DETECTION OF GROSS ERRORS IN CHEMICAL ENGINEERING MEASUREMENTS

František MADRON

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

Received April 7th, 1981

Detection of gross errors in chemical engineering measurements is studied, based on statistical analysis of redundant data. The error of third type has been defined as the unacceptably large error of a quantity which is the aim of measurement, while the gross error of measurement is not detected. The method of classification of directly measured quantities into disjoint subsets is described according to possible effective detection of occurence of gross errors of measurements by statistical analysis of measured data.

In chemical engineering measurements it is important to check the measured data as concerns the possible occurence of gross errors and systematic errors of measurements. Development of effective methods in this respect has a special significance at treatment of data obtained at industrial measurements when it is usually very difficult to obtain a completely reliable data. This problem is especially actual with increasing use of automatic measuring systems and process computers whose function might be adversely affected by gross or systematic errors of measurements.

Number of studies has been published proposing procedures for solution of these problems[1-8]. But it is not known what is the effectiveness of these methods at solution of practical problems. The aim of this contribution is to formulate possibilities and limitations of an analysis of measured data as concerns occurrence of gross errors of measurements.

The process of elimination of gross errors from measurements can be divided basically into three phases. The first one is detection of presence of errors where it is determined whether the gross errors are present. If this is so localisation of these errors follows which represents finding probable sources of errors. Elimination then means removal of these errors. Here is systematically considered only the first one of these steps, *i.e.* detection of presence of gross errors.

The procedure is based on comparison of measured data with the earlier information available on the studied process. By calculation from the measured data can be *e.g.* determined values of parameters which are considered as unprobable or impossible (extreme values of transfer coefficients, negative concentration *etc.*).

The most important method in the analysis of measured data as concerns the oc-

curence of gross errors in measurements is their confrontation with the exactly valid mathematical models. Under this term is most frequently considered the balance equations, sometimes it is also possible to include here the reliable thermodynamic data (phase equilibria) and according to conditions also other information. Next only cases when the mathematical model is represented by the system of algebraic and transcendent equations is considered.

The analysis of measured data is based basically on three types of information. These are at first those which can be called the theory of measured object. These information can be mostly expressed in the form of equations and inequalities of the mathematical model (balance and definition equations, callibration curves of measuring instruments *etc.*). Part of this prior information are also the in advance known constants (physical and mathematical constants, physico-chemical properties of compounds *etc.*). Second source of information are directly measured (primary) quantities obtained on basis of the measurement. These values are affected by errors in measurements and their treatment is based on the third group of information concerning the character and magnitude of measuring errors which are called the model of errors of measurement. The mathematical model and model of errors together form the stochastic model of the measured object.

The errors can be divided into two groups. Into the first one belong errors of the mathematical model. They are caused by imperfect knowledge of the measured object. This might concern *e.g.* escape (loss) of materials (the loss stream is not respected in balance equations), in chemical reaction the origin of unknown secondary product *etc.*

Into the second group belong errors in the model of errors. It is assumed *e.g.* that the errors in measurements are realisations of random variables with multidimensional normal distribution and given covariance matrix. The existence of a large gross error then contradicts this model. This type of errors is the most frequent one and when we are considering gross errors of measurements we are usually considering just these deviations from the assumed model of errors.

Before we begin with the method of error detection let us formulate another two assumptions on which the next analysis will be based.

First it is the assumption on knowledge of the error model. Mostly it is assumed that the errors have a normal distribution with zero mean value and with the known covariance matrix. Assumption on knowledge of the covariance matrix might seem too strict. But it is necessary to realize that as long there is no model of random errors of measurements so long it is not possible to define what are the gross errors and there it makes no sense to search for them. The method of estimation of the covariance matrix is given *g.e.* in the study[4].

Second is the assumption that in the mathematical model represented by a system of equations a certain part of measured data is redundant (in the absence of errors of measurements they could be uniquelly calculated by solving the equations of the

mathematical model). Redundancy of data is the necessary condition for application of the method of detection of gross errors of measurements described in this study.

THEORETICAL

In the case of redundant measurements, the equations of the mathematical model are not exactly satisfied by measured data. It is said, that the data are inconsistent. Thus the question is then solved whether it is possible to explain the inconsistency of data within the frame of the assumed model of errors or whether other errors are present on which no information is available. In the first case it is said that inconsistency is insignificant, in the second that it is significant. Determination of significant inconsistency is identical with detection of unknown errors. It is necessary to mention that the term data inconsistency has no relation to the concept of consistency of statistical estimation[12].

