INTEGRATED OPERATION SUPPORT SYSTEM (IOpSS)
THE DATA PRE-PROCESSING AND DATA RECONCILIATION MODULES

D. Aragón[1], P. A. Rolandi[2], J. A. Romagnoli[1],
[1] Chemical Engineering Department, Louisiana State University, Baton Rouge, LA, USA
[2] Process Systems Enterprise Limited, London, UK

ABSTRACT

This paper discusses the current developments within a novel environment to perform related model-based activities. In particular, the paper focuses in the modules corresponding to data pre-processing and dynamic data reconciliation. In terms of the former module, this work discusses the implementation of three approaches based on the minimum median distance (MMD), the moving median (MM), and the modified MT filter for the detection of outliers. Regarding the data reconciliation module, the error-in-variable method (EVM) was implemented in gPROMS as an important extension to the environment. Finally, the pre-processed data was used to evaluate the performance of the different outlier detection/cleaning methods in the dynamic EVM data reconciliation. Results show that the MM filter has the best performance among the outlier cleaning techniques, followed by the modified MMD method. It is demonstrated that the current EVM implementation is able to perform the reconciliation for complex non-linear dynamic modules and at the same time to estimate parameters and gross errors.

1. INTRODUCTION

Through out the years, chemical engineers have been searching for solutions to daily problems at plant scale. Obtaining a good set of data for further analysis is unquestionable since many plant decisions rely on the results of these analyses. One of the main purposes of having a "cleaned" data is to reconcile the process variables, this means to adjust the process measurements to obtain a better estimate of the process variables (flow rates, temperatures, compositions, etc.) in such a way that they are consistent with the mass and energy balances (Romagnoli and Sanchez, 2000). Once the reconciled data is obtained, more accurate parameters can be estimated. Different methods and objectives functions are available to perform these activities either separate or together, the latter is known as joint data reconciliation and parameter estimation (JPEDR). The objective function most widely used is the least squares function, which consists in minimizing the sum of the squared deviations between the measurements and the estimates.

The consolidation of the CAPE (Computer Aided Process Engineering) community in the 1990's, and the subsequent development of the CAPE-OPEN (CO) project paved the way for a paradigm shift which encouraged the development of a model-centric framework for support of process operations. The definition of model-based activities typical of industrial processes was tackled in a novel framework presented by Rolandi and Romagnoli (2004). In this work, they proposed an innovative methodology for simplifying the problem formulation by incorporating the Problem Definition Environment (PDE) to complement contemporary developments in Open Simulation/Optimization Architectures. The PDE is a software component responsible for the definition of a given model-based engineering problem through a friendly user interface; it allows the selection of variables, and the statement of the specific activity to be performed. Then, to solve the problem, it delegates the corresponding model-based activity to a powerful modeling and solution engine (MSE). Because of

these characteristics, the PDE offers the possibility of executing model-based activities typical of industrial applications providing additional consistency of results and using a single model representation.

In this work we will discuss our improvements to the framework in the data pre-processing and dynamic data reconciliation (DDR) modules. Regarding the DDR, the error-in-variable method (EVM) was implemented as an important extension to the framework previously developed. In EVM data reconciliation, errors are considered to be present in both input and output variables in contrast to the "traditional" data reconciliation, in which errors are considered to be present only in the output variables. In terms of data-preprocessing, we will discuss the implementation of an approach based on the Mean Minimum Distance (MMD) for the detection and median replacement for the rectification of outliers. Furthermore, the extension of the MMD method with median replacement applied to individual variables was considered, allowing the method to detect not only the time where the outlier is present, but also the individual variables contributing to the outlier. The moving median (MM) and the modified MT filter were also implemented. The pre-processed data was used in the reconciliation module to evaluate the performance of the different outlier detection/cleaning techniques.

1.1 Overview

The remainder of this paper is organized as follows. Section 2 contains an introduction to the CO standards. Section 3 presents a description of the environment of the Integrated Operation Support System (IOpSS) along with the methodology followed for its implementation, and discusses the proposed enhancements in the data pre-processing and data reconciliation modules. Section 4 describes the case study used in this work, along with the results obtained in both the data pre-processing and reconciliation. Finally, section 5 draws some conclusions and proposes topics for further work.

2. THE CAPE-OPEN PROJECT

The CO project was initiated with the ultimate purpose of allowing "complex process-modeling tasks and model-based applications to be performed successfully and cost-effectively via the collaborative use of software components from a wide variety of sources, possibly being executed on different computer hardware" (Braunschweig et al, 2000). To make this possible, two key concepts were defined within the project: the PMCs (process modeling components), and the PMEs (process modeling environments). The PMCs are well defined software components in charge of a specific function, such as the calculation of physical properties, or the provision the numerical methods to solve the underlying mathematical problems. On the other hand, PMEs provide the mechanisms for configuring individual elementary models and coordinating the calls among the necessary PMCs to solve the corresponding model-based activity.

After its completion, the CO gave raise to a new project, the Global CAPE-OPEN (GCO), aimed to develop additional standards to the modeling and simulation areas, to encourage the creation of software guided by those standards, and to form the "CAPE-OPEN Laboratories Network" (C.O.Lan) whose objective is to provide support to CO developers and to maintain the standards. It was expected that by the end of the GCO project, there would be a greater acceptance of the CAPE-OPEN standards for communication between components in process engineering (The GCO consortium, 2002).

3. THE INTEGRATED OPERATION SUPPORT SYSTEM

3.1 Architecture of the framework

In addition to the PDE introduced previously, Rolandi and Romagnoli proposed two more concepts for the communication between the actors in their model-centric framework: the Data Model Templates (DMTs), and the Data Model Definitions (DMDs). The DMTs are data structures defining the process variables to be used for further manipulation (a subset of all process variables), their nominal values and other properties related to their structural function. The identity of the variables to be used in the different activities is given by the DMT so that only those specifically identified by the DMT are shown to the user for their selection. More importantly, the DMT creates a link between the name of the variables as known by the user, and their names in the model implementation. In this way, the sensor tags (DCS tag name) will be presented to the user (e.g. a process engineer) and possible confusion with the variable names could be avoided. A given implementation of a DMT for data reconciliation is presented in figure 1.

