suspected of the presence of gross error if

$$|w_i| > u_{1-\alpha/2} , \tag{27}$$

where $u_{1-\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution. Usually, α is chosen from the interval $\langle 0{\cdot}01 - 0{\cdot}10 \rangle$. It should, however, be realized that in contrast to the test based on inequality (*25*), the probability of error of the first kind is not exactly known in this case (because of the larger number of quantities under test, the overall probability of this error is greater than α).

Analysis of Residuals of Mathematical Model Equations

An important property of the set of balance equations (*18*) is that each measured quantity appears in one or at most two equations. Hence, gross error of a measurement may be expected to give rise to an extreme value of residual for one or two equations in which the measurement is present. With this assumption, we can formulate algorithms for gross error identification, which are based on residuals of balance equations and on the balance scheme structure.

We shall form the normalized residuals

$$z_j = r_j/\sigma_{r_j}\,, \tag{28}$$

where $\sigma_{r_j} = (F_{r_{jj}})^{1/2}$. If measurements present in j-th equation are not subject to gross error, z_j has the standard normal distribution. The hypothesis that j-th equation is free of gross error will be rejected if

$$|z_j| > u_{1-\alpha/2}\,. \tag{29}$$

If we accept the hypothesis of the presence of gross error in j-th equation, the quantities appearing in this equation (*i.e.*, measurements of streams connected to the respective balance node) are the possible sources of error. A further reduction of the set of suspected quantities is achieved by examining pairs of neighbouring nodes and pseudo-nodes formed by combining neighbouring nodes. Table I gives, in an abbreviated form, an algorithm for identification of measurement subject to gross error published by Hlaváček and co-authors.[2]

Mah and coworkers[7] have proposed a more complicated method based on the following assumptions:

1) random errors are negligible compared with gross errors

2) a gross error for a given node occurs if and only if the measurement of stream connected with this node is subject to gross error or if there is a significant escape of mass from this node to the environment

3) there is no compensation of errors

4) all streams are measured

5) no parallel streams (more streams connecting two nodes) are present.

Mah and coworkers[7] have demonstrated that with these assumptions, gross errors and escape of mass into the environment can be identified provided that the streams subject to gross errors do not form a circuit in the given balance scheme. For details of this method, the reader is referred to the literature[7].

Successive Elimination of Measured Quantities

Let us assume that the test based on inequality (*25*) has revealed a gross error and that the gross error is the only one present. We shall now successively consider the individual directly measured quantities as unmeasured which, in practice, requires further reduction of the balance scheme and a change in model (*18*), and adjust the remaining measured values. If listing i-th quantity as unmeasured eliminates the gross error, this quantity is suspected of the presence of gross error.

It may happen that this procedure will not result in elimination of gross error. The most probable explanation then is that more gross errors are present. A logical extension of the described algorithm is to list pairs of quantities as unmeasured, selecting the pairs preferentially from those quantities for which marked reduction in the value of Q was found in the preceding step.

Comparison of the Methods

The methods will be compared for their effectiveness in identifying a gross error in the balance scheme shown in Fig. 1. This scheme has some features of a typical chemical technology process, particularly the presence of recycle and streams of different sizes. However, it contains no parallel streams (connecting two nodes) as these are known to be indistinguishable from the viewpoint of gross error identification. We shall assume that all streams in Fig. 1 are directly measured with a constant relative accuracy.

Table II gives simulated results of measurements: actual values of x_i, relative standard deviations of measurements γ_i, standard deviations σ_i, measurement errors e_i, and "measured" values x_i^+. The errors e_i were generated as realizations of random quantity with zero mean, standard deviation σ_i, and normal distribution. The measured values x_i^+ were adjusted by using Eqs (*19*) and (*20*). Results of the adjustment are summarized in Table III. Q was calculated from Eq. (*13*) to be 2·27, and hence the test based on inequality (*25*) revealed no gross error at 5% level of significance (the critical value of χ^2 distribution with four degrees of freedom for $\alpha = 0{\cdot}05$ was 9·49).

TABLE I

Algorithm for gross error identification[2]. 1 — an error indicated; 0 — no error indicated

Balanced node (N_iN_j)	N_i	N_j	Identification of error
1	1	1	All streams connected to nodes N_i and N_j
1	1	0	External stream connected to node N_i
1	0	1	External stream connected to node N_j
1	0	0	Contradictory result
0	1	1	Internal stream connecting N_i and N_j
0	1	0	Contradictory result
0	0	1	Contradictory result
0	0	0	No error in any of the streams

