Estrategias para Reconciliación de Datos Robusta

Claudia Llanos[1], Mabel Sánchez[1], Ricardo Maronna[2],

[1]Planta Piloto de Ingeniería Química, UNS-CONICET, Camino de La Carrindanga km 7,
8000 Bahía Blanca, Argentina
{cllanos, msanchez,} @plapiqui.edu.ar

[2]Dpto. de Matemática - Facultad de Ciencias Exactas,
Universidad Nacional de La Plata, Calle 50 y 115, La Plata 1900, Argentina.
rmaronna@retina.ar

Resumen. La optimización de las industrias químicas requiere de técnicas efectivas para reconciliación de datos y estimación de parámetros. Los estimadores clásicos se ven altamente influenciados por la presencia de valores atípicos en las mediciones, en cambio los estimadores robustos son insensibles a pequeñas cantidades de éstos. En este trabajo se presentan nuevas metodologías robustas para la reconciliación de datos que combinan las fortalezas de los estimadores monótonos y redescendientes. Se proponen dos estrategias de distinto grado de complejidad, las cuales se aplican para tres modelos de mediciones diferentes. Las medidas de desempeño consideradas son el error cuadrático medio, la cantidad promedio de Errores Tipo I y la potencia global. Se compara el desempeño de las estrategias propuestas para procesos cuya operación se representa mediante modelos lineales y no lineales.

Palabras claves: Reconciliación de datos, Estadística Robusta, Optimización en línea

1 Introducción

La reconciliación de datos y estimación de parámetros son tareas claves para la optimización de las industrias químicas en tiempo real. Dado que la presencia de errores sistemáticos en las observaciones distorsiona los valores de las mediciones reconciliadas y las estimaciones de los parámetros y variables no medidas, se han presentado diversas estrategias para reducir el efecto de dichos errores [1].

En los últimos años se desarrollaron metodologías de reconciliación de datos basadas en estimadores robustos. El enfoque robusto del análisis estadístico de datos se basa en formular estimadores que producen resultados confiables si los datos siguen una determinada distribución de probabilidad, o solo lo hacen de manera aproximada debido a la presencia de valores atípicos [2].

Una revisión crítica de la literatura de reconciliación de datos robusta indica que se han desarrollado estrategias adaptivas que producen estimadores de alta eficiencia, aun cuando la distribución de los errores presenta colas pesadas[3,4,5]. Sin embargo, tanto los resultados teóricos como numéricos en Estadística Robusta indican que el incremento en el esfuerzo computacional requerido por los estimadores adaptivos no

produce una mejora real en el desempeño[2], por lo tanto no serán aplicados en este estudio.

En este trabajo se presentan metodologías robustas para la reconciliación de datos que combinan las fortalezas de los estimadores monótonos y redescendientes. Los M-estimadores monótonos tienen solución única. En cambio los redescendientes son más robustos porque presentan un punto de quiebre mayor, y son más eficientes cuando se satisface el modelo del error y este corresponde a una distribución de colas pesadas. Entre los diversos tipos de M-estimadores redescendientes se seleccionan los de la familia bisquare.

Las técnicas desarrolladas se aplican para estimar los valores de las variables medidas y no medidas de procesos químicos, y también para identificar los valores atípicos. El desempeño de las metodologías propuestas se analiza utilizando ejemplos de aplicación presentes en la literatura sobre el tema. Las medidas de desempeño son el error cuadrado medio (MSE), la cantidad promedio de Errores Tipo I (AVTI) y la potencia global (OP).

2 Planteo del Problema

Consideremos que la operación de una planta operando bajo condiciones de estado estacionario se describe mediante el conjunto f de ecuaciones algebráicas no lineales

$$f(x,u)=0\ , \tag{1}$$

siendo $x= (x_1...,x_n)$ y $u=(u_1...,u_m)$ los vectores de las variables del proceso medidas y no medidas, respectivamente.

En ausencia de errores sistemáticos, se considera el siguiente modelo para las mediciones

$$y_{ij} = x_i + e_{ij}\ , \tag{2}$$

donde y_{ij} representa la medida de la variable i-ésima en el período de tiempo j-ésimo, x_i es el valor verdadero de dicha variable, y e_{ij} indica el error aleatorio de dicha medición. Se asume que los errores aleatorios son independientes, tienen media cero y desvío estándar σ_i. Se denota $y_j = \ y_{1j}, \dots, y_{nj}$ al vector de observaciones del intervalo de tiempo j.

