Impacts of the assumptions used for the simultaneous data reconciliation and parameter estimation problem

Rosario, T. C.[1]**, Kalid, R. A.**[2]**, Santana, D. D.**[3]

Abstract The simultaneous solution of data reconciliation and parameter estimation problems (SDRPE) involves an optimization problem build on statistical hypothesis. It is possible to use the reconciled quantities and the estimated parameters for the optimization of process plants or statistical process control, so it is important to evaluate the statistical significance of these results, which depends on the validity of the aforementioned hypotheses. Thus, this work assesses the impacts of the assumptions adopted for the SDRPE, in addition to the statistical interpretation of the results especially the behavior of residuals, through a case study applied to a heat exchange process.

Keywords: data reconciliation; parameter estimation; residual analysis; weighted least squares

1 Introduction

The data reconciliation (DR) technique bases on adjusting a set of measured quantities through an optimization problem so that they obey the conservation laws and constraints of a system (Crowe, 1996). The adjusted data can then be used for estimating parameters (PE) of this system. However, the simultaneous resolution of the aforementioned problems (SDRPE) is the most reliable method when there are uncertainties in the measured quantities regarding dependent and independent variables (Tjoa and Biegler, 1992).

The solution to the problems of data reconciliation and parameter estimation, regardless of whether they are solved separately or simultaneously, involves an optimization problem. The most common is to use the weighted least squares (WLS) (Schwaab and Pinto, 2007), which for the simultaneous problem gives:

$$\underset{\boldsymbol{z_R},\boldsymbol{\theta}}{min}(\boldsymbol{z_o}-\boldsymbol{z_R})^T\boldsymbol{U}_{\boldsymbol{zz}}^{-1}(\boldsymbol{z_o}-\boldsymbol{z_R}), \tag{1}$$

subject to:

$$g(\boldsymbol{z_R},\boldsymbol{\theta})=0, \tag{2}$$

where $\boldsymbol{z_o}$ is the vector of observed quantities, $\boldsymbol{z_R}$ is the vector of reconciled quantities, $\boldsymbol{U_{zz}}$ is the covariance matrix of the observed variables, $\boldsymbol{\theta}$ is the vector of parameters of the model, and $g(\boldsymbol{z_{R,}}\ \boldsymbol{\theta})$ are the constraints, including the conservation laws.

The optimization problem, (1), (2) is based on assumptions (premises) adopted during the development of the objective function (1) from the maximum likelihood procedure.

It is possible to find in the literature works that discuss the problem of simultaneous data reconciliation and parameter estimation, such as Tjoa and Biegler (1992) and Francken et al. (2009). However, it is a

[1] Tarciso de Castro Rosario (e-mail: tarcisodecastro@hotmail.com)
Federal University of Bahia, Salvador, Bahia, Brazil.

[2] Ricardo de Araújo Kalid (✉e-mail: kalid@ufsb.edu.br)
Federal University Southern of Bahia, Itabuna, Bahia, Brazil.

[3] Daniel Diniz Santana (e-mail : daniel.diniz@ufba.br)
Federal University of Bahia, Salvador, Bahia, Brazil.

common approach to do not present a residual analysis to investigate whether the assumptions adopted for the construction of the objective function were satisfied, which can limit the interpretation of the results from a statistical perspective. Thus, the purpose of this work is to discuss the importance of performing a residual analysis for the SDRPE results and to present appropriate tools to perform it. From this, it is possible to evaluate the impacts on the statistical interpretation of the results through a case study

This work is organized as follows: Section 2 presents the development of the objective function from the maximum likelihood procedure highlighting its premises. Section 3 presents the statistical evaluation needed to evaluate if such premises are fulfilled after an estimation procedure. Section 4 details the methodology of the case study to exemplify the statistical evaluation. Section 5 presents the results within the statistical interpretation. Finally, Section 6 offers some concluding remarks.