The methods of gross error identification were tested as follows. Gross errors were added to values of x_i^+ given in Table II successively for $i = 1, 2, \ldots, 9$, for each i at four levels, namely $+5\sigma_i$, $+10\sigma_i$, $+15\sigma_i$ and $+20\sigma_i$. The relative standard deviations were assumed to be 2% of the actual values, and hence the gross errors lay within 10 to 40% of the actual values of the measured quantity. The data so obtained were adjusted by using Eqs (*19*) and (*20*). The normalized adjustments were calculated from Eq. (*26*), and the normalized residuals were obtained from Eq. (*28*). The presence of gross error was tested by applying inequality (*25*). If a gross error was de-

TABLE II
Input data

i	x_i kg s^{-1}	γ_i %	σ_i kg s^{-1}	e_i kg s^{-1}	x_i^+ kg s^{-1}
1	100	2	2	1·7	101·7
2	90	2	1·8	1·8	91·8
3	20	2	0·4	0·3	20·3
4	10	2	0·2	0·5	10·5
5	15	2	0·3	−0·1	14·9
6	80	2	1·6	1·4	81·4
7	5	2	0·1	−0·05	4·95
8	5	2	0·1	0·05	5·05
9	15	2	0·3	0·06	15·06

TABLE III
Adjusted values

i	$\hat{x}_i$ kg s^{-1}	v_i kg s^{-1}	σ_{vi} kg s^{-1}	w_i
1	101·58	−0·12	1·73	−0·07
2	91·72	−0·08	1·50	−0·05
3	20·18	−0·12	0·32	−0·46
4	10·33	−0·17	0·12	−1·44
5	15·24	0·34	0·24	1·36
6	81·54	0·14	1·24	0·11
7	4·913	−0·037	0·026	−1·43
8	5·057	0·007	0·020	0·35
9	15·13	0·07	0·18	0·38

tected, the individual normalized adjustments and the normalized residuals were tested according to inequalities (*27*) and (*29*), respectively (in all cases at $\alpha = 0{\cdot}05$).

Three methods were used to identify gross error.

Method 1: Arrangement of normalized adjustments in order of decreasing absolute value. In this method, only those normalized adjustments were considered for which gross errors were detected by using inequality (*27*).

Method 2: Analysis of normalized residuals of balance equations. This approach involved two steps. The first step was to identify a set of streams connected to nodes for which the test based on inequality (*29*) revealed the presence of gross errors. In the second step, the range of suspected quantities was reduced by using the algorithm given in Table I. Where this led to contradiction (indicated by "C" in Table IV), the set of streams identified in the first step was taken as the result.

Method 3: Successive elimination of measured quantities. After listing a quantity as unmeasured, the remaining data were adjusted. Inequality (*25*) was applied to test the hypothesis that a gross error was absent.

Results of gross error identification are summarized in Table IV. The first column gives the number of stream subject to a gross error whose magnitude is listed in the second column. The column headed Q gives values calculated from Eq. (*13*) for data subject to gross errors. The next column lists results of Method 1, presented as a sequence of normalized adjustments w_i in order of decreasing magnitude. The following two columns relate to the method of analysis of residuals, one giving the nodes and the other the streams in which gross errors were identified. In some cases there is contradiction. The last column of the table lists measurements identified by Method 3 as suspected of the presence of gross errors.

DISCUSSION

The presence of gross error was detected in streams *1*, *2*, *3*, *4* and *6* for a gross error equal to as little as five times the standard deviation of random error in the given quantity. In streams *5*, *7* and *9*, the detectable gross error was ten times, and in stream *8* twenty times, the standard deviation.

Arranging the normalized adjustments in order of decreasing magnitude always placed adjustment to the stream subject to gross error as first in the sequence. The number of normalized adjustments which exceeded the corresponding critical value from inequality (*27*) ranged between *1* and *5*, increasing with the magnitude of gross error.

The method of analysis of residuals was successful in detecting gross errors in six cases, but in some of them only when large errors were involved. In the other cases there was contradiction as indicated by lines *6* and *7* of Table I.

