DATA RECONCILIATION

Georges Heyen
Department of Chemical Engineering, University of Liège, Belgium

Keywords
Measurement, redundancy, uncertainty, validation, error detection, error correction, data reconciliation, process operation, monitoring, state estimation, parameter identification, mathematical model.

Contents

Summary
Reliable process data, such as flow rates, compositions, temperatures, pressures and phase fractions, are the key to efficient operation of chemical plants. With the increasing use of computers in industry numerous data are acquired and used for on-line optimization and control. However, raw measurements are not accurate enough; they are affected by random or systematic errors, due to sensor drift, calibration errors, instrument malfunctions, leaks and so forth. Hence the measurements cannot satisfy exactly material and energy balances or other model constraints. The goal of data validation is to reconcile the contradictions between the measurements and their constraints, to estimate the true values of measured variables, to detect gross errors and solve for some unmeasured variables. Thus one can obtain the required process data with high accuracy and reliability, and generate consistent balances for accounting.
Algorithms used to correct random errors and allow closing process balances are discussed, both for steady state and dynamic systems.
Practical applications are described, and the benefits of data validation are illustrated.. Up to now steady-state data reconciliation and gross error detection began to be applied widely in industrial plants in 1980s. This technology is now a mature field with certain challenges remaining. The reconciliation for dynamic systems is an active development field; however, its on-line application to large industrial systems is still in its infancy.

1. Scope, aims and benefits of Data Reconciliation

Nowadays, industrial processes are more and more complex and difficult to master. They process large quantities of valuable goods, and thus should be run efficiently to avoid wasting raw materials and ensure a high product quality. The operation of many chemical plants involves also potentially dangerous operations: strict process monitoring is necessary to avoid unsafe operating conditions that could lead to fire, explosion or release of toxic components in the environment. The size of the equipment, the value of products they transform, the requirements for safety thus dictate that processes should be monitored and controlled efficiently.

1.1 Importance of measurements for Process Monitoring

Efficient and safe plant operation can only be achieved if the operators are able to monitor all key process parameters. Instrumentation is used to measure many process variables, like temperatures, pressures, flow rates, compositions or other product properties. Measuring these variables should allow the operators to verify that the equipment is operating according to the design. Without good measurements, the operators would be blind: similarly, to drive a car, one needs to see the road, locate the car position with respect to obstacles, and know its speed. When visibility is poor, the safe decision is to reduce speed, or even to stop the car. In the same way, when measurements do not allow assessing a plant operating condition, it cannot be run safely at maximal efficiency.

In practice, direct measurements do not provide always all the required information. What is needed is an estimation of some performance indicators. These are the variables that either contribute to the process economy (e.g. the yield of an operation), or are linked to the equipment quality (e.g. fouling in a heat exchanger or activity of a catalyst), to safety limits (e.g. departure from detonation limit) or to environmental considerations (e.g. amount of pollutant emissions). Most performance parameters are not directly measured, and are evaluated by a calculation based on one or several measured values.

For instance, a car driver is interested in knowing how much fuel is left in the tank. What is measured is the level in the tank. Thus some knowledge about the physical plant (here the shape and size of the tank) must be known to calculate the useful value (amount of fuel) from the raw measurement. In many cases, several independent measurements must be combined to assess the value of some process variable (e.g. the mileage for a vehicle is computed from the fuel consumption (based on the variation of the fuel tank level, or from a direct flow rate measurement) and from the distance traveled, obtained from an odometer).

However some difficulties arise when one considers experimental errors.

1.2 Sources of experimental errors

Experimental data is always affected by experimental errors. No sensor can be built that is absolutely exact and accurate. Besides uncertainty linked to the measuring device, errors can also arise from sampling or positioning the sensors (e.g. measuring local properties in a material that is not homogeneous), from inappropriate calibration, from transcription or reporting errors (during signal conversion for instance).

One should make a distinction between permanent bias or systematic errors, and random deviations. The overall error results from summing both contributions. Systematic errors are related to deficient instrumentation or inexact calibration: an example would be using erroneous weights or a chronometer that runs late. No matter how careful the measurement is carried out, the error will remain undetected, even if the measurement is repeated. The only way out is to compare the measurement with an independent assessment using a different sensor, and such a procedure allows then to calibrate properly the defective sensor. In other respects random errors are due to a multiplicity of causes, and may result from fluctuations in sampling or external perturbations (e.g. variation of atmospheric pressure, voltage fluctuations for electric instruments). They can be detected by repeating the measurement, and noticing that the outcome is different.

Measurement error is the sum of both contributions: systematic and random errors.

Good accuracy
High precision

Good accuracy
Low precision

Low accuracy
High precision
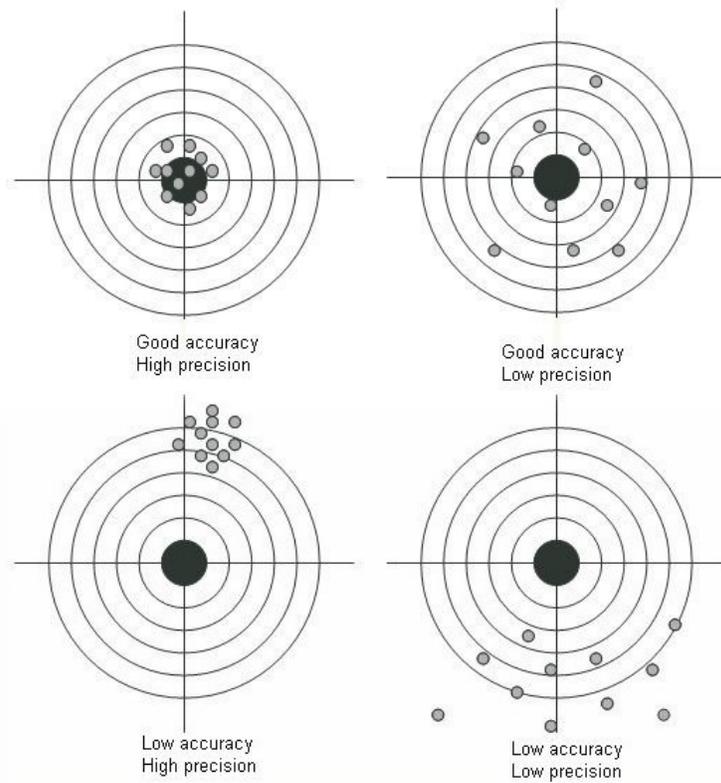
Low accuracy
Low precision

Figure 1 : Comparison between precision and accuracy

Figure 1 allows illustrating the difference between accuracy and precision, by comparing the measurement process with shooting at a target. Accuracy represents the systematic departure of the measurement with respect to the true value (usually unknown). For the shooting analogy, this could be corrected by adjusting the sight. For a measurement, inaccuracy results from instrument bias and improper calibration. Precision, for the shooting analogy, is related to the spread of the bullets on the target. Low precision results from imperfect instrumentation and variation in operating procedures. Precision is linked to the repeatability of the measurement: a clock can be very precise (exactly 3600 ticks every hour) and give systematically the wrong time.