DCS Tag Name	gPROMS Tag Name	FP	Value	MV	PC	LowBound	Guess	UppBound	EV	RV	IV	OF	OV	CV	Sel
KinPremultR1	SYSTEMCSTR.CSTR1.K	2	3.49E+07			1.00E+07	3.49E+07	9.00E+07	1						
kInPremultR2	SYSTEMCSTR.CSTR2.K	2	3.49E+07			1.00E+07	3.49E+07	9.00E+07	1						
UA-R1	SYSTEMCSTR.CSTR1.UA	2	7.50E+02			6.00E+02	7.50E+02	8.00E+02	1						
UA-R2	SYSTEMCSTR.CSTR2.UA	2	7.50E+02			6.00E+02	7.50E+02	8.00E+02	1						
Product_Value	SYSTEMCSTR.Pval	2	5.00E+02												
Cost_Reactants	SYSTEMCSTR.ReactCost	2	2.00E+00												
Coolant_Cost	SYSTEMCSTR.CoolCost	2	2.00E+01												
QF	SYSTEMCSTR.MV(3)			1			7.00E+00			1					1
QF-B	SYSTEMCSTR.BIAS(3)					0.00E+00	0.00E+00	0.00E+00							
QF-E	SYSTEMCSTR.ERROR(3)					-1.00E-01	0.00E+00	1.00E-01							
QIN-1	SYSTEMCSTR.MV(4)				1		5.00E+00			1			1		1
QIN-1B	SYSTEMCSTR.BIAS(4)					0.00E+00	0.00E+00	0.00E+00							
QIN-1E	SYSTEMCSTR.ERROR(4)					-1.00E-01	0.00E+00	1.00E-01							
QC-1	SYSTEMCSTR.MV(7)				1		2.50E+00			1			1		1
QC-1B	SYSTEMCSTR.BIAS(7)					0.00E+00	1.00E+00	2.00E+00							
QC-1E	SYSTEMCSTR.ERROR(7)					-1.00E-01	0.00E+00	1.00E-01							
QO-1	SYSTEMCSTR.MV(10)			1			5.00E+00			1					
QO-1B	SYSTEMCSTR.BIAS(10)					-2.00E+00	-1.00E+00	0.00E+00							
QO-1E	SYSTEMCSTR.ERROR(10)					-1.00E-01	0.00E+00	1.00E-01							
Least Squares	SYSTEMCSTR.LSQINT		0.00E+00								1				
TO-1I	SYSTEMCSTR.CSTR1.TOO		2.98E+02								1				

Figure 1. DMT for reconciliation activities

DMDs represent a valid model-based activity and its related plant dataset, acting as a map of the problem definition. The data in the DMT is used to create a DMD file or case study once the desired variables have been selected. Since the DMTs and DMDs are language-independent structures they can be incorporated within different PDEs.

Figure 2 sketches the interaction among entities. The *problem definition* is the first stage for the problem formulation; here, the PDE extracts from the DMT the available variables to be presented to the user who modifies their values according to his/her specific problem. Thereafter, the corresponding DMD is created using this information and the plant data. During the *problem translation* stage, the PDE constructs one or more problem input files (PIFs) which are a translation of the DMD into a language-specific high-level declaration guided by the semantic rules of the MSE under consideration. The last stage, the *problem initialization*, the problem is instantiated; this means that abstract model defined within the PDE is converted into concrete entities. Finally, the PME calls the necessary PMCs to execute

and solve the stated problem. The developed iOpSS is a PDE, and gPROMS (general PROcess Modelling System) is used as the MSE for the solution of the formulated problems. Therefore, one or more gPROMS input files (PIFs) are created for using during the actual problem solution stage (e.g. ESTIMATION, EXPERIMENT, PROCESS, and OPTIMIZATION entities).

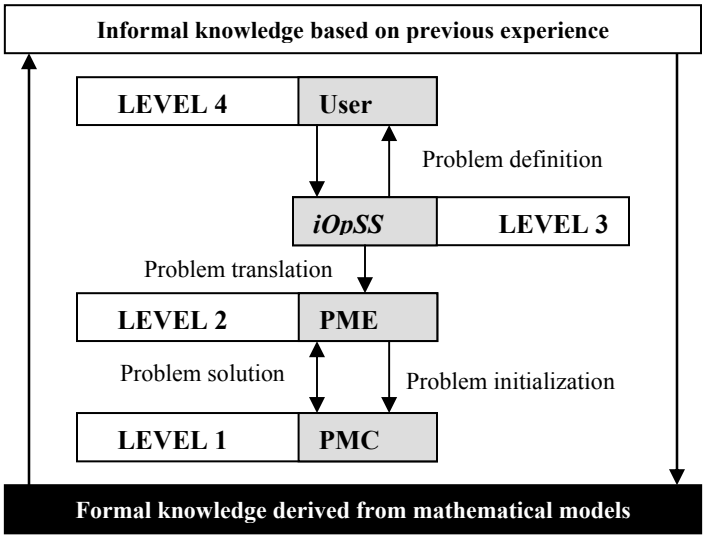

Figure 2. Interaction among entities in a model-based problem definition and solution. Adapted from Rolandi and Romagnoli (2004)

3.2 The environment of the iOpSS

The user-interface of the iOpSS has been designed in such a way that allows a flexible definition of complementary engineering problems. At the moment, such problems are represented within the iOpSS by four model-based activities (simulation, parameter estimation, data reconciliation, and optimization) sharing a common mathematical model of the system. Moreover, some activities may be combined to define more complex problems. For instance, it is possible to define problems of joint parameter estimation/data reconciliation, joint gross error estimation/data reconciliation, or joint parameter/gross error estimation/data reconciliation. Additionally, the environment allows the pre-processing of the data for further use in such activities, so that outliers can be removed from the dataset. Figure 3 depicts the main window of the iOpSS' environment.