El procedimiento de Reconciliación de Datos Clásica consiste en resolver un problema de cuadrados mínimos (LS) ponderados sujeto a restricciones, las cuales involucran a los balances de masa y energía del proceso considerado (Ec. 3). Se obtienen estimadores más precisos de las variables medidas redundantes y estimaciones para las variables no medidas observables en función de los valores ajustados de las mediciones redundantes y de los valores medidos de las mediciones no redundantes. Los estimadores son consistentes con las ecuaciones de conservación del proceso.

$$\begin{aligned} &\underset{x}{\text{Min}} \quad \left(y_j - x_j\right)^T \psi^{-1} \left(y_j - x_j\right) , \\ &\text{st.} \\ &\qquad f(x,u)=0 . \end{aligned} \tag{3}$$

Dado que los estimadores obtenidos utilizando la técnica LS son afectados por la presencia de valores atípicos en las mediciones, se han desarrollado estimadores robustos. Estos son insensibles a una proporción moderada de errores sistemáticos. Una familia de estimadores robustos muy popular es la familia de los M-estimadores [2].

Si r_{ij} es el residuo de y_{ij} definido por la Ec. 4, entonces un M-estimador del estado del sistema en el tiempo t se define como la solución (x, u) del problema de optimización representado por la Ec. 5

$$r_{ij} = \frac{(y_{ij} - \hat{x}_i)}{\sigma_i} , \tag{4}$$

$$\begin{aligned} &\underset{x}{\text{Min}} \quad \sum_{j=t-N+1}^{t} \sum_{i=1}^{n} \rho\left(r_{ij}\right) , \\ &\text{st.} \\ &\qquad f(x,u)=0 , \end{aligned} \tag{5}$$

donde ρ es la función de pérdida. Esta es creciente con $|r|$ y el caso $\rho\ r\ = r^2$ corresponde al estimador de la técnica LS. Se asume que el proceso se observa para un horizonte de datos de tamaño *N*; esto es, en el tiempo *t* el estimador se basa en los vectores de observaciones $y_{t-N+1}, \ldots, y_t$.

2.1 M-estimadores

Si llamamos ψ a la derivada de ρ , esto es $\psi = \rho'$, se pueden distinguir tres categorías importantes de M-estimadores:

A) El primer tipo corresponde a estimadores para los cuales la función de influencia ψ es una función creciente de r. Estos M-estimadores se llaman monótonos, ρ es una función convexa y, por lo tanto, no acotada. El caso más popular es el estimador de Huber, para el cual

$$\psi(r) = \begin{cases} r & \text{si } |r| \le c \\ c\,\text{sgn}(r) & \text{sino} \end{cases} , \tag{6}$$

donde sgn denota el signo de la función y c es una constante que regula la eficiencia del estimador. Los casos $c\to\infty$ y $c\to 0$ se asocian a los estimadores obtenidos mediante LS y a los estimadores de desviaciones absolutas mínimas, respectivamente.
B) El segundo tipo corresponde a los estimadores con función de pérdida no acotada y ψ tendiendo a cero en el infinito. Comprende estimadores de máxima verosimilitud para distribuciones de colas pesadas. Un caso importante es el asociado a la familia de distribuciones T (o Student), para el cual

$$\psi(r)=\frac{r}{r^2/c+1} \quad , \tag{7}$$

donde c corresponde a los grados de libertad. En general, si ψ tiende a cero en el infinito los estimadores se llaman redescendientes.
C) El tercer tipo corresponde a una ρ acotada. Estos estimadores también son redescendientes. Un caso importante es el de la función bisquare

$$\psi(r)=\begin{cases} r(1-(r/c)^2)^2 & \text{si } |r|\le c \\ 0 & \text{sino} \end{cases} \quad , \tag{8}$$

dado que $\psi(r)=0$ para $|r|>c$, estos estimadores rechazan completamente los valores atípicos.
De la comparación de las cualidades de estos tres tipos de estimadores surge que:

- ✓ Estimadores Monótonos (A): la solución de la Ec.5 tiene un solo mínimo local, por lo tanto el valor empleado para comenzar el proceso de iteración influye en el número de iteraciones pero no en el valor final. Por otro lado, no son insensibles a los valores atípicos grandes, y por esto pueden tener baja eficiencia para distribuciones de errores de colas pesadas.
- ✓ Estimadores redescendientes (B y C): Son eficientes para distribuciones de errores de colas pesadas. Pero el problema 5 puede tener varios mínimos locales, por lo que se requiere un buen punto inicial para asegurar la obtención de una buena solución.
- ✓ Estimadores con ρ acotada (C): rechazan completamente los valores atípicos grandes, y con una adecuada elección de la constante c puede alcanzar una alta eficiencia tanto para la distribución de error normal como para las distribuciones de colas pesadas.