Repeating measurements allows estimating their precision, by assessing the spread of their distribution around the average value (assuming that the measured variable remains constant during the measurement process). Thus we can expect that measurement redundancy is a way to improve the quality and reliability of the measurement results.

Random errors that always affect any measurement also propagate in the estimation of performance parameters. When redundant measurements are available, they allow the estimation of the performance parameters based on several independent data sets; this provides different estimates, which may lead to confusion if not properly interpreted. Data validation is the method applied to properly exploit measurement redundancy in order to improve the assessment of process variables.

1.3 How to achieve measurement redundancy

Measurement redundancy can be obtained in several ways.

A first approach is to repeat several times the same measurement using the same sensor. This is called temporal redundancy. By taking the average of the measurements, one can expect to decrease the uncertainty arising from random errors. In fact, statistics explain that the variance of the mean value $\sigma_{\bar{X}}^2$ is proportional to the inverse of the number of measurements N:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{N}$$

(1)

In a process whose variables are likely to fluctuate with time, measurement redundancy can also be achieved by installing multiple sensors in order to obtain several simultaneous measurements of the same variable(s). This procedure allows not only to reduce uncertainty by averaging the measured values, but also to detect gross errors resulting from sensor failures. Such an approach is used for some safety critical measurements, coupled with comparison software that implements a voting scheme in case contradictory measurements are obtained. However this type of sensor redundancy is costly and not applied systematically to all process variables.

Redundancy can also be achieved by using several measurements combinations and a process model to estimate the required process variables. To explain this method, we need first to evaluate the uncertainty of an estimate when several measurements are needed to assess a variable value.

An example is shown in Figure 2, where a weight is evaluated by summing two measurements. The variance of the estimate is obtained by summing the variance of both measurements.

$$W = W_1 + W_2$$

$$\sigma_W^2 = \sigma_{W_1}^2 + \sigma_{W_2}^2$$

(2)

For the more complex set up shown in figure 3, we need to use a model of the set up, and use the equilibrium condition to obtain the value of the weight W from the measurements W_1 and W_3:

$$W_1 = 2\left(W + W_3\right)$$

$$W = 0.5\ W_1 - W_3$$

(3)

$$\sigma_W^2 = 0.25\ \sigma_{W_1}^2 + \sigma_{W_3}^2$$

Figure 2 : use several measurements to assess a value

Figure 3 : Indirect evaluation by combining measurements
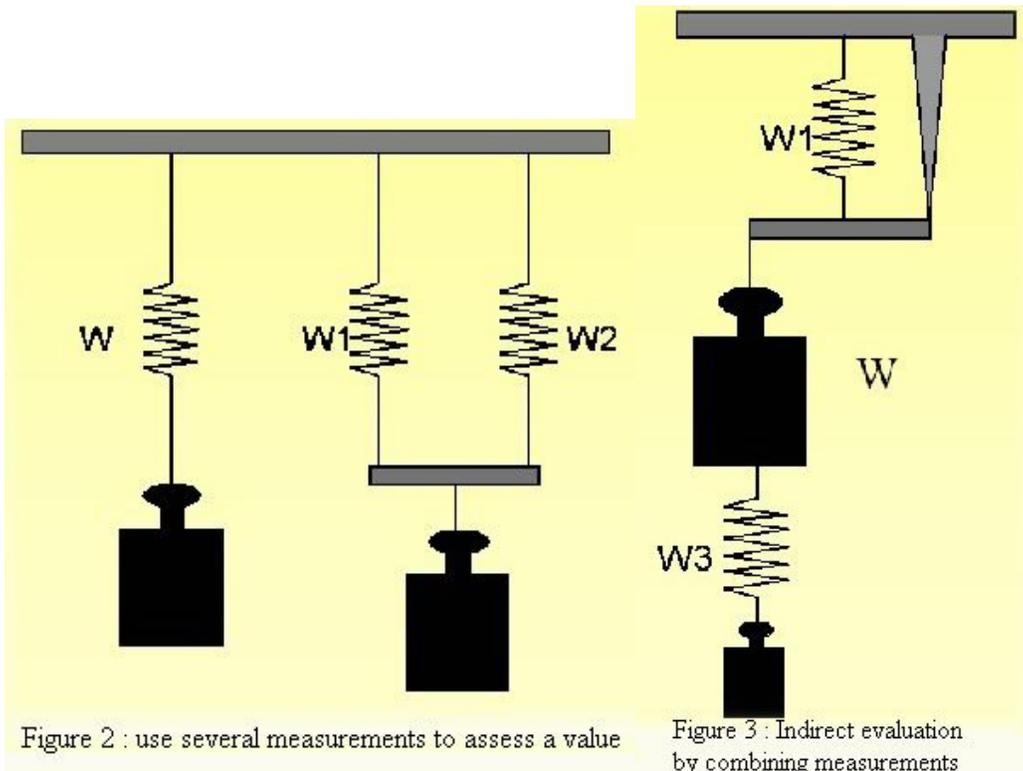
In general, if a variable W can be calculated using a model f and the value of several independent measured variables x_i, the variance of its estimate will be related to the measurement variances, using the following relationship obtained by linearizing the model f:

$$W = f\left(x_i\right) \qquad i = 1...n$$

$$\sigma_W^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_{x_i}^2 \qquad (4)$$

The variables appearing in the model need not to be of the same type. In fact, some process variables can be estimated in several independent ways. As an example, let us consider the case shown in figure 4.

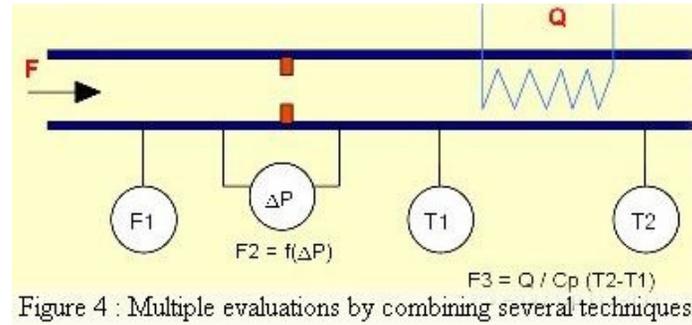Figure 4 : Multiple evaluations by combining several techniques

A flowrate in a pipe can be directly measured as F1 using a flow meter (e.g. using Doppler effect). The flowrate can also be estimated by measuring the pressure drop through an orifice, which will provide estimate F2. It can also be obtained from an energy balance, for instance by heating the fluid using electrical power and measuring the temperature increase. If the fluid specific heat C_p is known, the flowrate estimate F3 will be related to the power dissipated Q and the temperature increase by:

$$F3 = \frac{Q}{c_p\left(T_2 - T_1\right)} \qquad (5)$$

A data validation algorithm will provide a way to merge those independent estimates and pool their variances in order to provide a consistent value of the flowrate.

