L.V.D. : a software for mass balance equilibration and data validation

Didier Maquin, José Ragot

L.V.D.: a software for mass balance equilibration and data validation

D. Maquin, J. Ragot
Laboratoire d'Automatique et de Recherche Appliquée, Vandoeuvre (F)

Summary :

This paper deals with the problem of data validation in large scale steady state linear systems which are described by algebraic equations. The conditions for observability and redundancy are defined and a procedure is presented for classifying the system variables into observable, unobservable, redundant and no redundant variables. Reconciliation of the redundant process measurements is developped using a recurrent estimation technique which is well suited in real time data processing. Statistical tests based on the residuals or imbalances of the model constraints either individually or collectively allow to identify the sources or locations of the gross errors. The authors then present a software called L.V.D. as "Logiciel de Validation de Données" (Data Validation Software), which take into account all the major difficulties in the field of data validation for linear systems.

Zusammenfassung :

Die Störungen, die die Messketten eines Systems beeinflussen, erschweren die Benutzung der Messergebnisse; eine vorherige Verarbeitung ist nötig. Die Berichtigungsmethode für die Messdaten, die hier vorgestellt wird, benutzt den Überschuss der Informationen, der aus den linearen Bilanzgleichungen des Systems hervorgeht. Da dieses System grossdimensional ist, schlagen wir einen sequentiellen Lösungsweg vor, der einen originalen Algorithmus verwendet.

In der vorliegenden Arbeit werden Ausgleichsrechnungen mit fehlenden Messungen betrachtet. Das nötigt eine Klassifizierung der Variablen in wahrnehmbare oder nicht wahrnehmbare, überfüllte oder nicht überfüllte Variablen.

Ein Teil der Methoden der Materialbilanzierung bezieht sich auf Messungsfehler. Das spezifische Programm L.V.D. (für Logiciel de Validation de Données) ermöglicht eine Lösung der vorherigen erörterten Punkte.

Résumé :

Cette communication traite du problème de validation de données des systèmes linéaires de grande dimension opérant en régime statique. Ces systèmes sont décrits par des équations algébriques. Les conditions d'observabilité et de redondance sont définies et une procédure de classification des variables en variables observables ou non observables, redondantes ou non redondantes est présentée. La réconciliation des mesures redondantes du procédé est réalisée à l'aide d'un algorithme récursif d'estimation qui est bien adapté au traitement temps réel. Des tests statistiques basés sur l'analyse individuelle ou collective des écarts de bouclage de bilan des équations de redondance permettent d'identifier et de localiser les mesures entachées de biais. Les auteurs présentent ensuite un logiciel appelé L.V.D., Logiciel de Validation de Données, qui prend en compte l'ensemble des difficultés évoquées en ce qui concerne la validation de données des systèmes linéaires.

1 - INTRODUCTION

Industrial process control requires data acquisition at the very least cost that must be able to present pertinently the state of such a process. Before using these data, the user must take very elementary precautions amongst which the information coherence test is essential. Thus, data reliability is of great significance if these data are used effectively in process monitoring for operation optimization, control or identification. Measurements having undetected gross, random or biased errors result in false control of a process. Data validation is therefore situated between the process measurements acquisition and the decision to be taken. Data reliability can be secured through a balance equilibration. The measurements should be

reconciled in some "best" sense to obey conservation laws and other constraints that are required to be inforced.

This reconciliation poses several problems amongst which are the following :

- a model is no more than an approximation of the actual process. Its structure and parameters may have been selected or estimated from restrictive hypothesis. In addition, the models used are in general non linear,

- measurements carried out on the process variables do not present an accurate picture of the actual magnitudes because they are subject to errors (randoms errors or bias),

- for reasons of cost convenience or technical feasability, not every variables in a process is measured.

This study is limited to the case of linear systems operating under steady state conditions. Even though this hypothesis seems to be restrictive, this class of systems is frequently encountered in industrial applications. Moreover, the process models are exact as they are based on the material conservation laws. Process measurements are subject to two essential types of errors : random errors which are commonly assumed to be independantly and normally distributed with zero mean and gross errors which are caused by nonrandom events such as leaks or inadequate accounting of departures from steady state operations as well as by measurement biases and malfunctionning instruments.