Stochastic Model

Let us consider the mathematical model expressed by the system of J linear equations between I measured quantities x, and K not measured quantities y, in the form

$$f + Ax + By = 0\,, \tag{1}$$

where vectors $f\ (J.1)$ and matrices $A\ (J.I)$ and $B\ (J.K)$ are known. Vector x is known only approximately from measurements while there holds

$$x^{+} = x + e\,, \tag{2}$$

where x^{+} is the vector of measured values and e is the vector of errors of measurements. It is assumed that the errors e are random variables with the zero mean value and with the known positive definite covariance matrix F.

Next the case is considered where the vector x includes more measurements than is necessary for determination of not measured quantities y from Eqs (*1*).This assumption can be expressed mathematically by the next equations and inequalities[9]

$$r(A, B) = J\,; \quad r(B) = K\,; \quad I > J - K > 0\,, \tag{3}$$

where $r(.)$ is the rank of matrix.

Identification of Parameters of the Mathematical Model

In the presence of redundant measurements the measured values mostly do not fit

exactly the equations of the mathematical model (there does not exist such vector y that Eqs (*1*) are with the measured data exactly satisfied). Thus an effort is made to obtain adjusted values $\hat{x}$ defined by relation

$$\hat{x} = x^{+} + v\,, \tag{4}$$

where v is the vector of so called adjustments. Adjusted values must exactly satisfy equations (*1*) and adjustments must be in a certain sense minimal. Mostly is minimalised the quadratic form of adjustments

$$Q_{\min} = v^{\mathrm{T}} F^{-1} v\,. \tag{5}$$

Solution of this problem is obtained after introduction of the Lagrange multipliers k as the solution of the next system of equations[9]

$$\left(\begin{array}{c|c} AFA^{\mathrm{T}} & B \\ \hline B^{\mathrm{T}} & 0 \end{array}\right) \cdot \left(\begin{array}{c} k \\ \hline \hat{y} \end{array}\right) + \left(\begin{array}{c} f + Ax^{+} \\ \hline 0 \end{array}\right) = 0 \tag{6}$$

$$v = FA^{\mathrm{T}} k \tag{7}$$

$$\hat{x} = x^{+} + v\,. \tag{8}$$

The system of equations (*6*) can be solved with respect to vectors k and $\hat{y}$ either by use of inversion of matrix on the left hand side of Eq. (*6*) or it is possible to use the procedure proposed in the study[9] where the existence of the zero submatrix in this matrix is used and which requires inversion of matrices of smaller dimensions. Adjustments v and smoothed values $\hat{x}$ are then obtained by substituting the Lagrange multiplicators k into Eq. (*7*).

An important generalisation of the linear model is the quasilinear model. Quasilinear model differs from the linear one so that instead of the system of equations (*1*) there are J equalities

$$f(x, y) = 0\,, \tag{9}$$

where the elements of column vector f are generally nonlinear functions of elements of vectors x and y. Next let us assume: *1*) Functions $f_j(x, y)$; $j = 1, \ldots, J$ have continuous second partial derivatives according to x_i and y_k; $i = 1, \ldots, I$, $k = 1, \ldots, K$ *2*) The values of $x^{(0)}$ and $y^{(0)}$ are known which are so close to the actual values of x and y that in the Taylor expansion

$$f(x, y) = f(x^{(0)}, y^{(0)}) + C\,\Delta x + D\,\Delta y + \ldots, \tag{10}$$

where

$$x = x^{(0)} + \Delta x$$

$$y = y^{(0)} + \Delta y$$

$$C = [C_{ji}] = [\partial f_j/\partial x_i]$$

$$D = [D_{jk}] = [\partial f_j/\partial y_k],$$

the terms of second and higher orders can be neglected. Matrices of partial derivatives are then evaluated in points $x^{(0)}$ and $y^{(0)}$. In this way is the system of Eqs (*9*) transformed into the linear form (*1*) with solution given by the system of Eqs (*6*) to (*8*).

By the described procedure are (at the assumption of normal distribution of errors) obtained the maximum likelihood estimates of vectors x and y. Important is that the quadratic form of adjustments Q_{min} is a realisation of random variable with χ^2 distribution with $(J - K)$ degrees of freedom[9].