2. Exploiting redundancy
In order to apply data validation techniques, the information needed is:
- a process model, i.e. a set of mathematical equations that relate the values of all process variables;
- a set of measurements, providing experimental values for the process variables, or a subset of them;
- an estimation of the measurement uncertainty, in the form of a standard deviation for each measured value.

The model equations will allow either to calculate the values of unmeasured variables from the measurements, or to exploit structural redundancy to verify the measurement consistency, identify and possibly correct gross errors, and reduce the uncertainty affecting the variable estimates.

2.1 Variable classification
By analyzing the measurement set and the model structure, we can proceed to variable classification. Such an analysis is usually carried out by data validation software before attempting to solve a validation problem in order to verify the existence of a unique solution.
Variables are either measured or not.
Unmeasured variables are observable if their value can be inferred from the available measurements using the model equations. If no model equation is available to calculate a variable value from the measurements, the variable is not observable.
The value of measured variables is directly obtained from the measurement. However when several independent ways can be found to estimate a variable, it is said to be redundant. Its redundancy

number is equal to the number of measurements that contribute to the variable evaluation, but can be deleted simultaneously while saving a way to estimate the variable.

The level of redundancy is thus defined as the number of measurements which come in addition to the minimal set of measurements needed to be able to calculate the system.

The easiest way to understand this concept of redundancy is to examine an example where only an overall mass balance is considered, as shown in figure 5.

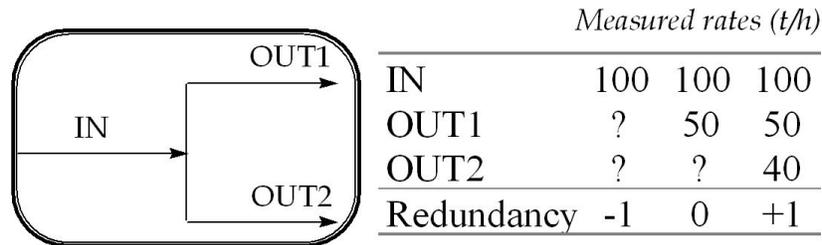| *Measured rates (t/h)*			
IN	100	100	100
OUT1	?	50	50
OUT2	?	?	40
Redundancy	-1	0	+1

Figure 5: The concept of redundancy

If the flow rate of stream IN is the only one measured, there is not enough information to fully describe the state of the system as nothing can tell how much goes through each outlet lines OUT1 and OUT2. The redundancy is negative, meaning information is missing.

If the inlet rate and one of the outlet rates is known, that we have just enough information to describe the system and the redundancy level is zero. The rate of OUT2 can be calculated as the difference between the rates of IN and OUT1.

Finally, if both outlet rates are measured in addition to the inlet rate, than we have more information than what is necessary to describe the system. We have a redundancy of level one, since there is one extra measurement beyond the minimal number required to calculate all flow rates.

2.2 Benefits of model based data validation

Benefits from data reconciliation are numerous and include:

- Improvement of measurement lay-out
- Decrease of number of routine analyses
- Reduced frequency of sensor calibration : only faulty sensors need to be calibrated
- Removal of systematic measurement errors
- Systematic improvement of process data
- Clear picture of plant operating condition and reduced measurement noise in trends of key variables
- Early detection of sensors deviation and of equipment performance degradation
- Consistent and closed plant balances for accounting and performance follow-up
- Safe operation closer to the limits
- Quality at the process level.

3. Mathematical formulation of the validation problem

The main aim of data reconciliation is to reduce uncertainty in the estimation of process variables, and to obtain estimates that are consistent with some model constraints, such as conservation equations. When measured values need to be corrected, the measurement precision is taken into account, so that less precise values are the most corrected. Model equations are also used to calculate values of unmeasured variables.

Model equations used in a data validation model are mainly conservation equations (mass and energy balance). Link equations relating the measured variables to the state variables are added to the model. For instance, component molar flow rates d_i are used as state variables, while the

measured variables are the total mass flow rate g and the component molar fractions x_i. thus the following link equations are included in the model :

$$d_i - x_i \sum_{j=1}^{Nc} d_j = 0 \qquad i = 1, N_c$$

$$g - \sum_{j=1}^{Nc} M_j d_j = 0$$

(6)

where Mj is the molar mass of component j.

Phase equilibrium equations are also useful to increase the redundancy level in a validation problem. They can be considered each time a stream can be reasonably assumed to be saturated (either at its dew point or bubble point).

However design equations should be avoided in data validation models, unless their purpose is just to obtain estimates of unmeasured performance parameters such as heat transfer coefficients. One main goal of data validation is to obtain high quality estimates, in order to monitor the process efficiency, e.g. by following trends in the heat transfer rate to detect fouling. Since the idea is to detect deviations from the design conditions, we should avoid correcting measurements by enforcing the design equations using specified values for the performance parameters.

3.1 Data validation for steady state systems

In a steady state system, process variables do not vary with time. Let us consider a system with n unmeasured variables $\{z_i, i=1,n\}$ and m measured variables x_i for which measurements y_i are available $\{x_i, y_i, i=1,m\}$. The variables are linked by a set of p algebraic model equations $\{f_j(\underline{x},\underline{z}), j=1,p\}$.

It is usually assumed that measurement errors follow a Gaussian distribution with zero mean and diagonal covariance matrix. Thus the precision of each measurement is characterized by its standard deviation σ_i.

The data validation problem results then in a constrained minimization problem:

$$\min_{\underline{x}, \underline{z}} \sum_{i=1}^{m} \frac{(y_i - x_i)^2}{\sigma_i^2}$$
$$s.t. \quad \underline{f}(\underline{x}, \underline{z}) = 0$$

(7)

Solving this optimization problem provides simultaneously the measurement error corrections and the estimates for unmeasured variables.

3.2 Solution for linear systems

The simplest data reconciliation problem deals with steady state mass balances, assuming all variables are measured, and results in a linear problem.

There x is the vector of *n* state variables, while y is the vector of measurements. We assume that random errors e=y-x follow a multivariate normal distribution with zero mean.

The state variables are linked by a set of *p* linear constraints:

$$A x - d = 0$$

The data reconciliation problem consists in identifying the state variables x verifying the set of constraints, and being close to the measured values in the least square sense, which results in the following objective function :

$$\min_{x} (y - x)^T W (y - x)$$

(8)

where W is a weight matrix.

The method of Lagrange multipliers allows obtaining an analytical solution:

$$x = y - W^{-1} A^T \left(A W^{-1} A^T \right)^{-1} (Ay - d)$$

(9)

It is assumed that there are no linearly dependent constraints.