TABLE IV

Results of methods of gross error identification

Stream number i	Gross error $+\sigma_i$	Q	Method 1 Sequence of w_i	Method 2 Gross error in nodes	Method 2 Gross error in streams	Method 3 Gross error in stream
1	5	18·1	1	1, 5	1	1
	10	56·7	1, 6, 2	1, 5	1	1
	15	109	1, 6, 2	1, 5	1	1
	20	171	1, 6, 2, 4	1, 5	1	1
2	5	16·8	2, 6	1, 2	2	2
	10	52·9	2, 6, 1	1, 2	2	2
	15	103	2, 6, 1	1, 2	2	2
	20	161	2, 6, 1, 5	1, 2	2	2
3	5	18·1	3, 9, 8	4 (C)	3, 8, 9	3, 8, 9
	10	54·4	3, 9, 8	4 (C)	3, 8, 9	3, 8, 9
	15	104	3, 9, 8	4, 1	3	3
	20	162	3, 9, 8	4, 1	3	3
4	5	16·1	4, 5, 7	3 (C)	4, 5, 7	4, 5, 7
	10	40·7	4, 5, 7	3 (C)	4, 5, 7	4, 5, 7
	15	74·1	4, 5, 7	3 (C)	4, 5, 7	4, 5, 7
	20	114	4, 5, 7	3 (C)	4, 5, 7	4, 5, 7
5	5	5·75	—	—	—	—
	10	33·2	5, 4, 7	3 (C)	4, 5, 7	4, 5, 7
	15	76·7	5, 4, 7	3 (C)	4, 5, 7	4, 5, 7
	20	130	5, 4, 7	3, 2	5	4, 5, 7
6	5	14·9	6, 7	2, 5	6	6
	10	46·7	6, 2, 1, 7	2, 5	6	6
	15	92·1	6, 2, 1, 7, 4	2, 5	6	6
7	5	9·35	—	—	—	—
	10	18·9	7, 5, 4	3	4, 5, 7	4, 5, 7
	15	31·5	7, 5, 4	3	4, 5, 7	4, 5, 7
	20	47·0	7, 5, 4	3	4, 5, 7	4, 5, 7
8	5	3·36	—	—	—	—
	10	5·47	—	—	—	—
	15	9·44	—	—	—	—
	20	15·2	8, 3, 9	4 (C)	3, 8, 9	3, 8, 9
9	5	8·85	—	—	—	—
	10	28·7	9, 3, 8	4 (C)	3, 8, 9	3, 8, 9
	15	59·7	9, 3, 8	4 (C)	3, 8, 9	3, 8, 9
	20	99·2	9, 3, 8	4, 5	9	9

The method of elimination of measured quantities correctly identified five streams subject to gross errors, but in three cases only for relatively high values of gross error. In the remaining cases the method indicated suspected quantities, always in groups of three.

Comparison of the methods shows that they are not much different in respect of their effectiveness in searching for gross error sources. The sets of suspected quantities identified by the methods of residual analysis and of elimination of measured quantities are virtually the same. The method based on values of normalized adjustments identifies somewhat larger sets, but its virtue lies in arranging the streams according to their suspiciousness. That the actual error source was placed in all the cases as first in the sequence must, however, be considered as coincidence. The above conclusions are in line with experience from use of the methods in eliminating gross errors of measurement in chemical industry practice.

We shall now examine why the method of analysis of residuals of balance equations for the individual nodes frequently yields contradictory results. Let us assume that measurement of a single stream is subject to gross error. The probability of detecting the gross error by applying inequality (*29*) to residuals of nodes connected by this stream depends on the magnitude of the gross error and on standard deviations of residuals for these nodes. If the standard deviations of the residuals are the same, the probabilities of detecting the gross error for the two nodes are approximately the same. If, however, the standard deviations differ significantly, the gross error will be found preferentially for the node with the smaller standard deviation. This is illustrated by the following example.

The vector of standard deviations of residuals of the balance scheme shown in Fig. 1

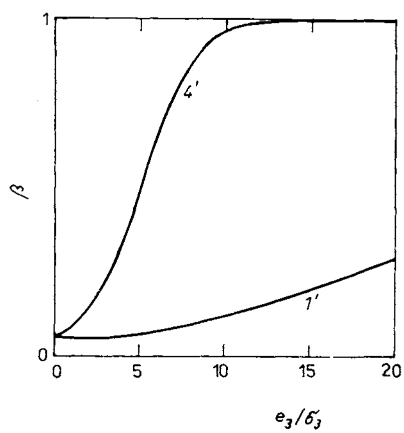

FIG. 3
Power curves of test for nodes *1'* and *4'* and stream *3*

is

$$\sigma_r^T = (2{\cdot}93;\ 2{\cdot}47;\ 0{\cdot}38;\ 0{\cdot}52;\ 2{\cdot}78)\,.$$

A significant difference is found, for example, for standard deviations of residuals of nodes *1'* and *4'*. The probability of detecting a gross error in stream *3* by applying inequality (*29*) to nodes *1'* and *4'* may be represented by means of a power curve of the test (the dependence of the probability of detecting a gross error on its magnitude). The power curves for the two nodes are shown in Fig. 3; the gross error in measurement of stream *3* is expressed here as a multiple of the standard deviation of stream *3* measurement. The power curves show that the probability of detecting a gross error of, say, $10\sigma_3$ is 97% for node *4'* while only 6% for node *1'* (accordingly, the error is not likely to be detected in the latter case). This is consistent with results presented in Table IV. Thus, it is seen that if the gross error in stream *3* measurement is in a certain interval, it cannot be expected to lead to detection of gross errors in both the nodes connected with this stream (assumption 2 presented by Mah and coworkers[7]).