2 Revisiting the objective function

The objective function of the SDRPE problem can be obtained from the maximum likelihood procedure (3), which aims to maximize the probability that the experimental data are as close as possible to its real values (Schwaab and Pinto, 2007; Bard, 1973).

$$\max L(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{t}) = \max p(\boldsymbol{z_o};\boldsymbol{z},\boldsymbol{U}_{zz},\boldsymbol{t}), \tag{3}$$

where $\boldsymbol{z}$ is the vector of the real (and unknown) values of the quantities, L is the likelihood function, p is a probability function and $\mathbf{t}$ is the time.

To solve (3) is necessary establish some assumptions. Table 1 describes the premises generally adopted.

Table 1. Description of the assumptions adopted for the development of the objective function.

Assumptions	Description
A1	The process is in steady state.
A2	The experiment is well done.
A3	The residues (difference between the observed data and the respective reconciliated values) are normally distributed.
A4	The model represents the process.
A5	The residues are not autocorrelated.

If the assumptions **A1**, **A2** and **A3** are valid: data in steady stated, trustworthy and adherents a Gaussian probability density function then form (3) can be written as follows:

$$p(\boldsymbol{z_o};\boldsymbol{z},\boldsymbol{U}_{zz}) = \frac{1}{\sqrt{2\pi\, det(\boldsymbol{U}_{zz})}} exp\left[-\frac{1}{2}(\boldsymbol{z_o}-\boldsymbol{z})^T\boldsymbol{U}_{zz}^{-1}(\boldsymbol{z_o}-\boldsymbol{z})\right]. \tag{4}$$

When considering the assumption **A4** the value of $\boldsymbol{z}$ can be described by $\boldsymbol{z_R}$, a value obtained from a representative model. Thus, (4) takes the following form:

$$p(\boldsymbol{z_o};\boldsymbol{z_R},\boldsymbol{U}_{zz}) = \frac{1}{\sqrt{2\pi\, det(\boldsymbol{U}_{zz})}} exp\left[-\frac{1}{2}(\boldsymbol{z_o}-\boldsymbol{z_R})^T\boldsymbol{U}_{zz}^{-1}(\boldsymbol{z_o}-\boldsymbol{z_R})\right]. \tag{5}$$

By replacing (5) in (3):

$$L(\boldsymbol{z_R},\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\, det(\boldsymbol{U}_{zz})}} exp\left[-\frac{1}{2}(\boldsymbol{z_o}-\boldsymbol{z_R})^T\boldsymbol{U}_{zz}^{-1}(\boldsymbol{z_o}-\boldsymbol{z_R})\right]. \tag{6}$$

The objective is to maximize (6). By applying the natural logarithm and performing mathematical operations on (6), the maximum likelihood procedures result in minimizing:

$$L(\boldsymbol{z_R},\boldsymbol{\theta}) = (\boldsymbol{z_o}-\boldsymbol{z_R})^T\boldsymbol{U}_{zz}^{-1}(\boldsymbol{z_o}-\boldsymbol{z_R}). \tag{7}$$

Therefore, using the objective function described by (7) in an optimization problem the statistical interpretations of the results implies that the results will satisfy the assumptions adopted in Table 1.

3 The importance of statistically evaluate of the reconciled values

According to Schwaab and Pinto (2007), estimation procedures do not end after the solution of the optimization problem, because it is necessary to analyze the quality of the results obtained through appropriate statistical tools. This strategy allows us to conclude whether the estimation procedure performed is satisfactory or there is a need to reevaluate the experimental data, process model, and assumptions, for example.

The statistical significance of the objective function used depends on the validity of the defined assumptions. Therefore, the results obtained are statistically valid if the defined assumptions are confirmed (Montgomery and Runger, 2003). According to Santana (2014), the fulfillment of the assumptions can be verified through the residual analysis applied to the obtained values.

Table 2 presents suggestions to check the validity of the adopted assumptions.

Table 2. Suggestions to check the validity of the adopted assumptions.