Usually W is taken as the inverse of the covariance matrix C of measurement errors, and the solution is thus:

$$x = y - CA^T \left(ACA^T \right)^{-1} \left(Ay - d \right)$$

$$= \left[I - CA^T \left(ACA^T \right)^{-1} A \right] y + CA^T \left(ACA^T \right)^{-1} d \qquad (10)$$

$$x = My + e$$

The estimates are thus related to the measured values by a linear transformation. Their precision can easily be obtained from the measurement precision (covariance matrix C) and from the model equations (matrix A and array d).

In case some variables z are not measured, the problem formulation becomes, for a linear model:

$$\min_{x,z} \left(y - x \right)^T C^{-1} \left(y - x \right) \qquad (11)$$

$$s.t. \quad Ax + Bz + d = 0$$

Analysis of the B matrix (a subset of the Jacobian matrix of the model, corresponding to the unmeasured variables) allows determining whether the validation problem has a solution. The size and rank of the matrix must be such that enough equations are available to calculate unmeasured variables from the validated measurements. A necessary condition is to have more model equations than unmeasured variables, thus p≥n. However this overall redundancy condition is not sufficient: if the measurements are not adequately located, part of the system could be validated, while other variables would be non observable.

Before solving the validation problem, some variable classification and pre-analysis is needed to identify unobservable variables and parameters, as well as nonredundant measurements Algorithms are available to analyze the structure of matrix B, by reordering variables and equations in order to assign an equation to each unmeasured variable. They allow checking the existence of a solution before attempting to run the calculation, and to classify the variables according to observability criteria. Measured variables can be classified as *redundant* (if the measurement is absent or detected as a gross error, the variable can still be estimated from the model) or *nonredundant.* Likewise, unmeasured variables are classified as *observable* (estimated uniquely from the model) or *unobservable*. The reconciliation algorithm will correct only redundant variables. If some variables are not observable, the program will either request additional measurements (and possibly suggest a feasible set) or solve a smaller sub-problem involving only observable variables. The preliminary analysis should also detect *over specified variables* (particularly those set to constants) and *trivial redundancy*, where a measured variable does not depend at all upon its measured value but is inferred directly from the model. Finally, it should also identify model equations that do not influence the reconciliation, but are merely used to calculate some unmeasured variables. Such preliminary tests are extremely important, especially when the data reconciliation runs as an automated process.

3.3 Nonlinear case

The general formulation described here above for a steady state nonlinear model can be solved numerically using general purpose nonlinear programming algorithms. SQP algorithms have been used efficiently for that purpose. Practical validation problems tend to be large (several hundreds of variables and equations), thus algorithms exploiting the sparse pattern of the model equations are favored.

An alternative is to use Lagrange method, and to transform the constrained optimization problem in a larger unconstrained problem:

$$\min_{x, z, \lambda} \quad L = \sum_{i=1}^{m} \frac{\left(y_i - x_i\right)^2}{\sigma_i^2} + \sum_{j=1}^{p} \lambda_j f_j\left(x, z\right) \tag{12}$$

The necessary condition for optimality is expressed by setting to zero the gradient of the Lagrangian; the validation problem thus results in solving a system of m+n+p algebraic equations:

$$\frac{\partial L}{\partial x} = C^{-1}\left(x - y\right) + A^T.\Lambda = 0$$

$$\frac{\partial L}{\partial z} = B^T.\Lambda = 0 \tag{13}$$

$$\frac{\partial L}{\partial \Lambda} = f(x, z) = 0$$

where A and B are partitions of the Jacobian matrix corresponding to derivatives of the model with respect to measured and unmeasured variables, and Λ is the array of Lagrange coefficients.
This last equation can be linearized as :

$$\frac{\partial L}{\partial \Lambda} = A.x + B.z + d = 0 \tag{14}$$

where A and B are partial Jacobian matrices of the model equation system :

The equation system is nonlinear and has to be solved iteratively. Initial guesses for measured values are straightforward to obtain. Process knowledge usually allows estimating good initial values for unmeasured variables. No obvious initial values exist for Lagrange multipliers, but solution algorithms are not too demanding with that respect. Newton's method is suitable for small problems, and requires solving successive linearizations of the original problem:

$$\begin{vmatrix} x \\ z \\ \lambda \end{vmatrix} = J^{-1} \begin{vmatrix} C^{-1}y \\ 0 \\ -d \end{vmatrix} \tag{15}$$

where the Jacobian matrix **J** of the equation system has the following structure:

$$J = \begin{vmatrix} C^{-1} & 0 & A^T \\ 0 & 0 & B^T \\ A & B & 0 \end{vmatrix} \tag{16}$$

This shows again that validated results for measured variables x and unmeasured variables z are directly related to the measurements y and the covariance matrix C. A linear approximation of this relationship (similar to equation 4) is readily obtained from equation. 15.

3.4 Reduction of uncertainty
Solving the data reconciliation problem provides more than validated measurements. A sensitivity analysis can also be carried out, based on linearizing the equation system. As shown by equation 4, reconciled values of process variables x and z are linear combinations of the measurements. Thus their covariance matrix is directly derived from the measurements covariance matrix.
Inspecting the variance of validated variables allows detecting what are the key measurements in the state identification problem. Some measurements may appear to have a very high impact on key validated variables and on their uncertainty: particular attention should be paid to these measurements, and it may prove wise to duplicate the corresponding sensors.

The uncertainty of validated values can be compared to the original uncertainty of the corresponding raw measurements. Their ratio measures the improvement in confidence brought by the validation procedure. Nonredundant measurement will not be improved by validation. The uncertainty of the estimates for unmeasured observable variables is also obtained.

Sensitivity analysis provides a list of all state variables whose estimates depend on a given measurement, as well as the contribution of the measurement uncertainty to the variance of the reconciled values. This information helps to locate critical sensors, whose failure may lead to troubles in monitoring the process. A similar analysis allows locating sensors whose accuracy should be improved in order to reduce the uncertainty affecting the major process performance indicators, and helps to improve the rational design of sensor networks.

3.5 Extension to dynamics

The data validation methods shown above are only valid for steady state systems. In practice they are also used to handle measurements for processes operated close to steady state, with small disturbances. Measurements are collected automatically and average values are calculated and further processed with a steady state validation algorithm. Such an approach provides useful results for monitoring slowly varying performance indicators, such as fouling coefficients in heat exchangers. Such parameters are needed to fine tune a steady state simulation model, e.g. before optimizing set point values that are updated once every few hours.

However a more rigorous approach can be developed. The dynamic validation problem is also expressed as a constrained minimization problem, where the objective function is usually the sum of squares of weighted corrections to the measurements, and the constraints are a set of differential and algebraic equations. In order to limit the problem complexity, the measurement set is restricted to a time window that is moved forward each time the validation problem is solved, and older measurements are discarded to make room for new ones.