Quantities y which were estimated by this method are sometimes called *quantities measured indirectly* on the contrary to those directly measured ones x. The directly measured quantities x can be further divided to those for which it is possible due to statistical adjustment to reach their correction (adjustable quantities) and thus also reduction of their variance and to nonadjustable quantities which have their corrections always equal to zero. The case has been eliminated from our considerations when some elements of the vector y could not be estimated (not determinable not measured quantities)[13].

The made classification of quantities is based only on properties of equations of the mathematical model. If also the error model is taken into consideration a considerably wider spectrum of quantities is obtained at detailed evaluation. Frequently there is found that the indirectly measured quantities y have so unfavourable statistical properties (large variances) that from the practical point of view they can be considered as not determinable. Similar conclusion may be reached also in the case of adjustable quantities when reduction of variance of the measured quantity at smoothing of data is negligible. In practice it is thus advantageous to classify the quantities from the point of view of their statistical properties (variances, covariance matrices).

Analysis of Measured Data

How wide is inconsistency of measured data can be judged basically from two types of random quantities. First, it is the vector of adjustments v and second the vector of residuals of equations r defined by relation

$$r = f + Ax^{+} + B\hat{y}. \qquad (11)$$

Vectors v and r are not mutually independent. Vector r can be uniquely expressed as a function of vector of corrections by relation

$$r = Av \tag{12}$$

and thus it suffices to consider only the vector v.

Next problem is the choice of function of adjustments, so called statistics which is used for testing of the hypothesis whether the model of errors is correct (hypothesis H_0).

At first it is necessary to decide on the number of statistics. There are two limiting possibilities. First one is to use as statistics all elements of vector v (multidimensional statistics) second one is to calculate from vector v a single statistics (single unidimensional statistics). Results of tests will differ by the magnitude of probability in both these possibilities of errors of Ist and IInd type (error of Ist type is rejection of hypothesis in the case of its correctness. The error of second type is acceptance of hypothesis in the case of its incorrectness). With unidimensional statistics there is the danger that the effect of isolated gross error will be covered by the effect of larger number of random errors with which is related the increase of probability of errors of IInd type. In the case multidimensional statistics is used probability of Ist type errors will rise.

In literature number of statistics has been proposed for detection of gross errors of measurement. From unidimensional statistics it is possible to mention the linear form of corrections[2], quadratic form of corrections[4] (quantity Q_{min}) and quadratic form of residuals[8]. To multidimensional statistics belong the vector of residuals[6], vector of residuals transformed to the vector of not correlated random variables[7] and vector of adjustments[7]. From the earlier made study[10,11] resulted that the use of linear form of corrections[2] and of vector of residuals[8] are less effective than other statistics which are comparable from this point of view.

In the following part the detection of gross errors is considered. The statistics Q_{min} (Eq. (5)) is used, whose application is the most simple one. The hypothesis H_0 (hypothesis on absence of gross error) is rejected at validity of the inequality

$$Q_{min} > \chi^2_{1-\alpha}(J - K)\,, \tag{13}$$

where $\chi^2_{1-\alpha}(J - K)$ is $100(1 - \alpha)$ percent quantile of χ^2 distribution with $(J - K)$ degrees of freedom.

Testing of the hypothesis H_0 from the standpoint of effectiveness at detection of gross error is usually characterised by operative characteristics of tests[12] which are dependences of probability of the error of IInd type on magnitude of gross error. But it is necessary to realize that non-rejection of the hypothesis H_0 informs us only that the data and mathematical model are not in contradiction. But it says nothing

on how good results can be expected on basis of the made measurement. Answering this question practically requires introduction of the term target quantity.

As the *target quantity* will be considered the quantity for whose determination is performed the whole measurement (yields of chemical reactions, parameters of mathematical models *etc.*). Next is introduced the term gross error of target quantity as the not acceptable deviation of the value obtained on basis of measurements from the actual (correct) value. The magnitude of this not acceptable deviation can be determined *e.g.* on basis of the requirement on accuracy of measurement of the target quantity. In the case the actual error of the target quantity is larger than the specified value it is said that the target quantity is affected by gross error.