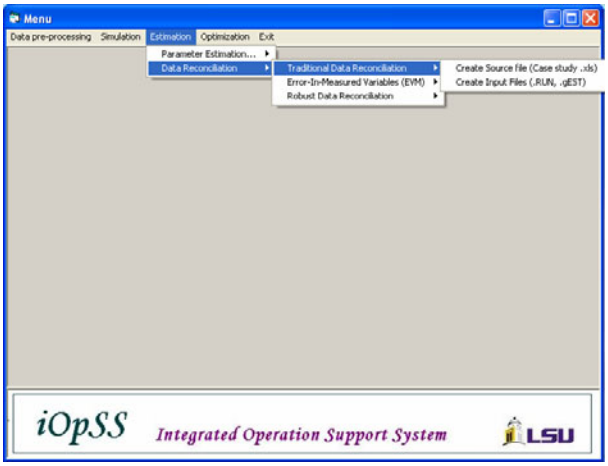

Figure 3. iOpSS user-interface

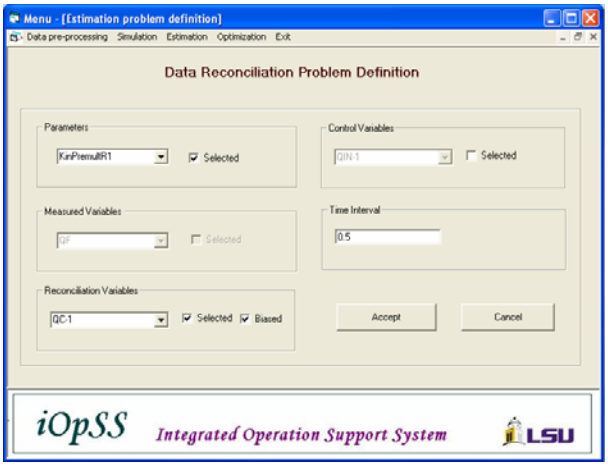

Figure 4. First stage in data reconciliation problem definition

Once an option has been selected, the application brings a window corresponding to the next level of information required. In the case of data pre-processing, for example, the user will be asked to select the desired window size and cleaning method. The estimation/reconciliation activities, on the other hand, require a series of consecutive windows for the selection of the estimation and reconciliation variables, and the statistical model for the variance, among others. Figure 4 shows the first stage in the data reconciliation activity.

3.3 The data pre-processing module

An outlier or gross error can be defined as a data point not representative of the statistical distribution where it belongs to. Usually, these observations have a strong influence in the analyses with their deletion causing significant changes in the estimates, confidence regions and other tests (Romagnoli and Palazoglu, 2005). To obtain better estimates, therefore, the outliers present in the data should be either eliminated, or incorporated to the bulk of data through the use of a suitable technique.

As it was mentioned previously, three methods were implemented in this module for the detection and rectification of outliers. All of them are based on a moving window which size is selected by the user. A more detailed description of these methods is presented next.

Median (MM) filter

The MM filter was first introduced by Tukey (1977), on his work on exploratory data analysis, and it is viewed as a smoothing technique. If a set of data points are equally spaced, it can be smoothed according to:

$$\text{Given data} = \text{Smooth} + \text{Rough} \tag{1}$$

The method is based on the selection of a subset of data from the sequence and the replacement of each element by its corresponding median value. The subset is moved throughout the entire sequence to replace every data point. In the medians of three, for example, three points are taken initially and the second element from them is replaced by their median. Then, the next three elements are taken into consideration with the first element being overlapped (the first element in the second subset is the last element from the first one), and the same procedure is applied. This is repeated until all data points from the sequence are considered.

The number of elements taken for the median calculation may be varied according to the type of data. It can be seen, however, that the application of this method leads to an inconvenience: the first and last point from the sequence, in the case of medians of three, do not have an assigned median. This problem was envisioned by Tukey, who also proposed several procedures to estimate the end values of the smoothed sequence. One of them, and the one used in this work, consists on selecting the median of three estimates: a) the actual end-value (no smoothed), b) the last smoothed value, and c) the result of an extrapolation to one step beyond the actual end-time.

The Mean Minimum Distance (MMD)

To detect outliers, this method bases in cluster theory, in which a set of data is divided in groups according to the distances of the elements to each other. Closer elements are said to belong to the same cluster, and therefore, they are more similar to each other than the elements from a different cluster (Chen and Romagnoli, 1998). For a dynamic process, the main structure is considered to be an elongated

cluster. When outliers are present, they are detected as points (or clusters) that do not belong to the underlying elongated one.

The distance of one object to its nearest neighbor is called the mean minimum distance (MMD), and it is the criteria to detect the main cluster. A data point (d-dimensional) is considered to be an outlier if the minimum distance between the measurement Y_i and any other in the window is less than twice the MMD (distance < 2 MMD). In a *d*-dimensional space, and considering that the variables may have different variation, the MMD is defined as:

$$MMD = \frac{1}{N}\sum_{i=1}^{N} \min_{i \neq j} \left[\left(\sum_{k=1}^{d} \frac{(y_{ik} - y_{jk})^2}{\upsilon_k} \right)^{1/2} \right] \tag{2}$$

Where N is number of objects ($Y_1, Y_2, ..., Y_N$), and υ_k is the k-th element of the covariance matrix.

On the contrary to the MM filter, the MMD is not applied to individual variables but to the entire dataset for a given time (d-dimensional space), this makes difficult the identification of individual outliers, this is, the determination of the variable which is contributing the most to the presence of the outlier. The knowledge of the specific variable contributing to the outlier is useful for further analysis, such as finding a faulty instrument or the model equation producing such errors. To address this issue, we propose the use of the MMD method in two stages. The first stage corresponds to the method as described above, meaning that the MMD is found for the d-dimensional space. The second stage is the application of the method to the individual variables so that the one contributing the most to the outlier is identified, and its value for that particular time is replaced with the median of the current window. It will be shown later that a better estimate is obtained when the criteria for the determination of the outliers is decreased from 2 to 0.5. This value is then considered as a tuning parameter.

The modified MMD for a one-dimensional space would look like:

$$MMD = \frac{1}{N}\sum_{i=1}^{N} \min_{i \neq j} \left[\left(\frac{(y_i - y_j)^2}{\upsilon_k} \right)^{1/2} \right] = \frac{1}{N}\sum_{i=1}^{N} \min_{i \neq j} \left(\frac{|y_i - y_j|}{\sqrt{\upsilon_k}} \right) \tag{3}$$

Modified MT filter

The MT filter method was first developed by Martin and Thompson (1982), to clean data that follows an autoregressive (AR) model. The algorithm is as follows:

1. Having the AR(p) model in state-space form, the filter computes robust estimates of the vector of measurements by means of a matrix M_t.
2. Matrix M_t is calculated recursively making use of the matrix of errors.
3. A robust prediction one-step ahead is calculated, and the cleaned data is estimated.

Liu et al (2004) proposed an extension of the MT filter-cleaner in which there is no need to know a priori the underlying model of the data. In their revised version of the algorithm, the AR(p) model is determined making use of the available data, and a moving window. The procedure is as follows:

1. Choose a dataset with a specific window size.

2. Select the order of the AR(p) model. This means to select the value of p.
3. Estimate the decorrelation model, using a robust variance and mean, and forming multivariate datasets for the calculation of the covariance matrix. The coefficients of the covariance matrix are used to calculate the autocorrelation coefficients. With these last coefficients, the Yule-Walker equations are solved to obtain the process model.
4. Construct the state-space form and apply the MT filter as described by Martin and Thompson.
5. Repeat the procedure for the next window until all the dataset has been cleaned.