Thus the dynamic validation problem can be formulated as:

$$\min_{x,z} \sum_{j=t_0}^{t_N} \left(y_j - x(t_j) \right)^T W_j \left(y_j - x(t_j) \right) \tag{17}$$

subject to

$$f\left(\frac{dx(t)}{dt}, \frac{dz(t)}{dt}, x(t), z(t) \right) = 0 \quad ; \quad x(t_0) = x_0 ; z(t_0) = z_0 \tag{18}$$

$$h\left(x(t), z(t) \right) = 0 \tag{19}$$

$$g\left(x(t), z(t) \right) \leq 0 \tag{20}$$

Measurements are collected at regular intervals over a time horizon from t_0 to t_N. The model expresses the variation of measured variables x(t) or unmeasured variables z(t), that are related by differential equations, algebraic equality and inequality constraints.

The problem can be solved using nonlinear programming algorithms, by embedding the numerical solution of the differential algebraic system in the evaluation of the objective function. An alternative is to approximate the x and z functions by a set of polynomials, and to enforce the differential relationships at some collocation points. This transforms the differential equations into a set of algebraic equations, which reduces the problem to the classical framework for validation using algebraic models.

4. Applications

We will first discuss a simple academic example to illustrate the features of validation techniques. Later two industrial success stories will be presented, and economical benefit obtained by using data reconciliation will be reported. More industrial applications are described in literature (see bibliography section below and web sites of validation software publishers)

4.1 Illustration of the method

To illustrate data validation, we consider first a simplified flowsheet of a cogeneration plant, combining a gas turbine and a steam generator. Figure 6 shows both the flowsheet and some results of a data validation run obtained with Belsim's VALI 4 software.
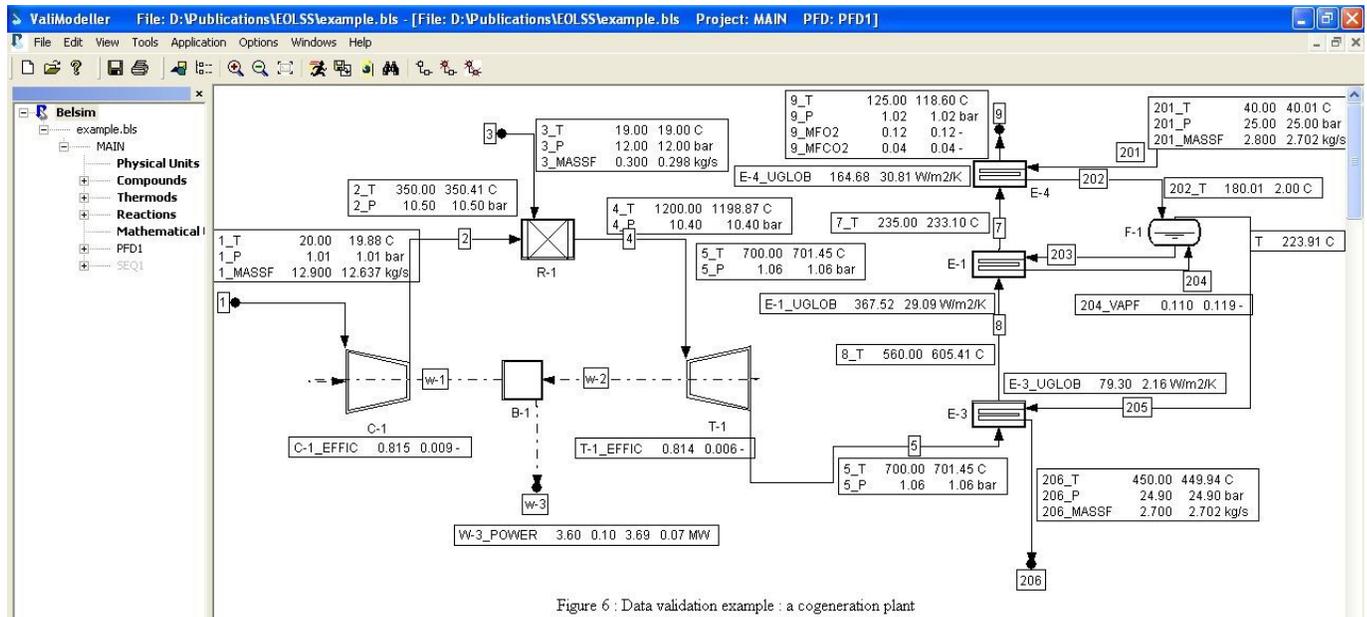

Figure 6 : Data validation example : a cogeneration plant

An air stream 1 is fed to compressor C1. Compressed air 2 reacts with natural gas 3 in combustion chamber R-1. Hot combustion gas 4 is expanded in T-1. Expansion work W-2 allows driving the compressor (work W-1) and the surplus W-3 is available as net work. Turbine exhaust 5 is cooled down in a series of heat exchangers and rejected to the stack as stream 9.

Boiler feed water in stream 201 is preheated in economizer E-4 and feeds drum F-1. A saturated liquid stream 203 is circulated to the vaporizer E-1 and returns to the drum as a vapor-liquid stream 204. Saturated steam 205 is further heated in E-3 to generate the superheated steam product 206.

This model involves mass and energy balances, reactions and phase equilibria. Isentropic efficiency parameters are evaluated from the compressor and the turbine models. Overall heat transfer coefficients are estimated for all heat exchangers. Thus this simple example covers all the main features of steady state validation discussed in this article.

Table 1 lists the main variables in this example. All streams are described internally using standard state variables, namely partial molar flowrates for all components, pressure and molar enthalpy. Most of those state variables are not directly measured (except pressures). In our example, some stream properties are supposed to be known exactly, such as the air and gas compositions. Some unit parameters are also given as exact numbers, such as the heat exchanger areas. Some extra variables need to be related to state variables because they are measured (temperatures, mole fractions) or because they should be estimated (heat transfer coefficients, compressor and turbine efficiency).

Figure 6 shows the user interface for typical data validation software. This is the interactive graphical user interface allowing the user to set up a problem description and run case studies, while in routine use, the validation program is run automatically in the background and exchanges measurements and validated results with process instrumentation systems and historians. Units and streams are created and linked by drawing a flowsheet using icons. Some key figures can be

displayed in result boxes directly on the flowsheet. For instance, validated variables can be compared to the corresponding measurements.

Table 1 : Data validation example : variable list

Variables specified as constants (25)

Stream / Unit	Variables
1	P, air composition (mole fractions Ar, CH_4, C_2H_6, CO_2, H_2O, N_2, O_2)
3	Gas composition (mole fractions Ar, CH_4, C_2H_6, CO_2, H_2O, N_2, O_2)
4	CH_4 and $C_2H_6 = 0$
E-1	Exchanger area
E-3	Exchanger area
E-4	Exchanger area

Variables measured (28)

Stream / Unit	Variables
1	Mass flow, T
2	T, P
3	Mass flow, T, P
4	T, P
5	T, P
7	P
8	P
9	T, P, mole fractions CO_2, O_2
201	Mass flow, T, P
202	T, P
203	Mass flow
204	Vapor fraction
206	Mass flow, T, P
W-3	Power

Non measured process parameters to be estimated (9)

Stream / Unit	Variables
7	T
8	T
W-1	Power
W-2	Power
C-1	Isentropic efficiency
T-1	Isentropic efficiency
E-1	Heat transfer coefficient
E-3	Heat transfer coefficient
E-4	Heat transfer coefficient

Table 2 shows the values and precisions of all measurements, and the validated results (with corresponding standard deviations).

