Melhoria da Performance em Reconciliação de Dados pela Eliminação de *Outliers* com Pré-Filtro por Predição Linear

José M. Araújo[1],Marcelo Embiruçu[2],Cristiano H.O. Fontes[2], Luiz R.P. de Andrade Lima[2], Ricardo A. Kálid[2]

[1]Instituto Federal da Bahia, Depto. de Tecnologia em Eletro-Eletrônica, Grupo de Pesquisa em Sinais e Sistemas. Rua Emídio dos Santos, S/N, Barbalho, Salvador-BA, Brasil, jomario@ifba.edu.br

[2] Programa de Engenharia Industrial, Escola Politécnica da UFBA, Rua Aristides Novis, 02, Federação, Salvador-BA, Brasil, {embirucu,cfontes,lelo,kalid}@ufba.br

Resumo: Data Reconciliation (DR) is a very important tool when one wishes adjusting measurement data from mass and/or energy balances. Application of linear prediction filtering on the data corrupted with errors of category so called outliers, harden to be removed, is considered here. The proposed methodology is carried out by simulation in a stationary DR scheme for a industrial reactor, and the results outperforms that encountered in specialized literature.

Keywords: Data Reconciliation, linear prediction, outlier.

1. INTRODUÇÃO

Reconciliação de dados (RD) é uma ferramenta que, dentre uma gama de aplicações, permite o ajuste ótimo de medidas e estimativas (valores mapeados), com base em redundância espacial e no modelo, geralmente balanços de massa e/ou energia [1]. Tópicos atuais sobre RD incluem reconciliação de dados em regime transitório ou dinâmica [2,3] e detecção de erros grosseiros simples e múltiplos [4]. A reconciliação de balanços de massas global em estado estacionário consiste em encontrar a solução do seguinte problema de otimização:

$$\min_{\hat{F}} \left(F - \hat{F} \right)^T W^{-1} \left(F - \hat{F} \right) \qquad s.a. \quad A\hat{F} = \mathbf{0} \tag{1}$$

Em que $\hat{F}$ é o vetor de estimativa fornecida pela RD, F é o vetor de variáveis medidas e W é a matriz diagonal de covariâncias (incertezas) e A é a matriz de incidências, resultante de balanço de massa. Restrições de desigualdade inerentes ao modelo são incluídas, de forma que a solução obtida atenda-as. Em reconciliação de dados estática, um tipo de restrição bastante comum é a de balanços de massa global, que fornece um conjunto de equações lineares, e o problema é chamado de RD linear em regime permanente. Restrições de balanço de energia ou concentração, por exemplo, tornam o problema não-linear (bilinear, trilinear, etc.). Em reconciliação de dados dinâmica, as restrições são formadas por um conjunto de equações diferenciais, por exemplo, as equações de estado do modelo dinâmico.

Um problema de importância central em reconciliação de dados é que a presença de erros sistemáticos (*bias*) ou do tipo *outliers*, pode afetar a solução de maneira severa, caso em que os valores reconciliados terão mérito mais pobre que os valores mapeados. Especificamente, no caso de *outliers*, caracterizados por valores espúrios que contaminam algumas amostras, diversos trabalhos tem contribuído com metodologias para sua detecção/eliminação [5,6]. A filtragem dos dados, anterior à reconciliação, também é bastante empregada [4], mas problemas inerentes a técnicas de filtragem, como atrasos, tornam-se inconvenientes e ainda podem comprometer a qualidade da RD, e outliers, por sua natureza, não são de fácil eliminação por filtragem clássica.

O presente trabalho propõe o uso de filtro baseado em detecção de ruídos impulsivos, bastante empregado em recuperação de gravações degradadas [7], cuja filosofia é baseada na detecção de erro após reconstrução dos dados por predição linear [8-10].

2. PRELIMINARES

Seja a sequência de dados y_k. A reconstrução desta sequência utilizando predição linear de ordem N tem a forma [8]:

$$\hat{y}_k = -\sum_{p=1}^{N} a_p y_{k-p} \tag{2}$$

O erro de predição, a cada amostra, é dado por:

$$e_k = y_k - \hat{y}_k = y_k + \sum_{p=1}^{N} a_p y_{k-p} \tag{3}$$

Os coeficientes de predição linear $a_1,...,a_N$, são determinados de forma a minimizar a média quadrática de e_k, dentro da janela de comprimento M. A solução do problema é bem conhecida e tem a forma [9]:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} r_0 & r_1 & \cdots & r_{N-1} \\ r_1 & r_0 & \cdots & r_{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N-1} & r_{N-2} & \cdots & r_0 \end{bmatrix}^{-1} \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix} \quad (4)$$