3 Estrategias de Reconciliación Robusta

La revisión crítica de las metodologías robustas de reconciliación de datos motiva el desarrollo de aproximaciones no adaptivas que combinen las fortalezas de los M-

estimadores monótonos y redescendientes. Entre ellos, se selecciona la función bisquare.

Se proponen dos estrategias, llamadas Método Simple (SiM) y Método Sofisticado (SoM).

3.1 Método simple

Compuesto por las siguientes dos etapas:

a) Etapa 1: Para cada medición i en el tiempo t se calcula un M-estimador de localización y_{it} basado en valores del horizonte $\{y_{ij}, j = t - N + 1, \dots, t\}$, el cual es la solución de

$$\underset{y}{\text{Min}} \sum_{j=t-N+1}^{t} \rho_{bis}\left(\frac{y_{ij}-\tilde{y}_i}{\sigma_i}\right), \tag{9}$$

donde ρ_{bis} corresponde a la familia bisquare. Se trata cada medición separadamente. De esta manera se dispone de un estimador inicial robusto y simple, llamado mediana de $\{y_{t-N+1}, \dots, y_t\}$, que permite aprovechar la redundancia temporal provista por las observaciones repetidas y_{ij} en el horizonte. Cabe notar que para la estimación de x_i el uso de redundancia proveniente de observaciones distintas a i es limitado (Maronna y Arcas [6]).

b) Etapa 2: El estimador se define como la solución (x, u) del Problema 10

$$\underset{x}{\text{Min}} \sum_{i=1}^{n} \rho_{Hub}\left(\frac{\tilde{y}_i-\hat{x}_i}{\sigma_i}\right),$$
$$\text{st.} \tag{10}$$
$$f(x,u)=0\ ,$$

donde ρ_{Hub} corresponde a la familia Huber.

3.2 Método Sofisticado

Incorpora una etapa extra al SiM que es la siguiente:

c) Etapa 3: usando como punto inicial la solución del Problema 10, se resuelve el Problema 11

$$\underset{x}{\text{Min}} \quad \sum_{j=t-N+1}^{t} \sum_{i=1}^{n} \rho_{bis}\left(\frac{y_{i,j}-\hat{x}_i}{\sigma_i}\right),$$

$$\text{st.} \qquad f(x,u)=0\ . \tag{11}$$

En este caso todas las mediciones del horizonte se usan en el problema de optimización.

4 Análisis Comparativo

Se lleva a cabo un análisis de desempeño de las estrategias robustas propuestas y el método clásico de reconciliación para dos procesos industriales. Estos corresponden a la red de vapor (SMN) de una planta de síntesis de metanol [7] y a una red de intercambio de vapor (HEN) [8].

El primer ejemplo involucra 28 corrientes que interconectan 11 unidades (ver Fig. 1), y se consideran medidos los caudales másicos de todas las corrientes. Se asume que el desvío estándar de los caudales es 2.5% del valor verdadero.

El segundo proceso (Fig. 2) involucra 15 corrientes que intercambian calor. Este caso comprende 30 variables (15 caudales másicos y 15 temperaturas) de las cuales 16 son medidas y 14 no medidas. Se formulan 17 balances de masa y energía alrededor de los intercambiadores de calor, mezcladores y divisores. Se asume que el desvío estándar de las temperaturas es 0.75K y los desvíos estándares de los caudales son el 2% de sus valores verdaderos.

Para comparar la capacidad de estimación y de detección/identificación de mediciones atípicas de las distintas técnicas, la eficiencia de los M-estimadores se fija en 95.5% con un ajuste de parámetros. Además el punto de corte de cada técnica, que es el valor a partir del cual las mediciones son consideradas atípicas, se ajusta por prueba y error de manera tal que el AVTI sea aproximadamente 0.05, cuando las mediciones no tienen errores sistemáticos.