Table 2 : Data validation example : data and results

```
BELSIM s.a.  19/01/2006  12:15    VALIDATION version 4.1.0.6w
 Web Site  : www.belsim.com

   TAG NAME       Measured   accuracy   Validated   accuracy   Penalty   Unit
 ===============================================================================
   1_MASSF        12.900      5.00    %  12.636      1.56    %    0.17   kg/s
   1_MFAR         0.90000E-02 CST        0.90000E-02                     -
   1_MFC1         0.0000      CST        0.0000                          -
   1_MFC2         0.0000      CST        0.0000                          -
   1_MFCO2        0.0000      CST        0.0000                          -
   1_MFH2O        0.10000E-01 CST        0.10000E-01                     -
   1_MFN2                     OFF        0.77600     0.333E-16           -
   1_MFO2         0.20500     CST        0.20500                         -
   1_P            1.0120      CST        1.0120                          bar
   1_T            20.000      1.00       19.883      0.996          0.01  C
   201_MASSF      2.8000      5.00    %  2.7023      1.66    %    0.49   kg/s
   201_P          25.000      5.00    %  25.000      5.00    %    0.00   bar
   201_T          40.000      1.00       40.013      1.00           0.00  C
   202_P          24.980      1.00    %  24.980      1.00    %    0.00   bar
   202_T          180.00      2.00       180.01      2.00           0.00  C
   203_MASSF      25.000      5.00    %  25.077      4.89    %    0.00   kg/s
   204_VAPF       0.11000     0.300E-01  0.11934     0.616E-02    0.10   -
   206_MASSF      2.7000      2.00    %  2.7023      1.66    %    0.00   kg/s
   206_P          24.900      1.00    %  24.900      1.00    %    0.00   bar
   206_T          450.00      3.00       449.94      3.00           0.00  C
   2_P            10.500      1.00    %  10.500      1.00    %    0.00   bar
   2_T            350.00      2.00       350.41      1.97           0.04  C
   3_MASSF        0.30000     2.00    %  0.29809     1.46    %    0.10   kg/s
   3_MFAR         0.0000      CST        0.0000                          -
   3_MFC1                     OFF        0.91000     0.430E-17           -
   3_MFC2         0.40000E-01 CST        0.40000E-01                     -
   3_MFCO2        0.0000      CST        0.0000                          -
   3_MFH2O        0.0000      CST        0.0000                          -
   3_MFN2         0.50000E-01 CST        0.50000E-01                     -
   3_MFO2         0.0000      CST        0.0000                          -
   3_P            12.000      2.00    %  12.000      2.00    %    0.00   bar
   3_T            19.000      1.00       18.999      1.00           0.00  C
   4_MRC1         0.0000      CST        0.0000                          kmol/s
   4_MRC2         0.0000      CST        0.0000                          kmol/s
   4_P            10.400      1.00    %  10.400      1.00    %    0.00   bar
   4_T            1200.0      3.00       1198.9      2.84           0.14  C
   5_P            1.0600      0.100E-01  1.0600      0.100E-01    0.00   bar
   5_T            700.00      3.00       701.45      2.77           0.23  C
   7_P            1.0300      2.00    %  1.0300      2.00    %    0.00   bar
   7_T                        OFF        233.08      2.04    %           C
   8_P            1.0400      2.00    %  1.0400      2.00    %    0.00   bar
   8_T                        OFF        605.41      0.405   %           C
   9_MFCO2        0.38000E-01 0.300E-02  0.37633E-01 0.155E-03    0.01   -
   9_MFO2         0.12000     0.500E-02  0.12255     0.340E-03    0.26   -
   9_P            1.0200      5.00    %  1.0200      5.00    %    0.00   bar
   9_T            125.00      5.00    %  118.58      3.24    %    0.10   C
   C-1_EFFIC                  OFF        0.81452     0.851E-02           -
   E-1_AREA       150.00      CST        150.00                          m2
   E-1_UGLOB                  OFF        367.63      29.1    %           W/m2/K
   E-3_AREA       60.000      CST        60.000                          m2
   E-3_UGLOB                  OFF        79.299      2.16                W/m2/K
   E-4_AREA       150.00      CST        150.00                          m2
   E-4_UGLOB                  OFF        164.71      30.8                W/m2/K
   T-1_EFFIC                  OFF        0.81373     0.625E-02           -
   W-1_POWER                  OFF        4.3491      0.786E-01           MW
   W-2_POWER                  OFF        8.0408      0.120                MW
   W-3_POWER      3.6000      0.100      3.6917      0.682E-01    0.84   MW

   TOTAL NUMBER OF TAGS :                    57
   NUMBER OF EQUATIONS :                     70
   NUMBER OF UNMEASURED VARIABLES :          64
   NUMBER OF MEASURED VARIABLES :            28
   NUMBER OF VARIABLES CONSIDERED CONSTANT :  25
   OBVIOUS NUMBER OF REDUNDANCIES :           6
   TOTAL NUMBER OF REDUNDANCIES :             6
   NUMBER OF TRIVIAL REDUNDANCIES :           0
   OBJECTIVE FUNCTION                 =   2.50929
   CHI-SQUARE                         =   14.0700
   SUM OF SQUARE RESIDUES             =   0.715683E-07
   NUMBER OF BOUNDS ACTIVATED BY SOLVER  =     1
   NUMBER OF VARIABLES CLOSE TO BOUNDS   =     3
```

Some variables are not corrected (e.g. 3_T) because no redundancy allows to improve their value. Some other variables, like O2 content in stream 9, are validated with a much better precision (improvement by a factor of ten with respect to measurement). The process indicators (shaft work, efficiency, transfer coefficients) are also reported, their value being provided with the associated standard deviation. Compressor and turbine efficiencies can be characterized with rather good precision (better than 1%).

Such good results are not achieved for the heat transfer coefficients. They are estimated with a poor precision: for economizer E-4, the validated value E-4_Uglob is 135 W/m²/K with a standard deviation σ=21, or 15% of the estimate.