When using algorithms based on analysis of residuals, we often arrive at a contradiction. In real cases, residuals of balance equations cannot be expected to have approximately the same standard deviations; in fact, the reverse is true as a rule, and the standard deviations of residuals may differ even by several orders of magnitude.

Comparing the methods under test as to their complexity and the work load involved, we arrive at the following conclusions. The least laborious method is that based on arrangement of normalized adjustments. The adjustments are available from data adjustment, and the same is mostly true for standard deviations of the adjustments, which are intermediate results in the calculation of standard deviations of the adjusted quantities (see Eq. (*22*)). A disadvantage of the method is that the conclusions drawn are not definitive (they give no information as to which quantities are sufficient to explain the presence of gross error).

The methods based on analysis of residuals do not involve adjustment of redundant data. However, identification of gross error requires an algorithm program which may be quite complicated if it is to be effective.

The method of elimination of measured quantities involves as many adjustments as there are directly measured quantities. In real cases, this may become a limiting factor even when the computation saving method described by Ripps[3] is taken into consideration.

CONCLUSION

The case analyzed here in detail, as well as experience from solutions of real problems of identifying gross errors of measurements in chemical plants, has shown that the proposed methods are helpful in detecting measurements subject to gross errors. It should, however, be borne in mind that only rarely are these methods capable of unambiguous identification of the cause of gross error. In most cases, they only allow a set of probable sources of error to be delineated, the detailed establishment of the gross error source being a matter of analysis of the measuring process itself.

Among the methods so far proposed for identifying gross errors that based on arrangement of normalized adjustments may be recommended. The results obtained may then be refined by the method of elimination of measured quantities applied only to quantities which have been singled out by the former method.

LIST OF SYMBOLS

$\mathbf{A}$	reduced incidence matrix of balance scheme
$\mathbf{A}_1$, $\mathbf{A}_2$	reduced incidence matrices of directly measured and directly unmeasured quantities, respectively
e_i; $\mathbf{e}$	error in i-th measured quantity; vector of quantities e_i
$\mathbf{F}$	covariance matrix of measurement errors
$\mathbf{F}_r$	covariance matrix of residuals of balance equations
$\mathbf{F}_{\hat{x}}$	covariance matrix of adjusted values
$\mathbf{F}_v$	covariance matrix of adjustments
$h(\cdot)$	rank of matrix
I	number of streams
I_1, I_2	numbers of directly measured and directly unmeasured streams, respectively
J	number of nodes (including the environment node)
$\mathbf{k}$	vector of Lagrangian multipliers
Q	weighted sum of squares of adjustments (Eqs (*12*) and (*13*))
r_j; $\mathbf{r}$	residual of j-th equation; vector of quantities r_j
$u_{1-\alpha/2}$	$100(1 - \alpha/2)$ percentile of standard normal distribution
v_i; $\mathbf{v}$	adjustment of i-th measured quantity; vector of quantities v_i
w_i	normalized adjustment (Eq. (*26*))
x_i; $\mathbf{x}$	flow through i-th stream; vector of quantities x_i
$\mathbf{x}_1$, $\mathbf{x}_2$	vectors of directly measured and directly unmeasured flows, respectively
z_j	normalized residual (Eq. (*28*))
α	significance level of test
β	power of test
$\chi^2(\nu)$	random quantity with chi-square distribution and ν degrees of freedom
$\chi^2_{1-\alpha}(\nu)$	$100(1 - \alpha)$ percentile of $\chi^2(\nu)$ distribution
ν	number of degrees of freedom
σ_i, σ_{rj}, σ_{vi}	standard deviations of quantities e_i, r_j, and v_i, respectively

Superscripts and overlays

$\sim$	measured value
$\wedge$	adjusted value

REFERENCES

1. Madron F., Výborný J., Vaněček V.: Chem. Prům. *29*, 609 (1979).
2. Hlaváček V., Václavek V., Kubíček M.: *Mass Balance and Simulation Calculations for Complex Chemical Technology Processes*. Academia, Praha 1979 (in Czech).
3. Ripps D. L.: Chem. Eng. Progr. Symp. Ser. 55, *61*, 8 (1965).
4. Nogita S.: Ind. Eng. Chem. Proc. Des. Develop. *11*, 197 (1972).
5. Kubáček L., Pázman A.: *Statistical Methods in Measuring*. Veda, Bratislava 1979 (in Slovak).
6. Madron F., Výborný J., Novotný P.: VZ-E-1021 Report, Research Institute of Inorganic Chemistry, Ústí nad Labem 1980 (in Czech).
7. Mah R. S. H., Stanley G. M., Downing D. M.: Ind. Eng. Chem., Process. Des. Develop. *15*, 175 (1976).

Translated by M. Škubalová.