2 - PROBLEM FORMULATION - METHODOLOGY

The structural information in a plant can be conveniently represented by a direct graph /1/. The nodes of which represent the process units as reactors, tanks, distillation columns, whilst the arcs represent streams of circulating matter. The mathematical model originated from mass conservation laws and in the linear case, is written under the exact structural form :

$$M X^* = 0 \tag{1}$$

where $M \in \mathbb{R}^{n.v}$ is the incidence matrix of the process graph with n the number of nodes and v the number of arcs,

$X^* \in \mathbb{R}^v$ is the vector of "true" values (unreachable to the measure).

The measurements are given by :

$$X = X^* + \varepsilon \tag{2}$$

where $\varepsilon \in \mathbb{R}^v$ is the vector of measurement errors which is assumed to be normally distributed with zero mean and a known diagonal variance matrix V.

With these hypothesis, the maximum likelihood estimator reduces to the least square estimator subject to the constraints of the model :

$$\min \ \Phi = \| \hat{X} - X \|^2_{V^{-1}}$$

$$\text{subject to the constraints} \quad M \hat{X} = 0 \tag{3}$$

The solution to this least square estimation problem is expressed as :

$$\hat{X} = P X \tag{4}$$

where

$$P = I - V M^T (M V M^T)^{-1} M \tag{5}$$

We shall refer to P as projection matrix.

3 - CLASSIFICATION OF THE VARIABLES - OBSERVABILITY

In the general case, not all the variables of a process are measured. Then, it is necessary to reduced the set of balance equations such that the reduced set involves no unmeasured flow rate and a maximum number of measured flow rates. This work has been previously accomplished by Vaclavek /2/, Mah /1/ or Romagnoli /3/. Basically, the proposed algorithms eliminate balances involving unmeasured feed or product flow rates and merge two balances with a common unmeasured flow rate. These graph oriented algorithms which are employed to "reduce the balance scheme" and to identify the redundant measurements, have a matricial interpretation which generally reduce to extract a regular part of the incidence matrix of the process graph. This can be done by using an echelon form transformation /4/.

4 - INCONSISTENCY TEST OF DATA

The procedure outlined in the previous paragraph have been found to work well for data containing only small errors. When the set of plant data contains gross erors, the least square procedure distributes the residual of the balance amongst all the measurements and transforms the data as a whole highly unreliable set. It is then necessary to examine the residuals of the reduced balance scheme.

Assume M to be the matrix of this reduced balance scheme; the R vector of imbalance is generated according to the following equation :

$$M\ X = R \tag{6}$$

Under the previous hypothesis relative to the measurements errors, one can demonstrate that the R vector is normally distributed with zero mean and a variance matrix V_r which is equal to $M\ V\ M^T$. Notice that V_r is not a diagonal matrix, then the different components of R are not independant.

In order to compare the components of the R vector, let us define a new vector R_n such as each component $R_{n(i)}$ is defined by :

$$R_{n(i)} = \frac{R(i)}{\sqrt{V_r(i,i)}} \tag{7}$$

Each component of R_n vector has a normal distribution with a zero mean and a unity variance. Then, a statistical test criterion of data inconsistency can be introduced. From a cumulative normal distribution table, for example, the probability of $R_n(i)$ being in the interval of -2 to 2 is read to be 0.95. Therefore, when $|R_n(i)| > 2$, we might say that the inconsistency is significant with a probability of 0.05. With this error probability it is predicted that there exists a systematic unbalance, such as an unsteady state in the system, unexpected efflux or influx, or gross error in measurements /5/.

5 - STATISTICAL ANALYSIS OF THE RESIDUALS

Statistical analysis of the residuals allows to answer to the following questions :

- are the data validation results, acceptable ?
- if they are not, what and where are the suspicious measurements ?