Assumptions	How to evaluate
A1	It must be verified by the analyst of the experiment.
A2	Cannot be proved, but only controlled by performing the experiments in a controlled and adequate way.
A3	Can be evaluated through hypothesis tests and graphical analysis.
A4	Can be ensured through conceptual and qualitative evaluation and through hypothesis tests and graphical analysis.
A5	Can be evaluated through hypothesis tests and graphical analysis.

4 Methodology

4.1 Case study

The case study discussed in this work is borrowed from Rosario et al. (2020) which compare two methods of solution (decoupled and simultaneous) for the SDRPE problem using a heat exchange system (Fig. 1). The decoupled solution solves both problems separately through a loop until convergence is achieved. In the simultaneous solution the problems are solved simultaneously by using a single optimization problem.

This work evaluates only the results obtained for the simultaneous solution which treats the SDRPE problem through a single optimization problem. In fact, the SDRPE is a coupled problem, i.e. all variables are included in a single objective function. The decoupled solution is just a mathematical artifice to help solve the problem (Rosario et al., 2020). Moreover, in order for the parameters to inherit the statistical information contained in the measurements, the data reconciliation and the parameter estimation must be done simultaneously instead of sequentially (Crowe, 1996).

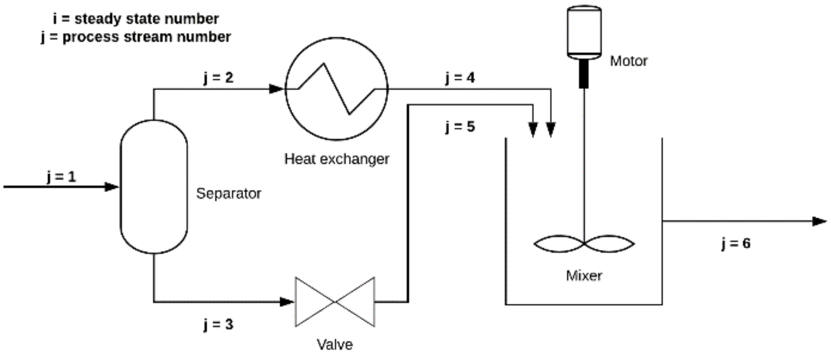

Fig. 1. Heat exchange process diagram. Font: adapted from Narasimhan and Jordache (2000).

4.2 Residual analysis

The residual analysis is performed based on statistical and graphical tools as depicted in Fig. 2.

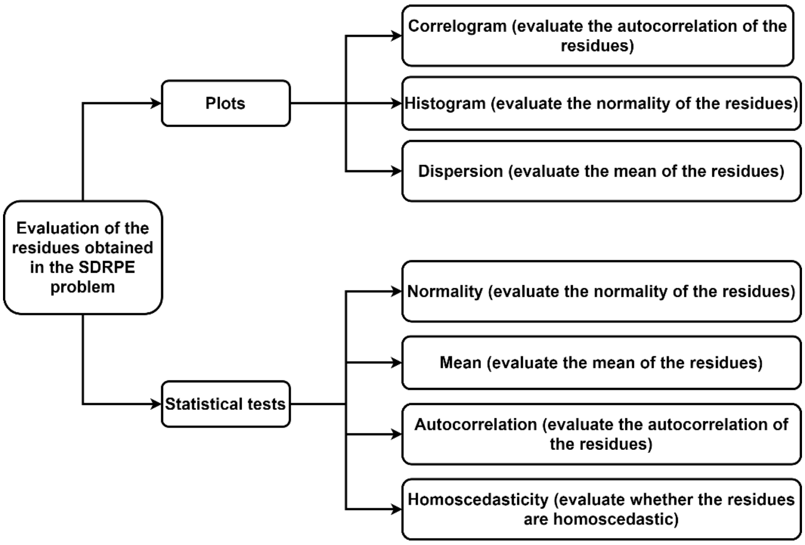
Evaluation of the residues obtained in the SDRPE problem
Plots
Correlogram (evaluate the autocorrelation of the residues)
Histogram (evaluate the normality of the residues)
Dispersion (evaluate the mean of the residues)
Statistical tests
Normality (evaluate the normality of the residues)
Mean (evaluate the mean of the residues)
Autocorrelation (evaluate the autocorrelation of the residues)
Homoscedasticity (evaluate whether the residues are homoscedastic)

Fig. 2. Tools used to evaluate the residues of the results obtained.