The next phenomena can be then defined:

Phenomenon B: gross error of measurement was not detected (hypothesis H_0 was not rejected);

phenomenon A: there is a gross error in the target quantity;

phenomenon $C = B \cap A$: gross error of measurement has not been detected but with the target quantity a gross error has been made. From the practical side of consequences it is necessary to avoid especially the phenomenon C which means that the result is affected by a gross error but we are considering it as correct. The phenomenon C will be further on denoted as the error of IIIrd type. The error of IIIrd type differs from the error of IInd type by not taking into account cases when the existing gross error of measurement has not been detected but no gross error in the target quantity has occured.

Example: Detection of Gross Error at Measuring of the Degree of Absorption

Let us consider the absorption unit shown in Fig. 1, where the measured quantities are also denoted. The inlet gas stream *1* containing two absorbed components and an inert gas is led into contact with pure absorption liquid (stream *4*). Vapour concentration of the absorbent in the gaseous phase can be neglected.
The primary quantities are measured:

x_{ij} $i = 1, 2$; $j = 1, 3$, mole fraction of i-th component in j-th stream (–);
c_{i2} $i = 1, 2$, concentration of i-th component in the stream *2* (kmol m^{-3});
V_2 volumetric flow rate of stream *2* (m^3 h^{-1})
Δp pressure drop on the orifice for measurement of flow rate of stream *1* (Pa);
T temperature at the orifice for measurement of flow rate (K);
p pressure of gas at the orifice for flow measurement (Pa).

As another primary quantity which is the source of error can be considered the constant of measuring orifice k. Molar flow rate of stream *1* (N_1) is given by relation

$$N_1 = k(\Delta p P/(h_n T))^{1/2}, \tag{14}$$

where h_n is the density of stream *1* at normal conditions given as function of normal densities of compounds *1*, *2* and of inert one (component *3*), given by relation

$$h_n = 1{\cdot}34x_{11} + 1{\cdot}27x_{21} + 1{\cdot}29x_{31} \tag{14a}$$

For measured and not measured quantities in total 6 independent equations can be written:

1) flow rate of stream *1* expressed by Eq. (*14*)

2) balance of component *1*

$$N_1x_{11} + V_2C_{12} + N_3x_{13} = 0 \tag{15}$$

3) balance of component *2*

$$N_1x_{21} + V_2C_{22} + N_3x_{23} = 0 \tag{16}$$

4) equation for concentration in the stream *1*

$$x_{11} + x_{21} + x_{31} - 1 = 0 \tag{17}$$

5) equation for concentration in the stream *3*

$$x_{13} + x_{23} + x_{33} - 1 = 0 \tag{18}$$

6) balance of inert

$$N_1x_{31} + N_3x_{33} = 0 \tag{19}$$

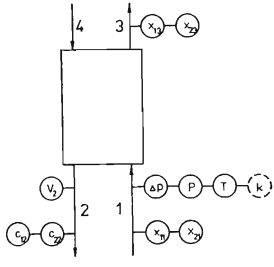

FIG. 1
Measuring absorption unit

In the problem there appear four non-measured quantities (flow rates of gaseous streams N_1 and N_3 and concentrations of inert x_{31} and x_{33}). The number of equations is 6 and there exist 2 redundant measurements.

As the target quantities are considered degrees of absorption of compounds *1* and *2* defined by relations

$$s_1 = -V_2 C_{12}/N_1 x_{11} \qquad (20)$$

$$s_2 = -V_2 C_{22}/N_1 x_{21} . \qquad (21)$$

When the calculation is started with correct values given in Table I the correct values of target quantities are determined as $s_1 = 0{\cdot}4968$ and $s_2 = 0{\cdot}8907$. By the methods of propagation of errors[12] the standard deviations of absorption degrees can be also calculated from adjusted values $\sigma_{s1} = 0{\cdot}0122$ and $\sigma_{s2} = 0{\cdot}0046$.

Effectiveness of detection of gross errors has been studied by simulation on the computer by the Monte Carlo method. To the values of primary quantities given in Table I, which represented the correct values were added randomly generated uncorrelated errors with normal distribution and zero mean values and variance coefficients (relative standard deviations) given in Table I. Moreover to one primary quantity was added a grooss error of constant value. One hundred of such simulated measurements were generated for each primary quantity and magnitude of gross error. Obtained data were smoothed while Eqs (*14*) to (*19*) were linearized into the form of Eq. (*10*).