Em que:

$$r_k = \sum_{j=0}^{N-1-k} y_j y_{k+j} \quad (5)$$

Para uma sequência de baixo conteúdo espectral, e_k tem comportamento suave. Porém, a presença de descontinuidades na seqüência de dados introduz picos no erro de predição. Isto também pode ser deduzido examinando a função de transferência entre o erro de predição e a seqüência verdadeira:

$$H(z) = \frac{E(z)}{Y(z)} = 1 + \sum_{p=1}^{N} a_p z^{-p} \quad (6)$$

Este é um modelo de um filtro passa-altas. Este fato pode ser utilizado para indicar a presença de descontinuidades bruscas, que tem larga aplicação em recuperação de gravações degradadas [7], utilizando um algoritmo de detecção cuja idéia principal é utilizada para concepção da metodologia do presente trabalho no pré-processamento dos dados para reconciliação. A Fig. 1 apresenta um exemplo de seqüência numa janela de comprimento 25, contendo uma variação brusca em uma de suas amostras. A seqüência em questão é aproximada usando predição linear de 2ª ordem, e o erro é mostrado na Fig. 2. É bastante simples a localização de tal amostra, até mesmo por inspeção visual. Na Fig. 3, observa-se a resposta em freqüência para o ganho do filtro calculada de acordo com a eq. 5.

3. METODOLOGIA PROPOSTA

A metodologia utilizada em [7] foi adaptada para detecção de outliers, com posterior aplicação de reconciliação de dados. O algoritmo utiliza dois limiares para varredura do erro de predição, LL (limiar de localização) e LD (limiar de detecção). O critério para determinação destes limiares é bastante subjetivo [7], mas é razoável utilizar 90% dos valores da seqüência e_k, devidamente ordenada, e estabelecer um valor de referência sobre a mediana destes valores. Levando-se em conta a existência de incertezas gaussianas em uma dada medida, adota-se aqui um único limiar L, considerando que *outliers* tem curta duração, na absoluta maioria dos casos apenas uma amostra:

$$L = 2\eta$$
$$\eta = \sqrt{madiana(e_{ks}^2)} \quad (7)$$

Em que e_{ks} é resultante do truncamento em 90% da sequência e_k ordenada de forma não decrescente. Então, uma amostra é considerada corrompida com *outlier* quando

$$|e_k| > L \quad (8)$$

Uma vez estabelecido este critério, as amostras corrompidas são descartadas, e podem ser substituídas utilizando alguma técnica de interpolação. No presente caso,

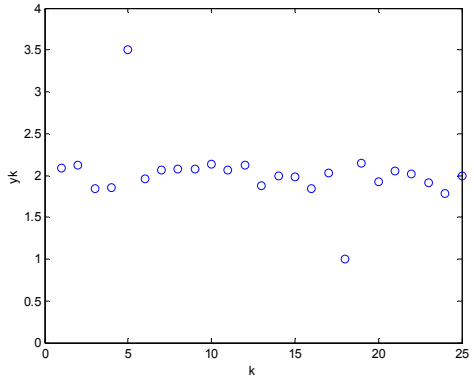

Figura 1 – Sequência de dados contendo incerteza gaussiana e duas amostras corrompidas com *outliers*.

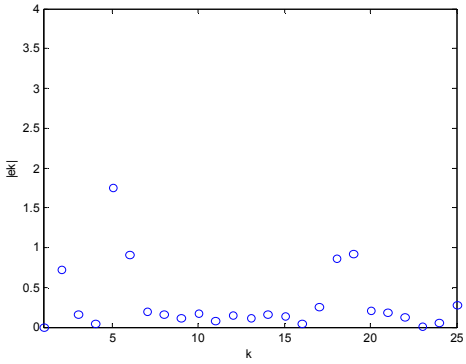

Figura 2 – módulo do erro de predição linear de ordem 2 para a sequência yk

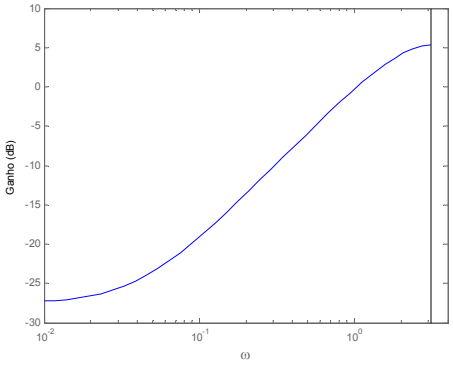

Figura 3 – resposta em frequência do filtro passa altas $E(z)/Y(z)$.

uma interpolação simples por média (linear) é suficiente, pois a quantidade de informação perdida é pequena.