5 Results and discussion

5.1 *Assumptions A1 and A2*

The assumption of a steady-state process (A1) depends on the behavior of the studied process. So, it is assumed that the analyst knows this information properly. Regarding the assumption of a well-done experiment (A2), it is related to how the experiment is conducted and can be controlled through the planning experiments technique. Therefore, it is the responsibility of the analysts to ensure that these assumptions are valid.

5.2 *Assumptions A3 and A4*

The validity of assumptions regarding the normally distributed residues (A3) and the representative model (A4) is evaluated based on the residues obtained after the resolution of the SDRPE problem.

Fig. 3a shows that the histograms of the residues do not present adherence to a normal PDF (probability density function), because the behavior deviates from the expected distribution (red line), in addition Fig. 3b suggests that the mean of the residues is not equal to zero.

Statistical tests are performed to confirm such analysis and are presented in in Table 2, considering a 95% probability of coverage. It is confirmed that the residues are not adherent to a normal PDF. Additionally, the Ttest allows to confirm that the residues do not have a zero mean, and the homoscedasticity tests showed that probably the residues are not homoscedastic. So, the representative model assumption can be questioned.

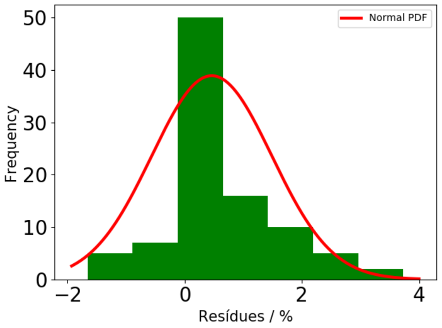

Fig. 3a. Histogram of the relative residues.

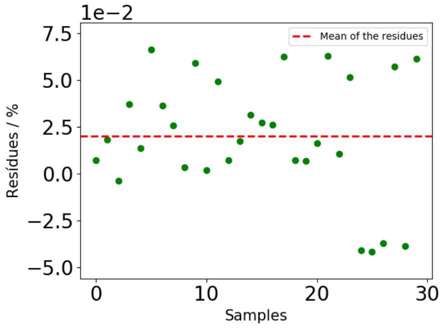

Fig. 3b. Relative residue dispersion for the model estimates.

Table 3. Results of the tests of normality, mean and homoscedasticity applied to residues.

Test	Null hypothesis	p-value
Lilliefors	Residues adhere to normality	$2{,}00 \times 10^{-15}$
Ttest	Mean equal to zero	$7{,}61 \times 10^{-6}$
White test (Lagrange multipliers)	Residues are homoscedastic	$1{,}22 \times 10^{-2}$
White test (Test F)	Residues are homoscedastic	$9{,}10 \times 10^{-3}$
Bresh Pagan (Test F)	Residues are homoscedastic	$1{,}79 \times 10^{-7}$

5.3 Assumption A5

The autocorrelation plots (Fig. 4) present peaks that exceeded the interval limits of 95% confidence, suggesting that the residues are autocorrelated. Consequently, there is no independence between the measurements. The results presented in Table 4, performed for 95% probability coverage, also allow questioning the validity of the assumption that the residues are not autocorrelated.