Se presentan tres casos de estudio para los dos ejemplos de aplicación. En los casos de estudio 1 y 2, los errores de las mediciones se generan usando distribuciones F= N(0,1) y F=(1-ε) N(0,1)+εN(0,K^2) con ε=0.1 y K ∈ {2,5,10}, respectivamente. Para el caso de estudio 3, las mediciones se generan al azar agregando errores de valor constante ($R\sigma_i$) a los valores de las variables verdaderas con probabilidad ε=0.1. El rango de los valores de R es [1-5]. Se efectúan 10000 pruebas de simulación para cada caso de estudio, y la ventana del horizonte de datos se fija en N=10.

Las medidas de desempeño usadas en este análisis son: MSE, AVTI y OP, las cuales se estiman como se muestra a continuación:

$$\text{MSE}=\frac{1}{\text{SIM}}\sum_{k=1}^{\text{SIM}}\sum_{i=1}^{n}\left(\frac{\hat{x}_i-x_i}{\sigma_i}\right)^2, \tag{12}$$

$$AVTI=\frac{\#(\text{errores sitematicos incorrectamente identificados})}{SIM}\ , \qquad (13)$$

$$OP=\frac{\#(\text{errores sistematicos correctamente identificados})}{\#(\text{errores sistematicos simulados})}\ , \qquad (14)$$

donde SIM es el número de simulaciones.

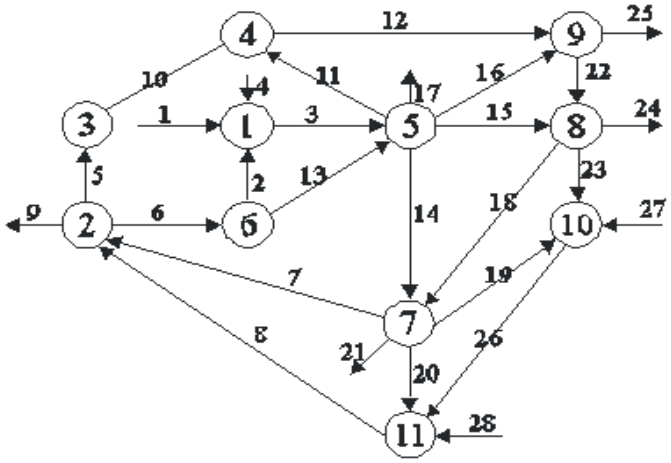

Fig. 1. Red de Vapor.

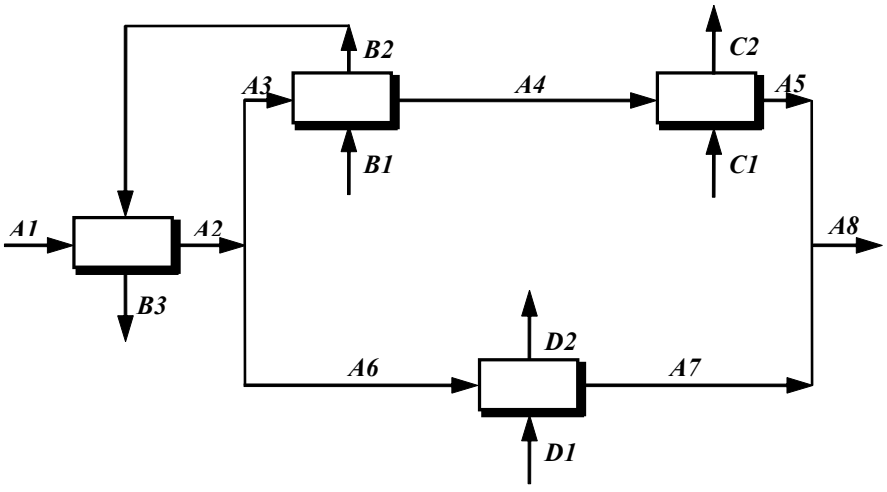

Fig. 2. Redes de intercambio de calor.

5 Resultados y Discusión

Las Tablas 1, 2 y 3 presentan los resultados de las simulaciones para cada caso de estudio.

Tabla 1. Resultados del caso de estudio 1.

	AVTI			MSE x 10^2		
	LS	SiM	SoM	LS	SiM	SoM
Ej1	0.0500	0.0499	0.0499	6.074	6.384	6.387
Ej2	0.0499	0.0501	0.0497	8.102	8.516	8.517

Tabla 2. Resultados del caso de estudio 2.