3.4 The dynamic data reconciliation module

Data reconciliation refers to the process of obtaining better measurement estimates so that they comply with the mass and energy balances. The general dynamic estimation problem can be stated as (Rolandi, 2004):

$$\begin{aligned} &\min_{\theta,\beta,\omega,\gamma} \varphi(\tilde{z}(t), z(t), \sigma(t)) \\ &F(\dot{x}(t), x(0), y(0), p, \theta, \beta) = 0, \ \ t \in [0, t_f] \\ &I(\dot{x}(t), x(0), y(0), u(0), p, \theta, \beta) = 0 \\ &\sigma(t) = \sigma(\tilde{z}(t), z(t), \omega, \gamma), \ \ t \in [0, t_f] \\ &\theta^{\min} \le \theta \le \theta^{\max} \\ &\beta^{\min} \le \beta \le \beta^{\max} \\ &\omega^{\min} \le \omega \le \omega^{\max} \\ &\gamma^{\min} \le \gamma \le \gamma^{\max} \end{aligned} \tag{4}$$

The variable $\varphi(\cdot)$ represents the objective function to be optimized, and it is dependent on the model predictions, $z(t)$, the experimental observations, $\tilde{z}(t)$, and the variance model, $\sigma(t)$, which is at the same time a function of the parametric variables ω and γ, the model predictions and the experimental observations. $F(\cdot)$ represents the set of differential-algebraic equations (DAEs) with x and y denoting the differential (state) and algebraic variables respectively. $I(\cdot)$ denotes the initial conditions that must be satisfy by the model for the parametric, state, algebraic and input ($u(t)$ and p) variables. The last four equations correspond to the upper and lower bound of the decision variables (θ, β, ω and γ).

The objective function can take several forms depending on the nature of the mathematical model, and on the decision variables of interest. The latter will define the name given to the general problem stated by equations (4) to (8). For instance, when only the model parameters, θ, are taken into consideration the problem will be called *parameter estimation*; if the measurement biases, β, are the only decision variables involved, the problem will now be called *gross error estimation*. One of the most flexible objective functions used is the maximum likelihood function (which can be manipulated further to obtain the ordinary least square (OLS) or the weighted least square (WLS) functions):

$$\varphi(z(t), \tilde{z}(t), \sigma(t)) = \frac{1}{2}\left(N \ln(2\pi) + \sum_{j}^{NV} \sum_{k}^{NM_j} \left(\ln\left(\sigma_{j,k}^2\right) + \left(\frac{\tilde{z}_{j,k} - z_{j,k}}{\sigma_{j,k}} \right)^2 \right) \right) \tag{5}$$

The experimental observations, or measurements, $\tilde{z}(t)$, are related to the model prediction, or reconciled values $z(t)$ through a relationship of the form:

$$\tilde{z}(t) = z(t) + \varepsilon + \beta \tag{6}$$

Where ε and β represent the random errors and measurement bias correspondingly, and the latter is assumed to be constant for most practical applications. Hence, the measurement biases could be treated as parameters within the objective function. Furthermore, the estimation of the random errors, ε, should be included as decision variables in the general estimation problem defined by equation (4).

Classical or traditional reconciliation assumes that the input variables are error-free, $\varepsilon = 0$, in which case a distinction between input (*ip*) and output (*op*) variables is necessary:

$$z^{ip} = \tilde{z}^{ip} + \beta^{ip} \tag{7}$$

$$z^{op} = \tilde{z}^{op} + \varepsilon^{op} + \beta^{op} \tag{8}$$

However, error-free input variables may lead to biased estimators if this assumption is not met (Albuquerque and Biegler, 1995). Therefore, the incorporation of measurement errors in the input variables has been studied (Britt and Luecke, 1973, Kim et al, 1990, Albuquerque et al, 1997, Kim et al 1991, Albuquerque and Biegler, 1995) and was called error-in-variables method (EVM). Since error-free input variables cannot be assumed in most chemical processes, it is important for a framework used by industry engineers to be in accordance to this demand.

The iOpSS data reconciliation module possesses the flexibility to handle different estimation problems in both traditional and EVM reconciliation. Formulation of DR, parameter estimation and gross error estimation problems, or any combination of these activities, is possible upon request of the user.

3.5 Implementation procedure

Typically, an estimation problem is declared in gPROMS by means of the ESTIMATION and EXPERIMENT entities supported on the MODEL and PROCESS entities. As its name indicates, the MODEL entity provides the model equations against which the parameter estimation and data reconciliation are to be performed. The PROCESS entity provides initial values, and other information necessary for the problem to be well specified (i.e. degrees of freedom, initial states of transition networks and initial conditions). In the gPROMS language/architecture, the ESTIMATION entity describes the problem to solve, that is, the parametric variables/parameters to be estimated/decision variables, their bounds and the variance model of the measurement devices associated with experimental measurements. The EXPERIMENT entity contains all the measurements available for a given experimental run; the variance model can be specified here as well (gPROMS advanced user manual, 2006).

While parameter estimation activities are directly supported by gPROMS, traditional data reconciliation activities can be reformulated as a general estimation problem and, therefore, implemented and solved in gPROMS, as discussed in Rolandi (2004). On the other hand, the general EVM data reconciliation problem is not currently supported by the gPROMS language and, consequently, it was necessary to bend the language rules to succeed in the goal of using gPROMS as the modelling and solution engine (MSE) of this framework. Since data reconciliation is by itself an

optimization activity, this concept/notion was used to reformulate the reconciliation problem using gPROMS OPTIMIZATION entity. Naturally, several inconveniences arose from this choice; one of them was the incorporation of the measurements in the problem definition since the optimization activity in gPROMS does not support the EXPERIMENT entity. Concurrently, the MODEL entity was also modified to account for the objective function, the measurements, errors, and biases, in such a way that it could be used by the remaining activities (simulation, parameter estimation, traditional data reconciliation, and optimization).

The problem was solved in gPROMS by means of a sequential solution algorithm. Although at the moment the activities are performed assuming constant variance, the model can be extended to include other variance models.

4. CASE STUDY: TWO CSTRS IN SERIES

4.1 Problem description

The process was initially proposed by Hennin (1991) and reproduced in Bahri (1995). It consists in two CSTRs connected in series, and an additional stream of fresh feed is mixed with the output stream of the first reactor to conform the feed to the second reactor, as shown in figure 5.