Sensitivity analysis allows finding a way out. Table 3 shows the sensitivity report for variable E-4_Uglob. It lists all measurements whose inaccuracy contributes significantly to the resulting uncertainty of the target variable. For instance, the first line in the table means that 39% of the uncertainty arises from the standard deviation assigned to measurement of 9_T (temperature of stream 9). If we find a way to improve that measurement, the estimation of the heat transfer coefficient will also be improved. The relative gain indicates that variable 9_T has been validated: its uncertainty has been reduced by 37.12%. Its measurement has also been corrected as shown by the penalty figure that refers to the contribution of 9_T to the objective function:

$$Penalty = \frac{(y_i - x_i)^2}{\sigma_i^2} = 0.25 \qquad (21)$$

Checking the data in table 2, we can see that measurement 9_T is not reliable: its standard deviation is 5% of the measured value, or 9°C. It can and should be upgraded in order to improve the reliability of the heat transfer coefficient estimation.

The value in the column labeled DerVal indicates the sensitivity of the estimated heat transfer coefficient to a change in the measured value. If the measurement increases by one degree, the estimate of E-4_U will increase by 0.39 W/m2/K.

Table 3 : Sensitivity analysis for heat transfer coefficient in unit E-4

Variable			Tag Name	Validated value	Absolute accuracy	Relative accuracy	Penalty	P.U.
UGLOB	U	E-4	Computed E-4_UGLOB	135.49	20.694	15.27%		W/m2/K

Measurement			Tag Name	Contribute	Deriv.Val.	Rel.Gain	Penalty	P.U.
T	S	9	9_T	38.74%	-0.64696	37.12%	0.25	C
MASSF	R	205	206_MASSF	30.10%	210.24	16.21%	0.23	kg/s
MASSF	R	3	3_MASSF	13.45%	-1185.6	29.15%	3.55	kg/s
T	S	5	5_T	5.09%	-1.5557	7.90%	0.66	C
MASSF	R	201	201_MASSF	4.48%	31.278	67.68%	0.28	kg/s
MASSF	R	1	1_MASSF	2.81%	-5.3825	67.94%	0.07	kg/s
T	S	4	4_T	1.57%	0.86495	5.61%	0.36	C
T	S	202	202_T	1.32%	1.1897	0.01%	0.00	C
POWER	S	W-3	W-3_POWER	1.30%	-23.618	30.32%	3.42	MW
T	S	2	2_T	0.46%	-0.70498	1.65%	0.11	C

A variant of this example shows how a gross error in a measurement can be detected. Compared to the base case, the measured value of the fuel flow rate has been increased from 0.30 to 0.34 kg/s, all other measurements remaining constant. This leads to larger discrepancies in the system balance and to larger corrections of several measurements. The objective function (weighted sum of squares of the corrections) becomes larger than a threshold value and does not pass the F test. The validation program then searches for the most likely measurement to explain the error, and correctly identifies 3_MASSF. The validation program runs again ignoring this erroneous measurement and

obtains the results shown in figure 7. They do not differ much from the results obtained with the original measurements, and the fuel flowrate has been correctly estimated form redundant information.

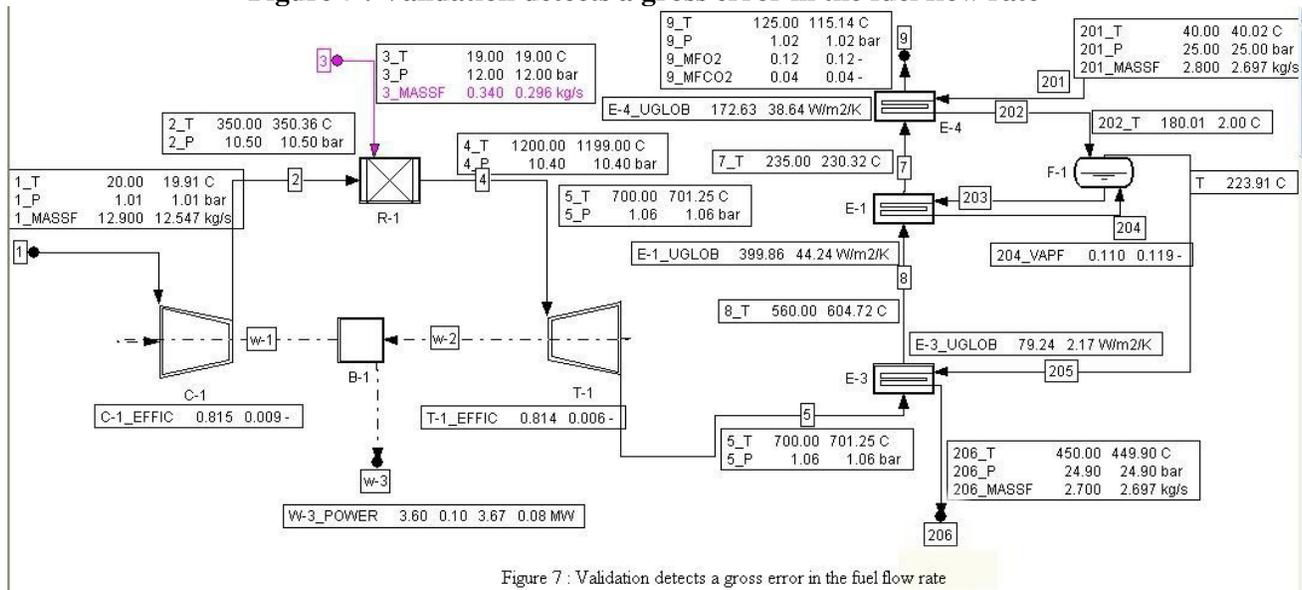**Figure 7 : Validation detects a gross error in the fuel flow rate**

Figure 7 : Validation detects a gross error in the fuel flow rate

4.2 Monitoring steam to carbon ratio in a hydrogen plant

Some reports on industrial applications of data validation have been published. Two examples will be shortly described here.

An on-line implementation of data validation enabled to safely operate a hydrogen plant within tighter limits, giving the opportunity to save a 0.5 Mio Euro/year (ERE refinery, Germany). Fluctuations in the feed gas composition as well as measurement errors led to an uncertainty (30%) in estimating the steam/carbon ratio in the reformer. Operating with too little steam would increase the risk of carbon deposition on the catalyst, forcing a plant shut down. Using too much steam would be costly, since energy is wasted to produce it, and extra cooling is required to condense the excess steam. So the choice of a correct steam to carbon ratio results from a trade-off between safety and economy.

Using validation software on-line, the operators are now able to determine the steam to carbon ratio within one percent accuracy. The validation procedure identifies every hour several key performance indicators, such as heat transfer coefficients, conversion in reactors, unmeasured flow rates, catalyst activity, total energy consumption per ton of H_2 produced, safety margin with respect to carbon deposition. Drift in sensors is also monitored, allowing early maintenance.

This information helps the refinery to improve the follow-up of the unit and to better focus the maintenance of its instrumentation, while at the same time operating the unit at the optimal steam to carbon ratio.

4.3 A nuclear power plant

The priority of nuclear power plant operators is to run their plant as close as possible to the licensed reactor power in order to maximize the generator power. In order to meet this objective, plant operators must have the most reliable evaluation of the reactor power. The definition of this power is based on a heat balance using several measured process parameters among which the boiler feed water flow rate is the most critical value.