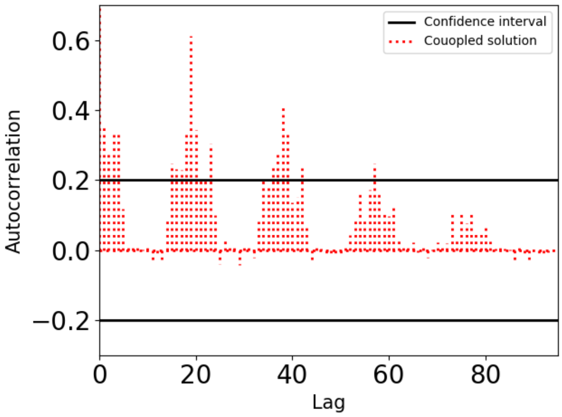

Fig. 4. Autocorrelation of the relative residues.

Table 4. Results of the autocorrelation tests applied to residues.

Test	Null hypothesis	p-value
Ljung-Box (Chi-2)	Not autocorrelated residues	0,059
Ljung-Box (Box-pierce)	Not autocorrelated residues	0,063

5.4 Impacts of residual analysis

The unsuitability of the assumptions **A3**, **A4**, and **A5** does not invalidate the results obtained, but makes it impossible to interpret the statistical significance of the adjusted model using the weighted least squares method. Alternatively, one can investigate the presence of outliers in the process and solve the problem again by removing such points or develop an objective function that does not incorporate such assumptions.

6 Conclusion

This work discussed the need for statistical analysis of the residuals after the simultaneous data reconciliation procedures and its impact on the interpretation of results. The case study exemplified such procedures in which it is not possible to confirm the assumptions applied to develop the objective function. In fact, three of the premises adopted for the development of the weighted least squares objective function are not met, so nothing can be concluded about the statistical significance of the results obtained for the case study evaluated. However, questioning the statistical significance of the results do not imply that they are wrong or that the adjusted model cannot be used.

As a suggestion for future work, one has: (a) the impacts of the of gross errors or outliers in the process on the fulfillment of such assumptions, (b) methodologies to compare the results obtained from different procedures, when conflicting conclusions are obtained from the hypothesis testing, and (c) development of objective functions that solve the Maximum Likelihood problem without incorporating the assumptions of normally distributed and not autocorrelated residues.

7 References

Crowe, C.M., 1996. Data reconciliation - Progress and challenges. J. Process Control 6, 89–98. https://doi.org/10.1016/0959-1524(96)00012-1

Francken, J., Maquin, D., Ragot, J., Bèle, B., 2009. Simultaneous data reconciliation and parameter estimation. Application to a basic oxygen furnace. IFAC Proc. Vol. 2. https://doi.org/10.3182/20090921-3-TR-3005.00017

M. Schwaab and J.C. Pinto, 2007. Análise de Dados Experimentais - Volume I Fundamentos de Estatística. e-papers, Rio de Janeiro.

Montgomery, D.C., Runger, G.C., 2003. Applied Statistics and Probability for Engineers, 3rd ed, European Journal of Engineering Education. John Wiley & Sons, United States of America. https://doi.org/10.1080/03043799408928333

Narasimhan, S., Jordache, C., 2000. Data Reconciliation & Gross Error Detection. Gulf Publishe Company, Houston.

Rosario, T.C., Kalid, R.A., Santana, D.D., 2020. Simultaneous Data Reconciliation and Parameter Estimation Applied to a Heat Exchange Process, in: Thomé, A.M.T., Barbastefano, R.G., Scavarda, L.F., dos Reis, J.C.G., Amorim, M.P.C. (Eds.), Industrial Engineering and Operations Management. Springer International Publishing, Cham, pp. 311–323.

Santana, D.D., 2014. Interpretação da região de abrangência na estimação de parâmetros.

Tjoa, I.B., Biegler, L.T., 1992. Reduced successive quadratic programming strategy for errors-in-variables estimation. Comput. Chem. Eng. 16, 523–533. https://doi.org/10.1016/0098-1354(92)80064-G

Yonathan, B., 1973. Nonlinear Parameter Estimation. Academic Press, New York.