The answer to the second question corresponds to detect and locate bad data.

Now, let us define the residuals vector :

$$E = X - \hat{X} = V\,M^T (M\,V\,M^T)^{-1}\,M\,X \tag{8}$$

Under the null hypothesis that the measured value for all the stream contain no systematic error, E has a normal distribution with a zero mean and a variance matrix equal to $\hat{V} - V$.

As for the imbalance, let us define a normalized residuals vector E_n such that :

$$E_n(i) = \frac{E(i)}{\sqrt{V(i,i) - \hat{V}(i,i)}} \quad (9)$$

Each component of E_n can be compared with a critical test value E_c. $|E_n(i)| > E_c$ denotes stream i as a bad stream. A recommended critical value is 2.

The global quality of the adjustement can also be appreciate by the analysis of the minimum value of the objective function :

$$\Phi_r = \| \hat{X} - X \|^2_{V^{-1}} = E^T V^{-1} E \quad (10)$$

Since each of the v element of E is distributed according to the Gaussian distribution, we conclude that the random variable Φ_r, being the sum of squares is best characterized by a chi square distribution. Notice that only n componants amongst v are independant. The degree of freedom of this chi square distribution is the same as the number of constraints equations. Statistical analysis of this value will inform about the quality of the adjustment.

6 - THE SOFTWARE L.V.D.

Data validation requires a simple tool which take into account all the difficulties previously described ; in this way, L.V.D. has been developped in the "Laboratoire d'Automatique et de Recherche Appliquée" /6/.

Experience has proved that this software was sufficiently efficient to employ it in industrial environment. The results presented here, illustrate somme aspects of L.V.D.

A network representation of a chemical process consisting in 20 nodes and 53 streams is shown in Figure 1. The locations of the sensors are indicated on this network by a cross.

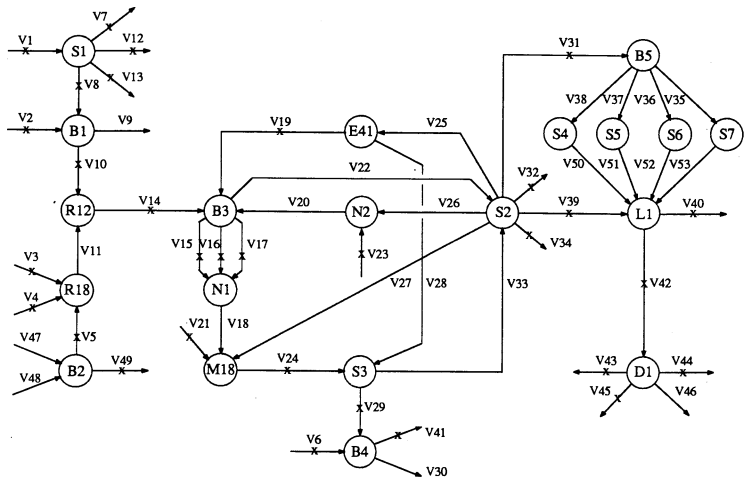

Figure 1 : Network representation of a chemical process

The classification issue from observability allows to determine 4 redundancy equations which are shown in Table I, as well as 7 equations allowing the deduction of streams V9,V11,V18,V27,V30,V46 and V48, 13 streams are indeterminable (V20,V22,V25,V26,V28,V35,V36,V37,V38,V50,V51,V52 and V53).

Equation	Streams					
1	+	V40	V42			
	-	V39	V31			
2	+	V14	V21	V23		
	-	V29	V34	V39	V31	V32

Equation	Streams					
3	+	V10	V3	V4	V5	
	-	V14				
4	+	V1				
	-	V7	V12	V13	V8	

Table I : redundancy equations

The measured values and their accuracy are summarized in Table II. Notice some negative values relative to pseudo streams which have been added on some nodes for taking into account change of the level in big tanks (streams V7,V32,V34,V43 and V49). Indeed, these changes can be assimilated to pseudo flow rates.