Successful use of data validation has been reported for this purpose. The operators have been able to monitor and quantify the deviation between the actual and the measured water flow rate. In

agreement with the authorities in charge of safety, they now recalibrate the flow rate sensors on basis of the reconciled value, as soon as a deviation becomes significant. This enables the power plant to work close to its maximum capacity throughout the whole year, allowing an increase of production valued at 1.0 Mio US$/year. In addition, the use of validation also made the annual heat cycle testing obsolete (saving 0.3 Mio US$/year) and reduced the cost for mechanical and instrumentation maintenance up to 0.7 Mio US$/year, saving a total of 2 Mio US$/year.

5. Conclusions

Data reconciliation is now a mature technology for processes operating close to steady state, for which a model based on steady state conservation laws is adequate. Developments are still going on in several areas:

- Most implementations of data reconciliation assume that measurement errors follow a Gaussian distribution with zero mean and known covariance (usually diagonal covariance is assumed); the implications of those assumptions and the benefit expected from more complex hypothesis still deserve some attention.
- Model-based data reconciliation ignores any model uncertainty; however some constraint equations (eg energy balances) involve empirical relationships, such as the physical property models, that are not totally accurate; a data validation framework accounting for model uncertainty is still to be developed.
- Data reconciliation is useful for process monitoring, but it could be used also for the rational design of measurement systems. Typical questions that can be asked are: where to locate sensors, when to make them redundant, how to minimize measurement cost for a prescribed accuracy, or how to maximize accuracy for a given measurement cost.
- On-line data reconciliation based on steady state models involves the solution of a large set on nonlinear equations. Robust solution is usually sought by using the previous solution as the initial guess for the next problem; this strategy fails when the process structure or the measurement set change (shutdown or start up of a unit, sensor failure, or delayed measurements). Algorithms taking care of such conditions can still be improved;
- Algorithms to detect failing sensors (in order to ignore them) and discriminate between process upsets and perturbed measurements can be improved.

Glossary

Data reconciliation: a procedure to calculate a minimal correction to measured variables, to make them verify a set of model constraints, such as material and energy balances.

Data validation: a set of procedures aiming to improve the reliability and accuracy of plant measurement, combining gross error detection and data reconciliation

Degrees of freedom: number of independent variables that must be fixed in a model to allow its unique solution

Gross error detection: a procedure to detect the existence of systematic errors in sets of measurements and to identify the faulty sensors

Performance indicator: process variables providing information on the quality of some process operation; it can be linked to economy, safety, wear, emissions, etc

Redundancy: number of measurements in excess of the number of degrees of freedom of the model, or difference between the number of equations and the number of unmeasured variables

SQP: sequential quadratic programming, algorithms allowing the iterative solution of constrained optimization problem; at each iteration, constraints are linearized and the objective function is approximated by a quadratic function in the vicinity of the current estimate of the solution.

Nomenclature

A, B : coefficient matrix for linear model; partitions of Jacobian matrix for nonlinear model

C : covariance matrix

c_p : specific heat of a material (J kg^{-1} K^{-1})

f : a model constraint equation

d_i : partial molar flowrate of component i (mole s^{-1})

g : mass flowrate (kg s^{-1})

L : Lagrangian objective function

N : number of variables in a model

m : number of measured variables

n : number of non measured variables

p : number of constraint equations in the model

P : pressure (Pa)

Q : thermal power (W)

t : time (s)

T : temperature (K)

W : weight matrix

y_i : the value of a measurement

z_i : a non measured variable in a model

λ : Lagrange coefficients

σ_i : standard deviation of variable i

Bibliography

Albuquerque J.S., L.T. Biegler, *Data reconciliation and Gross-Error Detection for Dynamic Systems*, AIChE Journal, (42) 2841-2856 (1996) [good introduction paper on error detection for dynamic systems]

Bagajewicz M.J., Design and Retrofit of Sensor Networks in Process Plants, AIChE J., 43(9), 2300-2306 (2001) [strategy to design or upgrade measurement systems]

Crowe C.M., *Observability and redundancy of process data for steady state reconciliation*, Chem. Eng. Sci. 44, 2909-2917 (1989)   [variable classification and redundancy checks]

Heyen G. , Kalitventzeff B *Process monitoring and data reconciliation,* in Computer Aided Process and Product Engineering , Puigjaner, L. Heyen, G. (eds.), Wiley VCH (2006)   [review article, including considerations on the rational design of sensor networks]

Kalitventzeff B., Heyen G., Mateus M.,   *Data Validatio: a Technology for Intelligent Manufacturing,* in Computer Aided Process and Product Engineering , Puigjaner, L. Heyen, G. (eds.), Wiley VCH (2006)   [focus on industrial applications and benefits for process monitoring]

Kalman R.E., A new Approach to linear Filtering and Prediction Problems, Trans. ASME J. Basic Eng. 82D, 35-45 (1960) [seminal paper on model based observation for linear dynamic systems]

Liebman M.J., T.F. Edgar, L.S. Lasdon, *Efficient Data Reconciliation and Estimation for Dynamic Processes using nonlinear Programming Techniques*, Computers and Chemical Engineering, 16, 963-986 (1992)   [seminal paper on dynamic data reconciliation]

Musch H., List T., Dempf D., Heyen G., *Online estimation of Reactor Key Performance Indicators : An Industrial Case Study ,*in*, Computer-Aided Chemical Engineerin*g *volume 18*, Elsevier Science (2004)   [description of an industrial application for online estimation of reactor performance]

Narasimhan S., C. Jordache, *Data Reconciliation and gross Error Detection, an intelligent use of Process Data*, Gulf Publishing Company (2000)   [good introduction, clearly written with examples]

Romagnoli J.A., M.C. Sanchez, *Data Processing and Reconciliation for chemical Process Operations*, Academic Press (2000)   [thorough theoretical and mathematical developments]

Vaclavek V., *Studies on System Engineering : Optimal choice of the balance measurements in complicated chemical engineering systems*, Chem. Eng. Sci. 24, 947-955 (1969) [a pioneering paper on the subject]

Veverka, V.V., Madron, F., *Material and energy balancing in the process industries. From microscopic balances to large plants, Computer-Aided Chemical Engineerin*g *volume 7*, Elsevier Science (1996) [clear introduction to the topic]

Additional information to be found on the web sites of some commercial software publishers :

Belsim VALI 4 : http://www.belsim.com/Products_main.htm

DATACON : http://www.simsci-esscor.com/us/eng/simsciProducts/productlist/datacon/DATACON.htm

SIGMAFINE : http://www.osisoft.com/Sigmafine_Brochure_screen.pdf

DATREC : http://www.technip.com/english/pdf/DATREC.pdf

RECONCILER : http://www.ris-resolution.com/reconciliation.shtml

ADVISOR : http://www.aspentech.com/brochures/Aspen_Advisor.pdf