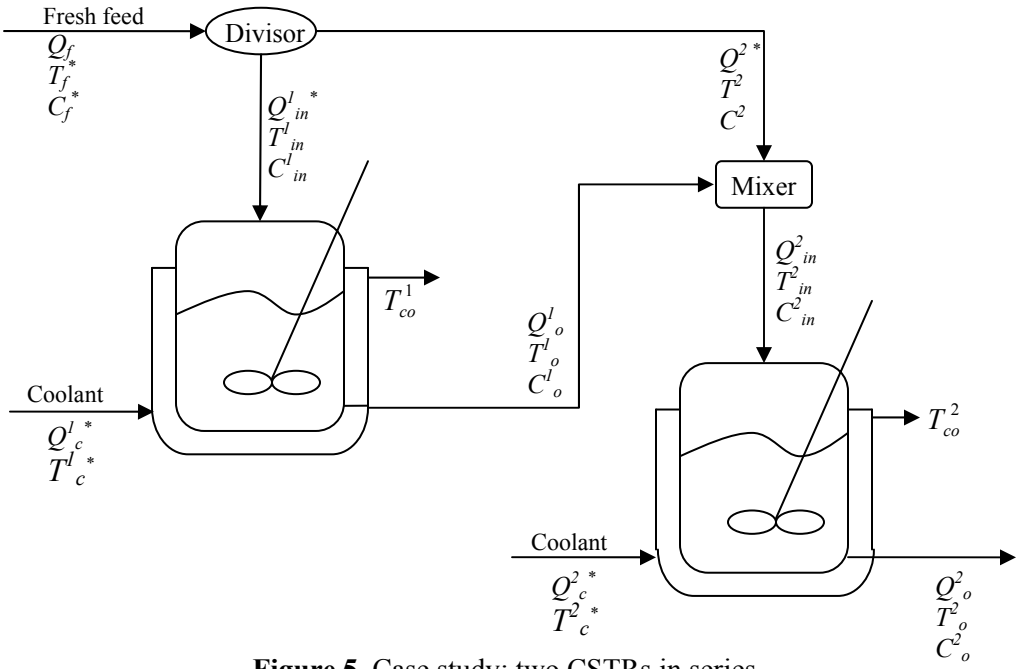

Figure 5. Case study: two CSTRs in series.

The simple exothermic reaction A → B is occurring in both reactors. The process was assumed to be at constant density and well-mixed. The model is the following:

Reaction rate:

$$r = k_o e^{-E/RT} \cdot C \qquad (9)$$

Mass balance of species A:

$$\frac{dC_A}{dt} = \frac{Q_{in}}{V}(C_{in} - C_o) - r \qquad (10)$$

Energy balance:

Overall mass balance:

$$\frac{dT}{dt} = \frac{Q_{in}}{V}(T_{in} - T_o) - \frac{r \cdot \Delta H_r}{\rho \cdot C_p} - \frac{U_a}{V \cdot \rho \cdot C_p}(T_o - T_{co}) \quad (11) \qquad Q_{in} = Q_o \quad (12)$$

Mass and energy balances in the mixer:

$$Q_{in}^2 C_{in}^2 = Q_o^1 C_o^1 + Q^2 C^2 \quad (13)$$
$$Q_{in}^2 T_{in}^2 = Q_o^1 T_o^1 + Q^2 T^2 \quad (14)$$

Where Q is Volumetric flow rate (m^3/h), T is Temperature (K), C *is* Concentration of species A ($kmol/m^3$), U_a is Heat transfer coefficient ($kcal/h.K$), ΔH_r is Heat of reaction ($kcal/kmol$), E is Activation energy ($kcal/kmol$), R is Gas constant ($kcal/kmol.K$), k_o is Reaction constant (h^{-1}), C_p is Heat capacity ($kcal/kmol.K$), ρ is Molar density ($kmol/m^3$), V is Reactor volume (m^3), and r is Reaction rate ($kmol/m^3.h$).

Table 1. Subscripts and superscripts

Subscript		Superscript	
f	Feed stream	*1*	Reactor 1
in	Input stream	*2*	Reactor 2
o	Output stream	*	Input variable
c	Coolant		

The process has 24 variables from which eight are input variables and four are state variables. The parameters of interest are the kinetic constant of reactor 1 (K_1) and heat transfer coefficient of reactor 2 (Ua_2). All variables were assumed to be measured.

Once the model was implemented in gPROMS language, a simulation was performed for a total time of 15 hours with a reporting interval of 6 min. The values of the input variables used to simulate the process are listed next:

$Q^1_{in} = 7\ m^3/h$ $\quad Q^1_c = 1\ m^3/h$ $\quad T^1_c = 283\ K$ $\quad T_f = 298\ K$
$Q_2 = 2\ m^3/h$ $\quad Q^2_c = 2.5\ m^3/h$ $\quad T^2_c = 283\ K$ $\quad C_f = 10\ kmol/m^3$

And the parameters:

$U_a = 750\ kcal/h.K$ $\quad k_o = 3.49308E7\ h^{-1}$ $\quad E = 11.843\ E3\ kcal/kmol$

Noise was then added to the simulated data to account for random variations. This dataset was modified later to add biases in three output variables and two input variables, and a proportion of 3.3% of outliers in all variables (generated randomly from Gaussian distributions), creating six new datasets as detailed in table 2:

Table 2. Datasets created and presence of measurement bias and outliers

Dataset name	Biased variables					Presence of outliers
	Q^1_o	T^2_o	C^2_{in}	Q^1_c	T_f	
NB						
NBO						x
B	x	x	x			
BO	x	x	x			x

BA	x	x	x	x	x	
BAO	x	x	x	x	x	x

The magnitude of the biases and their relative values are shown in table 3.

Table 3. Magnitude of measurement biases

Variable name	Bias magnitude	Relative value (%)[1]
Q^1_o	-1.0	20
T^2_o	12.0	4
C^2_{in}	0.9	10
Q^1_c	1.0	40
T_f	-10.0	4

[1] Reference value equal to the highest value reached by the variable during the time frame for time-varying variables or nominal value for time-invariant variables

As indicated by Rolandi (2006), the discrete representation (by the raw data pool) of the transient behavior of the of process variables forces the reconstruction of their continuous trajectories (Reconstruction of process trajectories, RPT). Moreover, the dimensionality of the plant data set is important when simulating combined discrete/continuous processes, and it should be reduced whenever possible. Therefore, the parameterization strategy and the nominal interval window should be carefully selected so that the data follows the dynamic nature of the variables and the accuracy of the solution is not affected in a great extent. In this work, the datasets were averaged to have an experimental value every 30 minutes, reducing its dimension from 150 to 30 points.