Para as primeiras amostras, dois problemas são críticos: o atraso próprio do filtro e a possibilidade de existência de *outliers* nas primeiras amostras. Para contornar estes problemas, a técnica de detecção é aplicada duas vezes, uma na janela em sentido direto e a segunda vez com a reversão da janela. Na secção à seguir, o método é utilizado junto com um problema de reconciliação da dados, e é verificada a melhoria dos resultados em relação à reconciliação sem a aplicação da técnica.

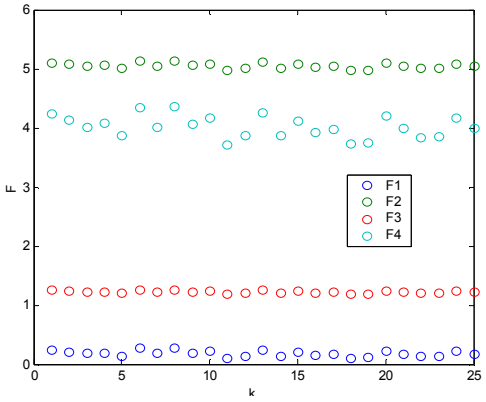

Figura 4 – Vazões mapeadas para o reator de quatro correntes de exemplo 4.

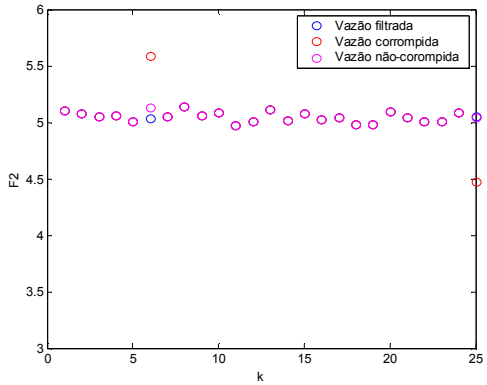

Figura 5 – Vazão da corrente 2 de entrada com outlier simulado para $k = 6$ e $k = 25$.

4. EXEMPLO E DISCUSSÃO

O exemplo mostrado nesta seção foi apresentado em [6] onde um reator com quatro correntes, duas de entrada e duas de saída, é considerado, em que a matriz de balanço de componentes, a matriz de ponderação (covariâncias das vazões) e os valores verdadeiros de vazão são, respectivamente:

$$A = \begin{bmatrix} 0,1 & 0,6 & -0,2 & -0,7 \\ 0,8 & 0,1 & -0,2 & -0,1 \\ 0,1 & 0,3 & -0,6 & -0,2 \end{bmatrix}$$

$$W = \begin{bmatrix} 0,0029 & 0 & 0 & 0 \\ 0 & 0,0025 & 0 & 0 \\ 0 & 0 & 0,0006 & 0 \\ 0 & 0 & 0 & 0,04 \end{bmatrix}$$

$$T = \begin{bmatrix} 0,1739 & 5,0435 & 1,2175 & 4,00 \end{bmatrix}^T$$

Os valores mapeados numa janela de 25 amostras foram gerados com base em $\boldsymbol{T}$ e $\boldsymbol{W}$, utilizando o software SIMULINK. A Fig. 4 mostra estas sequências. Para aplicação da metodologia, as amostras 6 e 25 da corrente 2 foram contaminadas com *outliers*, conforme a Fig. 5, sendo os valores corrompidos dados abaixo:

$$\boldsymbol{F}_6^c = [0,2652\ 5,5868\ 1,2590\ 4,3392]^T$$

$$\boldsymbol{F}_{25}^c = [0,1709\ 4,4735\ 1,1742\ 3,9856]^T$$

A reconciliação de dados foi feita para a amostra 25 da vazão da corrente 2 considerando dois casos: sem contaminação ($\overline{F}_{25}$) e após contaminação ($\overline{\overline{F}}_{25}$). Os resultados obtidos foram:

$$\overline{F}_{25} = [0,1737\ 5,0363\ 1,2157\ 3,9943]^T$$

$$\overline{\overline{F}}_{25} = [0,1621\ 4,7004\ 1,1346\ 3,7279]^T$$

Claramente, a contaminação com *outlier* introduz um impacto severo no resultado da reconciliação, trazendo uma estimativa de baixa qualidade para as vazões do processo em estudo. A aplicação da metodologia proposta com posterior reconciliação de dados resulta em:

$$\overline{\overline{F}}_{25}^{f} = [0,1738\ 5,0394\ 1,2164\ 3,9968]^T$$

Que é um resultado consistente com as características do modelo. A tabela 1 exibe a comparação entre os resultados obtidos sem detecção de *outliers* (S), com a presente metodologia (MP) e das propostas em [6] (C), baseada em gráficos QQ e transformações não-lineares de limitação e em [11] (RS) que utiliza eliminação serial pela detecção de erros grosseiros. É possível notar que os resultados são muito consistentes entre si, o que torna claro o mérito da metodologia abordada.