The boundaries for gross errors of the target quantity were chosen as 1·96 multiples of their standard deviations, *i.e.* 0·0239 with s_1 and 0·0090 with s_2. Gross error with the target value

TABLE I

Correct values of directly measured quantities

Measured quantity	Number	Correct value	Coefficient of variance
x_{11}	1	0·03160	0·03
x_{21}	2	0·00566	0·03
x_{13}	3	0·01600	0·04
x_{23}	4	0·00062	0·04
C_{12}	5	4·013	0·02
C_{22}	6	1·272	0·02
V_2	7	0·320	0·02
Δp	8	2 013	0·03
T	9	284	0·02
P	10	99 300	0·02
k	11	−0·1091	0·01

was thus detected when the values of absorption degrees were out of the intervals 0·4968 ± 0·0239 or 0·8907 ± 0·0090.

The obtained results are demonstrated in Figs 2 to 4. In Fig. 2 are empirically determined operative characteristics of testing the hypothesis H_0. These are relative frequencies f of the not observed gross error for individual primary quantitites in dependence on magnitude of gross error e_g. Gross errors of primary quantities are here expressed as multiples of standard deviations of corresponding primary quantities. It can be seen from this figure that the operative characteristics differ considerably for individual primary quantities. While *e.g.* for primary quantities 1 and 2 (Table I) are the operative characteristics favourable, for quantities 4 there holds that the gross error probably will not be found even if it is actually great. This unfavourable property is typical of those quantities which do not have significant effect on statistics Q_{min}.

Extreme in this respect are quantities with zero corrections (measured nonadjustable quantities) with which it is not possible to detect the gross error. Moreover it is possible to state that the operating characteristics of the test have a similar form (they are decreasing monotonously) and practical effectiveness of detection of gross errors from them is not obvious.

In Figs 3 and 4 are given relative frequencies of occurence of the case when there is a gross error in the target quantity but this is not detected by the analysis of in-

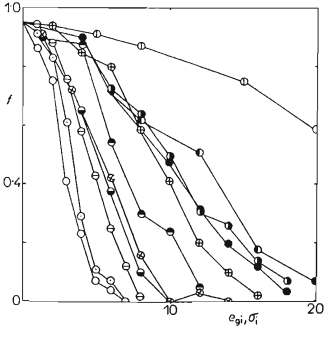

FIG. 2

Operative characteristics of the test. ○ x_1; ⊙ x_2; ⊖ x_3; ⓘ x_4; ⊕ x_5; ⊗ x_6; ◒ x_7; ◓ x_8; ◐ x_9; ◑ x_{10}; ● x_{11}

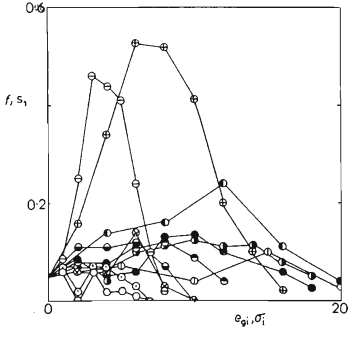

FIG. 3

Relative frequency of origin of IIIrd type errors for quantity s_1

consistency (error of IIIrd type). In contrast to operating characteristics the dependence of error of IIIrd type is passing through a maximum. There results from Figs 3 and 4 that with a limited number of primary quantities (quantities 3, 5 and 9 with s_1 and quantity 4 with s_2) there might originate an undetected gross error with the target quantity at a relatively large probability. From other primary quantities this danger practically does not exist.

The practical conclusion which can be made on basis of this study is that statistical analysis of inconsistency of measured data together with independent checking of measurements No 3, 4, 5 and 9 (parallel independent measurements) can practically eliminate possibility that some of target values could be affected by a gross error.

DISCUSSION AND CONCLUSIONS

Statistical analysis of inconsistency of measured data is an effective method of detection of gross errors of measurements. Their effectiveness at solving practical problems can be best characterised by the dependence of probability of error of IIIrd type on magnitude of gross errors of individual primary quantities.

The set of directly measured quantities is in this case divided into two subsets. In the first one are the primary quantities with which there is a small probability of an error of IIIrd type (*e.g.* smaller than 0·1) at an arbitrary large gross error of measurements. With primary quantities belonging into this subset the statistical analysis of inconsistency of measured data is sufficient for verification of the assumption that gross errors of measurements are not present or do not have a significant effect on the value of target quantity.