All outlier detection/filtering techniques were applied to each of the datasets containing outliers. The size of the moving window was 3 for the MM filter, 15 for the MMD and 5 for the modified MT filter. Each dataset was used in one or more activities according to the presence of biases, outliers or both, as indicated in table 4. The base case or case 0 corresponds to the reconciliation using the simulated dataset (NN) before introducing any error, bias or outlier.

Table 4. Datasets for traditional and EVM data reconciliation

Datasets	Case #		Parameters to estimate	Measurement biases to estimate				
	Traditional	EVM		Q^1_o	T^2_o	C^2_{in}	Q^1_c	T_f
B, BO	1	2						
B, BO	2	5	K_1, Ua_2	x	x	x	x	x
NB, NBO	3	1	none	x	x	x	x	x
BA	--	3	none	x	x	x		
BA	--	4	K_1, Ua_2	x	x	x		

4.2 Results

Data pre-processing

The efficiency of the modified MMD method using both criteria (twice the MMD, and 0.5 time the MMD), is shown in table 5. It can be seen that the second criteria increases the detection rate by approximately 159%. It is important to highlight that the window size of 15 was selected as the best one

after a series of trials previously performed. The efficiency as shown for the modified MMD is not suitable for the filters (MM, MT) because they change not only the outliers, but all the dataset.

Table 5. Efficiency of outlier detection/rectification methods

Method	Efficiency (detected/total)	Misidentification (detected but not outlier/total)
Modified MMD	0.37	0
Modified MMD criteria 0.5 (MMD-0.5)	0.95	0

Figure 6 shows the comparative behavior of the filtering techniques and the MMD method the state variable T_o^1. This is shown only for the NBO dataset since the other datasets behave in the same manner. A detailed section of the dataset is shown in figure 6a for a better picture. The results show that, in effect, the modified MMD has a lower performance than the MMD with a criterion of 0.5, and than the filtering techniques, which are able to clean more number of outliers. It can be seen that while the MM3 filter was able to bring all outliers to the bulk of the data, the modified MT filter was able only to reduce the magnitude of the outliers and degrading some of the data points following the outlier. This may be explained by taking into consideration the fact that the modified MT filter is based in an autoregressive model, which is a linear approximation, and therefore, may not be suitable for the model followed by this specific process.

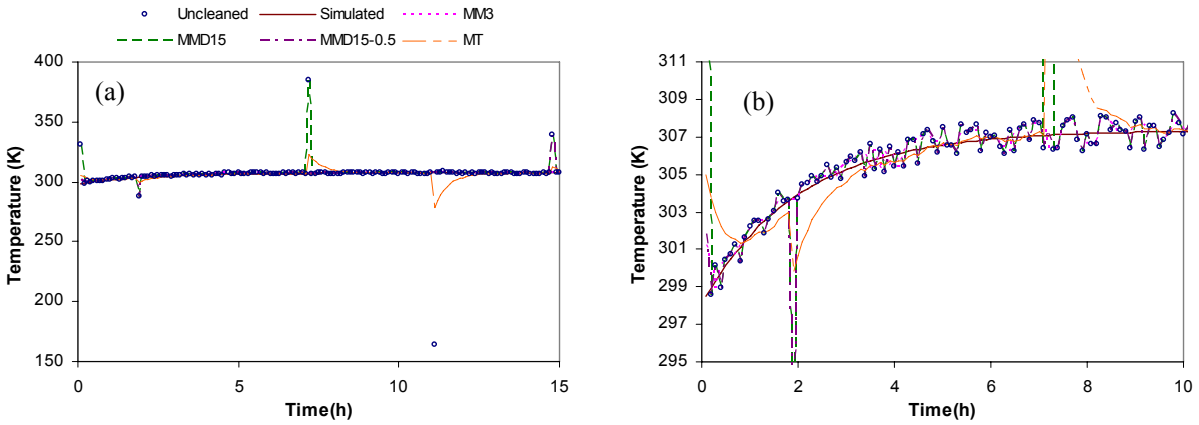

Figure 6. Outlier detection/cleaning performance for T^1_o (a)Complete dataset (b)Detail section

Data reconciliation

Table 6 shows the parameter and measurement bias estimates in the case of traditional data reconciliation for each case. It can be seen that the best estimates are obtained when no outliers were present (1 B and 2 B). If outliers were present, the MM3 filter and the MMD15-0.5 showed the best estimates. This is in agreement with the results for data pre-processing where these techniques presented a better performance than the other two methods.

While the kinetic constant is estimated with great accuracy in the traditional data reconciliation, the heat transfer coefficient is not, especially in those cases where the measurement bias estimates deviate considerably from their true values (case 2). As expected, the estimates differ significantly from

their true values when outliers are present and no cleaning method is employed for both reconciliation techniques (1 BO, 2 BO).

Table 6. Parameter and biases estimates in traditional data reconciliation

Case	File	Method	Parameters		Measurement biases		
			K^1_o	Ua_2	Q^1_O	C^2_{IN}	T^2_O
1	B	None			-0.995	0.896	11.783
	BO	None			-0.924	1.153	0.000
		MM3			-1.007	0.897	11.709
		MMD15			-1.013	1.070	10.559
		MMD15-0.5			-0.993	0.900	11.837
		MT			-0.903	0.992	9.654
2	B	None	3.493E+07	727.872	-0.996	0.897	11.673
	BO	None	3.480E+07	1190.366	-0.907	1.134	4.648
		MM3	3.493E+07	723.555	-1.005	0.897	11.588
		MMD15	3.493E+07	874.039	-0.918	1.075	10.873
		MMD15-0.5	3.493E+07	725.626	-0.990	0.899	11.723
		MT	3.492E+07	781.238	-0.920	0.998	9.991

As observed in table 7, the EVM exactly estimates the bias in all of three output variables and presents acceptable results in the bias of input variables. The kinetic constant was estimated correctly except for those cases where outliers were present and no cleaning technique was used (4 BAO, 5 BAO). However, the heat transfer coefficient was underestimated in all cases.