		AVTI			OP			MSE x 10^2		
	K	LS	SiM	SoM	LS	SiM	SoM	LS	SiM	SoM
Ej1	2	0.0505	0.0517	0.0519	0.0556	0.0552	0.0552	7.889	7.539	7.515
	5	0.1429	0.0524	0.0516	0.4400	0.4381	0.4383	20.558	7.953	7.899
	10	1.4892	0.0484	0.0484	0.6985	0.6978	0.6979	66.012	7.679	7.635
Ej2	2	0.0534	0.0493	0.0491	0.0654	0.0614	0.0615	9.423	9.354	9.337
	5	0.1155	0.0497	0.0493	0.4591	0.4506	0.4511	18.621	9.647	9.602
	10	0.8417	0.0483	0.0489	0.7105	0.7062	0.7066	51.402	9.449	9.409

Tabla 3. Resultados para el caso de estudio 3.

		AVTI			OP			MSE x 102		
	R	LS	SiM	SoM	LS	SiM	SoM	LS	SiM	SoM
Ej1	1	0.0461	0.0490	0.0419	0	0	0	6.983	7.719	7.777
	2	0.0625	0.0693	0.0693	0	0	0	11.585	12.705	12.683
	3	0.0995	0.0941	0.0902	0.0002	0.0016	0.0016	19.270	15.506	14.831
	4	0.1689	0.1205	0.1084	0.2492	0.5761	0.6033	30.038	12.527	11.539
	5	0.3072	0.1346	0.1278	0.9003	0.9934	0.9946	43.890	9.338	8.928
Ej2	1	0.0479	0.0493	0.0490	0	0	0	8.677	9.376	9.406
	2	0.0581	0.0603	0.0597	0	0	0	11.743	12.759	12.762
	3	0.0832	0.0716	0.0716	0.0013	0.0034	0.0035	16.849	14.770	14.410
	4	0.1310	0.0831	0.0724	0.4184	0.6824	0.7030	23.996	12.904	12.068
	5	0.2212	0.0864	0.0736	0.9279	0.9916	0.9957	33.185	10.719	9.968

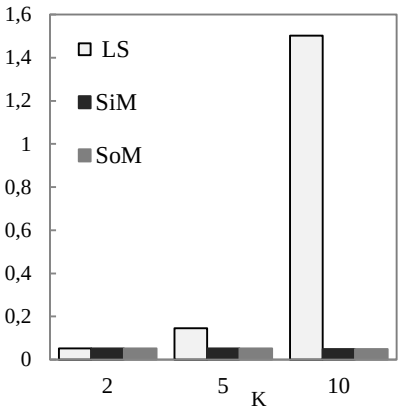

Fig. 3a. AVTI del SMN para distintas contaminaciones de la normal.

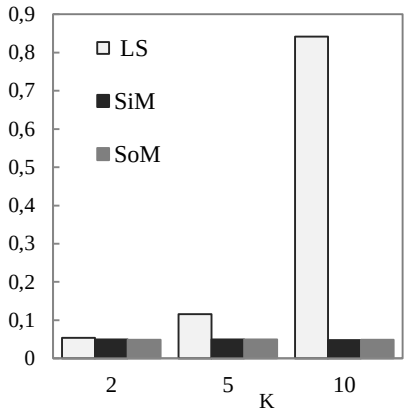

Fig. 3b. AVTI del HEN para distintas contaminaciones de la normal.

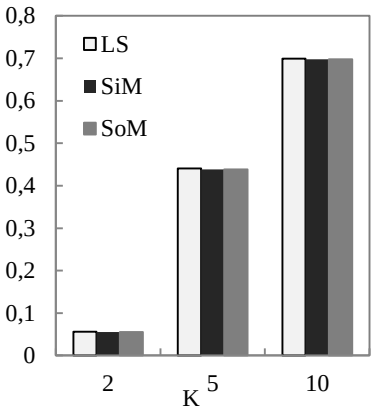

Fig. 4a. OP del SMN para distintas contaminaciones de la normal.

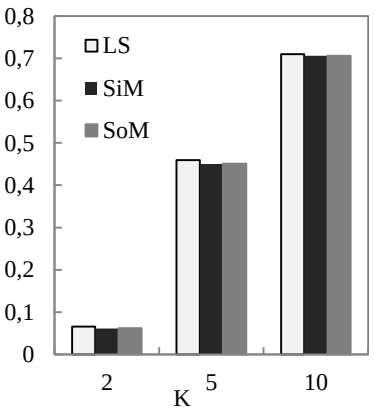

Fig. 4b. OP del HEN para distintas contaminaciones de la normal.