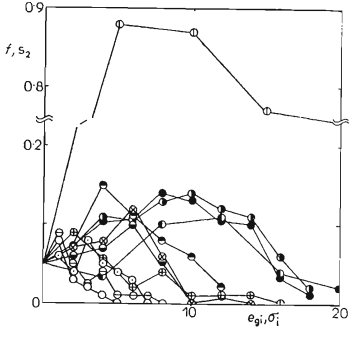

FIG. 4
Relative frequency of origin of IIIrd type errors for quantity s_2

In the second subset are then the remaining primary quantities. With them the analysis of inconsistency of measured data is not guaranteeing detection of gross errors which could have the consequence in gross error of target quantity. Into this subset belong *e.g.* those measured nonadjustable quantities which have an effect on target quantities. If we want to eliminate consequences of gross errors of measurements with the quantities of this subset an independent checking of individual measuring devices and methods is necessary.

The possibility of detection of measurements affected by gross errors should be taken into consideration at preparation of measurements and at location of measuring points so that the statistical analysis of inconsistency of data would verify whether the target quantities are correct. This concerns especially measurement of those quantities whose checking by direct method is difficult. Effective detection of gross errors is the assumption of their succesful localisation and elimination from the measuring process.

LIST OF SYMBOLS

$\boldsymbol{A}$	matrix ($J.I$) defined by Eq. (*1*)
$\boldsymbol{B}$	matrix ($J.K$) defined by Eq. (*1*)
C_{i2}	concentration of *i*-th compound in the stream 2 (kmol m^{-3})
$\boldsymbol{f}$	vector ($J.1$) defined by Eq. (*1*)
I	number of directly measured (primary) quantities
J	number of equations in system (1)
k	proportionality constant of the measuring orifice
K	number of not measured quantities (vector $\boldsymbol{y}$)
N_1, N_3	flow rates of streams 1 and 3 (kmol h^{-1})
Δp	pressure drop on orifice (Pa)
P	pressure (Pa)
$Q_{\min}$	quadratic form of adjustments
$\boldsymbol{r}$	vector ($J.1$) of residuals
s_1, s_2	absorption degree of components *1* and *2*
T	temperature (K)
$\boldsymbol{x}$	vector ($I.1$) of primary quantities
x_{ij}	mole fraction of *i*-th compound in *j*-th stream
$\boldsymbol{v}$	vector ($I.1$) of adjustments
$\boldsymbol{y}$	vector ($K.1$) of indirectly measured quantities
$\chi^2_{1-\alpha}(J-K)$	$100(1-\alpha)$ percent quantiles of χ^2 distribution with $(J-K)$ degrees of freedom
σ	standard deviation

Superscripts

$+$	measured quantity
$\wedge$	adjusted quantity
-1	inverse matrix
T	transposed matrix

REFERENCES

1. Rips D. L.: Chem. Eng. Progr. Symp. Ser. *61*, 8 (1965).
2. Nogita S.: Ind. Eng. Chem. Process Des. Develop. *11*, 197 (1972).
3. Václavek V., Vosolsobě J., Plesar P., Fiedler M.: Chem. Prům. *25*, 170 (1975).
4. Madron F., Veverka V., Vaněček V.: AIChE J. *23*, 482 (1977).
5. Madron F.: This Journal *45*, 32 (1980).
6. Mah R. S. H., Standley G. M., Downing D. M.: Ind. Eng. Chem., Process Des. Develop. *15*, 175 (1976).
7. Knepper J. C., Gorman J. W.: AIChE J. *26*, 260 (1980).
8. Romagnoli J. A., Stephanopoulos G.: Chem. Eng. Sci. *35*, 1067 (1980).
9. Kubáček L., Pázman A.: *Štatistické metódy v meraní*. Veda, Bratislava 1979.
10. Britt A. I., Leucke R. H.: Technometrics *15*, 233 (1973).
11. Madron F.: National CHISA Conference, Vysoké Tatry 1980.
12. Himmelblau D. M.: *Process Analysis by Statistical Methods*. Wiley, New York 1970.
13. Hlaváček V., Václavek V., Kubiček M.: *Bilanční a simulační výpočty složitých procesů chemické technologie*. Academia, Prague 1979.

Translated by M. Rylek.