Table 7. Parameter and biases estimates in EVM data reconciliation

Case	File	Method	Parameters		Biases				
			K^1_o	Ua_2	Q^1_O	C^2_{IN}	T^2_O	T_f	$Q_c{}^1$
2	B	none			-1.000	0.900	12.000		
	BO	none			-1.000	0.900	12.000		
		MM3			-1.000	0.900	12.000		
		MMD15			-1.000	0.900	12.000		
		MMD15-0.5			-1.000	0.900	12.000		
		MT			-1.000	0.900	12.000		
3	BA	none			-1.000	0.900	12.000	-10.210	0.900
	BAO	none			-1.000	0.900	12.000	-9.398	0.900
		MM3			-1.000	0.900	12.000	-10.201	0.900
		MMD15			-1.000	0.900	12.000	-10.197	0.900
		MMD15-0.5			-1.000	0.900	12.000	-10.216	0.900
		MT			-1.000	0.900	12.000	-9.651	0.900
4	BA	none	3.4937E+07	600	-1.000	0.900	12.000	-10.189	0.900
	BAO	none	3.4862E+07	600	-1.000	0.900	12.000	-9.221	0.900
		MM3	3.4937E+07	600	-1.000	0.900	12.000	-10.175	0.900
		MMD15	3.4937E+07	600	-1.000	0.900	12.000	-10.211	0.900
		MMD15-0.5	3.4937E+07	700	-1.000	0.900	12.000	-10.213	0.900
		MT	3.4934E+07	600	-1.000	0.900	12.000	-9.640	0.900
5	B	none	3.4938E+07	700	-1.000	0.900	12.000		
	BO	none	3.4834E+07	700	-1.000	0.900	12.000		
		MM3	3.4943E+07	700	-1.000	0.900	12.000		
		MMD15	3.4937E+07	700	-1.000	0.900	12.000		

		MMD15-0.5	3.4942E+07	700	-1.000	0.900	12.000	
		MT	3.4936E+07	700	-1.000	0.900	12.000	

The values for the absolute errors are shown in tables 8 and 9 for the cases under study. The results presented in these tables confirm that the best results are obtained when no outliers were present (cases B and BA). When outliers were present, significant improvements on the absolute errors were observed for the MM filter and the modified MMD (MMD15-0.5). However, the norms are higher for the EVM so these trajectories are expected to deviate more from their true values.

Table 8. Absolute error for traditional data reconciliation

Case	File	Method	L1 norm	L2 norm
1	B	none	88.599	51.650
	BO	none	5132.304	262884.542
		MM3	113.040	64.366
		MMD15	911.276	5125.541
		MMD15-0.5	138.504	353.426
		MT	1081.512	6664.863
2	B	none	88.126	48.787
	BO	none	2202.266	29460.451
		MM3	111.070	61.122
		MMD15	936.207	5264.210
		MMD15-0.5	138.834	350.626
		MT	1077.610	6610.421

Table 9. Absolute error for EVM data reconciliation

Case	File	Method	Error L1 norm	Error L2 norm
2	B	none	672.951	1570.753
	BO	none	3664.189	82292.733
		MM3	693.609	1685.537
		MMD15	1853.106	20993.941
		MMD15-0.5	722.313	2033.022
		MT	2248.561	23061.834
3	BA	none	929.723	3759.426
	BAO	none	3476.611	74004.600
		MM3	943.690	3849.409
		MMD15	1933.950	21563.256
		MMD15-0.5	994.692	4359.326
		MT	2576.617	26709.173
4	BA	none	935.664	4465.523
	BAO	none	3457.352	73191.402
		MM3	953.148	4538.674
		MMD15	1982.884	23027.007
		MMD15-0.5	994.214	4522.045
		MT	2628.957	28415.164
5	B	none	683.495	1767.854
	BO	none	3645.476	82680.331
		MM3	698.964	1866.152
		MMD15	1866.543	21444.510

		MMD15-0.5	733.556	2246.867
		MT	2267.301	23590.083

The plant (uncleaned) vs. the reconciled data is shown in figures 7 to 10 for selected cases and the state variable T^{1}_{o}. It can be observed that when no errors are considered in the input variables and no biases are present, the reconciled values for the traditional and the EVM data reconciliations are practically the same (Figure 7). On the other hand, when no errors are considered in the input variables, the reconciled values using traditional DDR are slightly better than those corresponding to the EVM reconciliation (see fig. 8). However, the trajectories of the reconciled data for all reconciliation cases are very similar when MMD15 or MT was used (see fig. 9).

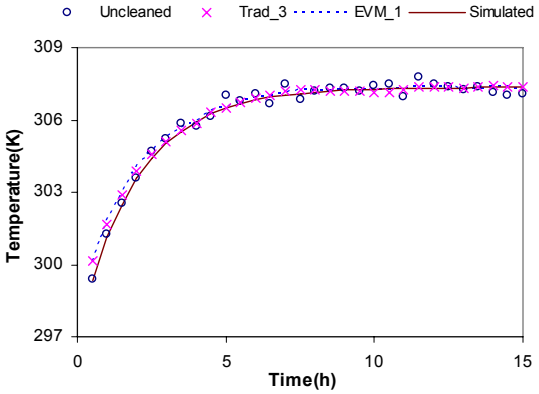

Figure 7. Reconciled data for case NB T^{1}_{o}

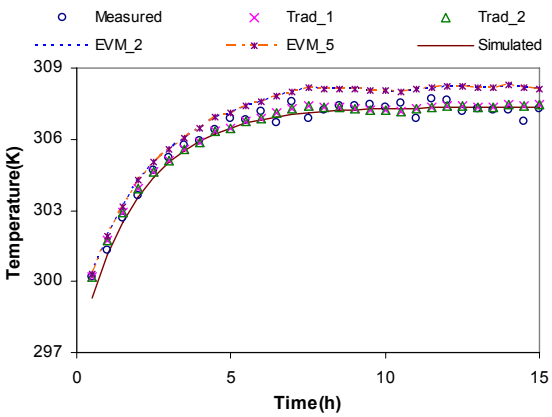

Figure 8. Plant vs reconciled data for cases bias (BO) MM3 of variable T^{1}_{o}

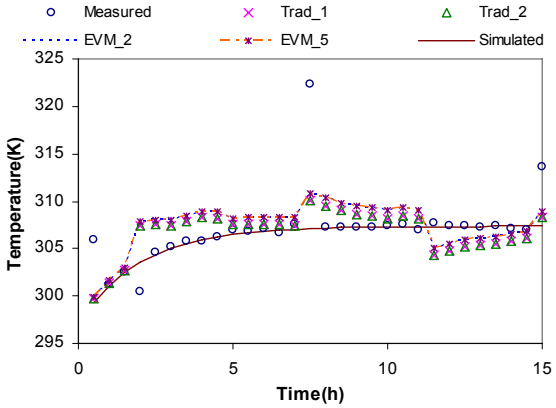

Figure 9. Reconciled data for cases bias all (BO) MMD15 of variable T^{1}_{o}

As it was expected from the data pre-processing module, the reconciled values are better for the datasets where the MM3 filter and the MMD15-0.5 were applied. This can also be seen from table 6, where the estimated parameters for the methods just mentioned, were closer to the nominal values.