Name of the stream	Value of the measurement	Kind of weight	Value of the weight
V1	35,786	% Measur.	6,00 %
V2	1,919	% Measur.	8,00 %
V3	0,384	% Measur.	8,00 %
V4	0,000	% Measur.	8,00 %
V5	412,675	% Measur.	6,00 %
V6	38,627	% Measur.	6,00 %
V7	-5,202	% Measur.	12,00 %
V8	26,491	% Measur.	6,00 %
V10	27,590	% Measur.	6,00 %
V12	8,503	% Measur.	8,00 %
V13	4,860	% Measur.	6,00 %
V14	432,210	% Measur.	8,00 %
V15	24,561	% Measur.	12,00 %
V16	24,564	% Measur.	12,00 %
V17	0,000	% Measur.	12,00 %
V19	14,402	% Measur.	8,00 %

Name of the stream	Value of the measurement	Kind of weight	Value of the weight
V21	0,177	% Measur.	7,00 %
V23	0,111	% Measur.	7,00 %
V24	50,787	% Measur.	8,00 %
V29	48,696	% Measur.	8,00 %
V31	121,540	% Measur.	7,00 %
V32	-13,831	% Measur.	12,00 %
V34	-17,784	% Measur.	12,00 %
V39	404,678	% Measur.	7,00 %
V40	50,800	% Measur.	8,00 %
V41	89,574	% Measur.	8,00 %
V42	366,897	% Measur.	8,00 %
V43	26,093	% Measur.	12,00 %
V44	218,097	% Measur.	7,00 %
V45	222,454	% Measur.	7,00 %
V47	304,698	% Measur.	7,00 %
V49	34,373	% Measur.	12,00 %

Table II : measured values

Stream	Measurement	Estimation	Standard deviation	Estimated Std. Dev.	Correction
V1	35,786	36,859	1,074	0,732	3,00 %
V2	1,919	1,919	0,077	0,077	0,00 %
V3	0,543	0,543	0,022	0,022	0,00 %
V4	0,000	0,000	0,000	0,000	0,00 %
V5	412,675	429,712	12,380	7,270	4,13 %
V6	38,827	38,827	1,165	1,165	0,00 %
V7	-5,202	-5,293	0,312	0,305	1,75 %
V8	28,678	27,989	0,860	0,697	2,40 %
V9		2,242		1,084	
V10	27,590	27,666	0,828	0,826	0,28 %
V11		430,255		7,270	
V12	9,456	9,323	0,378	0,365	1,41 %
V13	4,860	4,840	0,146	0,145	0,41 %
V14	432,210	457,921	17,288	7,255	5,95 %
V15	24,561	24,561	1,474	1,474	0,00 %
V16	24,564	24,564	1,474	1,474	0,00 %
V17	0,000	0,000	0,000	0,000	0,00 %
V18		49,125		2,084	
V19	14,402	14,402	0,576	0,576	0,00 %
V20					
V21	0,177	0,177	0,006	0,006	0,00 %
V22					
V23	0,111	0,111	0,004	0,004	0,00 %
V24	50,787	50,787	2,031	2,031	0,00 %
V25					
V26					
V27		1,485		2,910	

Stream	Measurement	Estimation	Standard deviation	Estimated Std. Dev.	Correction
V28					
V29	48,696	47,948	1,948	1,930	1,54 %
V30		-2,799		4,233	
V31	121,540	115,949	4,254	4,108	4,60 %
V32	-13,831	-13,967	0,830	0,829	0,98 %
V33					
V34	-17,784	-18,009	1,067	1,064	1,27 %
V35					
V36					
V37					
V38					
V39	381,324	326,288	13,346	7,736	14,43 %
V40	50,800	51,262	2,032	2,017	0,91 %
V41	89,574	89,574	3,583	3,583	0,00 %
V42	366,897	390,975	14,676	7,402	6,56 %
V43	26,093	26,093	1,566	1,566	0,00 %
V44	218,097	218,097	7,633	7,633	0,00 %
V45	222,454	222,454	7,786	7,786	0,00 %
V46		-75,669		13,271	
V47	304,698	304,698	10,664	10,664	0,00 %
V48		159,387	13,071	4,594	
V49	34,373	34,373	2,062	2,062	0,00 %
V50					
V51					
V52					
V53					

Table III : results of equilibration

Table III shows the results of equilibration where column 3 contains the estimated values. Columns 4 and 5 allow to compare the accuracy of the measurements and that of estimated values. In order to make the accuracies more homogeneous, all of them have been converted in standards deviation. The last column shows a correction ratio of the measured values. Nul terms in this column correspond to the measurements which have not been corrected.