5. CONCLUSÕES

Uma metodologia para melhoria de desempenho em reconciliação de dados foi apresentada. A mesma baseia-se em filtragem dos dados antes da reconciliação, por meio de técnicas de detecção e reconstrução dos dados por predição linear. Um exemplo foi apresentado, em um caso típico de reconciliação de dados linear em regime permanente, e os resultados foram de boa qualidade, compatíveis com outras metodologias classicamente utilizadas. O potencial da metodologia para uso em reconciliação dinâmica é muito grande, e deverá ser investigado em próximos trabalhos.

AGRADECIMENTOS

Os Autores gostariam de agradecer às suas Instituições e Grupos, e também aos revisores anônimos pelas valiosas sugestões para melhoria do manuscrito.

REFERENCIAS

[1] Kuehn, D. R. and Davidson, H. Computer control. II. Mathematics of control. Cem.Eng.Progress, v. 57, p. 44-47, 1961.

[2] Zhou, L., Su, H., and Chu, J. A study of nonlinear dynamic data reconciliation. International Conference on Systems, Man and Cybemetics. In: Proceedings of the IEEE International Conference on Systems, Man and Cybemetics 02, 1360-1365, 2004.

[3] Lou, K.-Y. and Huang, H.-P., A wavelet enhanced integral approach to linear dynamic data reconciliation. Journal of Chem.Eng.Jpn., v. 38, p. 1035-1048, 2005

[4] Narasimhan, S. and Jordache, C. Data reconcilation & gross error detection: An intelligent use of process data, 2000.

[5] Chen, J. and Romagnoli, J. A. A strategy for simultaneous dynamic data reconciliation and outlier detection. Computers Chem.Eng., v. 22, p. 559-562, 1998.
[6] Chen, J., Romagnoli, J. A., and Bandoni, A. Outlier detection in process plant data. Cem.Eng.Progress, v. 22, p. 641-646, 2007.
[7] Biscainho, L. W. P. Restauração digital de sinais de áudio provenientes de gravações musicais degradadas. Tese de Doutorado, COPPE-UFRJ, Rio de Janeiro, 2000.
[8] Makhoul, J. (1975). Linear prediction: a tutorial review. Proceedings of the IEEE, v. 63, p. 568-580.
[9] Mendel, J. M. Lessons in Estimation Theory for Signal Processing: New Jersey, Prentice-Hall, 1995
[10] Hayes, M. Statistical Digital Signal Processing and Modeling: New York, John Wiley and Sons, 1996.
[11]Romagnoli, J. A. e Stephanopoulos, G. Rectification of processing and treatment data in the presence of gross errors. Chem. Engng Sci. 36, 1849-1863, 1981.

Tabela 1: comparação entre métodos de reconciliação de dados para a amostra $k = 25$.

Método	Vazões Reconciliadas	Desvios (%)
S	$[0{,}1621\ 4{,}7004\ 1{,}1346\ 3{,}7279]^T$	$[-6{,}7855\ -6{,}8028\ -6{,}8090\ -6{,}8025]^T$
MP	$[0{,}1738\ 5{,}0394\ 1{,}2164\ 3{,}9968]^T$	$[-0{,}0575\ -0{,}0813\ -0{,}0903\ -0{,}08]^T$
C	$[0{,}1742\ 5{,}0508\ 1{,}2192\ 4{,}0058]^T$	$[0{,}1725\ 0{,}1447\ -0{,}3778\ 0{,}1450]^T$
RS	$[0{,}1751\ 5{,}0775\ 1{,}2256\ 4{,}0270]^T$	$[0{,}6900\ 0{,}6741\ 0{,}66539\ 0{,}6750]^T$

Valor verdadeiro: $[0{,}1739\quad 5{,}0435\quad 1{,}2175\quad 4{,}00]^T$

Legenda: S – sem eliminação de outliers; MP – metodologia proposta; C – Método de Chen; RS – Método de Romagnoli e Stephanopoulos