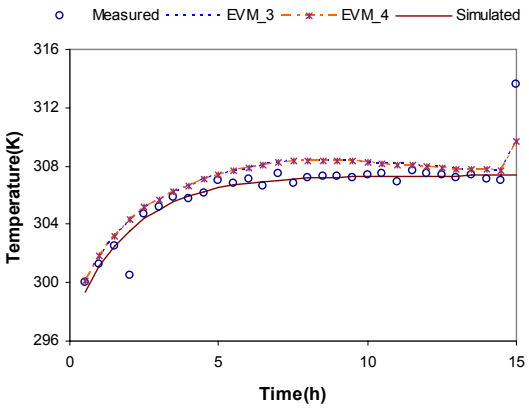

Figure 10. Plant vs. reconciled data for T^{1}_{o} , files BAO MMD15-0.5

5. CONCLUSIONS

Model-Centric Technologies (MCTs) have emerged as a powerful tool for the formulation and solution of model-based activities. Since the establishment of the CAPE community and with the emersion of MCTs, efforts towards the development of applications incorporating such features have increased. A novel framework for a unified and consistent formulation of model-based activities within a single environment initiated by Rolandi and Romagnoli (2006) is an excellent example of such efforts.

Two major extensions of the work initiated by Rolandi and Romagnoli were developed. The first one corresponds to the implementation of a pre-processing module in which the outliers could be decreased or removed. The mean minimum distance (MMD) method and its extension, the moving median (MM) filter and the modified MT filter were discussed. It was shown that the MM filter presents better performance in the cleaning of outliers, followed by the MMD when the criterion for the identification of an outlier is lowered from 2 to 0.5 times de mean minimum distance. This value is then considered as a tuning parameter for the adjustment of the detection rate. Although the modified MT filter indeed reduces the magnitude of the outliers, it showed to be less effective than the MMD and MM for this particular process. The non-linearity of the model might cause this behavior.

EVM (Error-in-variables method) dynamic data reconciliation was implemented as the first step towards the development of a more complete and robust framework for the definition of different model-based activities. The iOpSS in combination with gPROMS was able to adequately formulate and execute estimation and reconciliation activities for non-linear dynamic processes. Good measurement biases were estimated when estimation of parameters was not considered, and acceptable values were obtained when they were estimated in conjunction. It was shown that better estimates were obtained when the MM filter or the modified MMD method (criterion of 0.5 MMD) were used as cleaning techniques. When compared to the traditional data reconciliation, the EVM presents a decrease in the accuracy of parameter estimates although not in the measurement biases. Hence, the inclusion of robust methodologies should be considered as an additional extension to the framework. Another challenging task is the incorporation of capabilities for the formulation of on-line model-based control procedures. The former extension is currently being studied, and the latter is being considered for future work.

REFERENCES

Agiulli, F., S. Basta, and C. Pizzuti, "Distance-based detection and prediction of outliers," *IEEE Trans. Knowledge and Data Eng.*, **18**(2), 145 (2006).

Alburqueque, J.S., and L.T. Biegler., "Data reconciliation and gross-error detection for dynamic systems," *AIChE J.,* **42**(10), 2841 (1996).

Alburqueque, J.S., L.T. Begler, and R.E. Kass, "Inference in dynamic error-in-variable measurement problems," *AIChE J.*, **43**(4), 986 (1997).

Bahri, P.A., PhD. Thesis, Chemical Engineering Dept., The University of Sydney, Sydney, Australia (1995).

Beck, J.V. and K.J. Arnold, *Parameter estimation in engineering and science*, John Wiley & Sons, NY (1977).

Beckman, R.J., and R.D. Cook, "Outlier…..s," *Technometrics,* **25**(2), 119 (1983).

Braunschweig, B.L., C.C. Pantelides, H.I. Britt, and S., Sama, "Process modeling: the promise of open software architectures," *Chem. Eng. Prog.*, **96**(9), 65 (2000).

Britt, H.I. and Luecke, R.H. (1973). Estimation of parameters in nonlinear, implicit models. *Technometrics,* **15,** 233 (1973).

Chen, J., and J.A. Romagnoli, "A strategy for simultaneous dynamic data reconciliation and outlier detection," *Compt. Chem. Eng.,* **22**(4/5), 559 (1998).

Global CAPE-OPEN, final report. The GCO Consortium (2002). Taken from: http://www.ims.org/images/projects/gco/GCO-FinalReport.pdf (2005).

Kim, I.-W., M.J. Liebman, and T.F. Edgar, "A sequential error-in-variables method for nonlinear dynamic systems," *Compt. Chem. Eng.,* **15**(9), 663 (1991).

Kim, I.-W., M.J. Liebman, and T.F. Edgar., "Robust error-in-variables estimation using nonlinear programming techniques," *AIChE J.,* **36**(7), 985 (1990).

Liu, H., S. Shah, and W. Jiang, "On-line outlier detection and data cleaning," *Comp. Chem. Eng.,* **28**, 1635 (2004).

Martin, R.D. and D.J. Thompson, "Robust-Resistant spectrum estimation," *Proc. IEEE,* **70**(9), 1097 (1982).

PSE Ltda. (2006). gPROMS advanced user guide. Released 3.0.0.

Rolandi, P.A., PhD.Thesis, Chemical Engineering Dept., The Univeristy of Sydney, Sydney, Australia (2005).

Romagnoli, J.A. and M.C. Sánchez, *Data processing and reconciliation for chemical process operations,* Academic Press, London (2000).

Romagnoli, J.A., P.A. Rolandi, Y.Y. Joe, Z.Q. Ding, and K.V. Ling, "On data processing and reconciliation trends and the impact of technology," *International symposium on advanced control of chemical processes, ADCHEM*, 2006.

Romagnoli, J.A. and A. Palazoglu, *Introduction to process control,* CRC Press, Boca Ratón, FL (2006).

Tukey, J.W. *Exploratory data analysis,* Addison-Wesley, Reading, MA (1977).