Name of the stream	Residual	Normalized residual	Fault probalility
V1	-1,073	-1,367	82,83 %
V3	0,000	-1,694	90,98 %
V4	0,000	0,000	0,00 %
V5	-17,037	-1,700	91,09 %
V7	0,091	1,367	82,83 %
V8	0,689	1,367	82,83 %
V10	-0,760	-1,700	91,09 %
V12	0,133	1,367	82,83 %
V13	0,200	1,367	82,83 %
V14	-25,711	-1,638	89,87 %
V21	0,000	-2,806	99,50 %
V23	0,000	-2,204	97,25 %
V29	0,748	2,863	99,58 %
V31	5,591	5,061	100,00 %
V32	0,136	2,863	99,58 %
V34	0,225	2,863	99,58 %
V39	55,036	5,061	100,00 %
V40	-0,462	-1,900	94,26 %
V42	-24,078	-1,900	94,26 %
Residual criterium value :		27,676	
Chi square probability :		100,00 %	

Table IV : Analysis of the residuals

After this equilibration operation, L.V.D. then authorizes to analyze the residuals (estimated value - measured value). Table IV summarizes these results where one can see the residual, the normalized residual (formulae 9) and the associated fault probability. With a confidence level of 0.95, the upper bound of the confidence interval for the residual criterium value with four degree of freedom is equal to 9.49. On the example, one can see that this bound is largely overstep. The whole results cannot then be considered as correct. Therefore, it is necessary to analyse the residual indivudually. Notice that, if we choose a critical value equal to 2, seven streams seem to be suspicious. The two largest residuals are those of streams V31 and V39, or one can demonstrate that the stream that supported the largest residual is the more suspicious, meanwhile in some cases, it is not possible to discriminate the part of each stream (in our example, the streams V31 and V39 have exactly the same part between node S2 and the pseudo node issue from the merge of nodes B5, S4, S5, S6, S7 and L1). In that case, it is necessary to delete the measurements of all the streams that cannot be discriminated.

Then if we compute a new equilibration without these measurements, it remains 3 redundancy equations, the value of the residual criterium value falls to 2 and the associated chi square probability to 43.75 %. This new probability is smallest than 95 %, then the obtained results are considered to be satisfactory. We conclude that, amongst measurements of the streams V31 and V39, at least one of them is biased. L.V.D. is able to do automatically the previous analysis and to list the more suspicious measurements.

7 - CLOSING REMARKS

Throughout this communication, the authors have presented a short overview of the important problem of data validation in the linear case. The presented software, L.V.D., has been initially developped to answer to research laboratory problems. It has then be installed in industrial environment (ORKEM, Rhône Poulenc, ELF...) essentially to diagnose the working of a process by the historically analysis of the results.

Literature cited

/1/ R.S.H. MAH, M. STANLEY, D. DOWNING : I.E.C. Proc. Des. Dev. 15, (1976), 175.
/2/ V. VACLAVEK : Chem. Eng. Sci. 24, (1969), 947.
/3/ J. ROMAGNOLI, G. STEPHANOPOULOS : Chem. Eng. Sci. 36, (1981), 1849.
/4/ M. DAROUACH : Thèse de doctorat d'état, NANCY, 1986.(In french)
/5/ S. NOGITA : I.E.C. Proc. Des. Dev. 11, (1972), 197.
/6/ D. MAQUIN : Thèse de l'Université de Nancy I, NANCY, 1987. (In french)