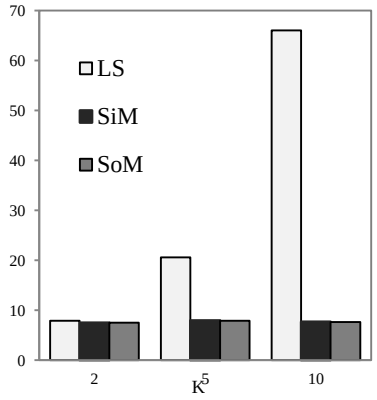

Fig. 5a. MSE* 10^2 del SMN para distintas contaminaciones de la normal.

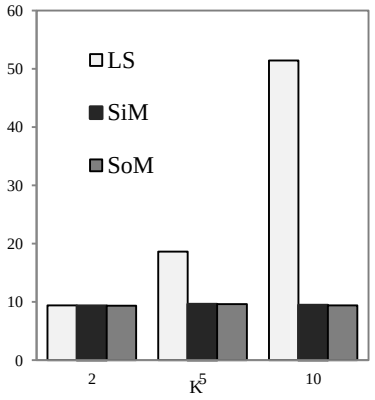

Fig. 5b.MSE*10^2 del HEN para distintas contaminaciones de la normal.

En las siguientes figuras se observan las medidas de desempeño analizadas, para el caso lineal y no lineal, con mediciones generadas al azar.

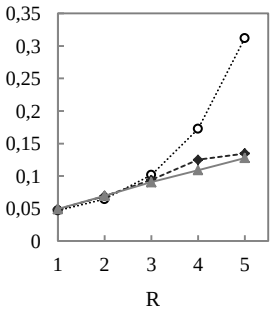

Fig. 6a. AVTI del SMN, LS(○), SiM(♦) y SoM(▲).

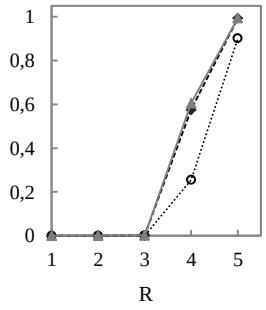

Fig. 6b.OP del SMN, LS(○), SiM(♦) y SoM(▲).

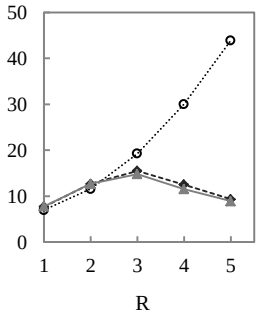

Fig. 6b.MSE*10 ^{2} del SMN, LS(○), SiM(♦) y SoM(▲).

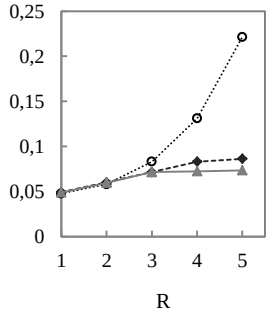

Fig. 7a. AVTI del HEN, LS(○), SiM(♦) y SoM(▲).

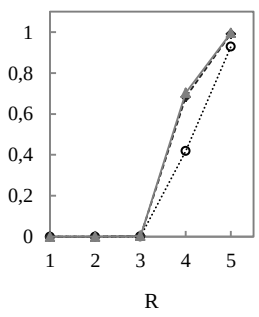

Fig. 7b.OP del HEN, LS(○), SiM(♦) y SoM(▲).

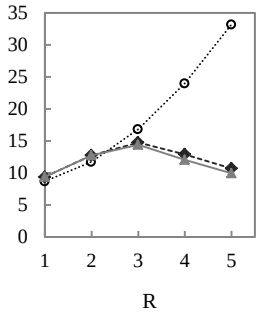

Fig. 7c.MSE*10 ^{2} del HEN. LS(○), SiM(♦) y SoM(▲).

Los resultados del Caso de Estudio 1 (mediciones con distribución normal) se presentan solo con el fin de mostrar que el AVTI de los tres procedimientos es aproximadamente igual a 0.05 cuando la distribución de probabilidad de las mediciones es la ideal. Tal como es de esperar, el MSE de la estrategia LS es inferior al de los métodos SiM y SoM, mientras que dicha medida de desempeño presenta valores similares para los métodos robustos.

Con respecto al Caso de Estudio 2 (mediciones con distribución normal contaminada), la capacidad de identificar correctamente los errores sistemáticos aumenta con el incremento del parámetro K, y es similar para los tres métodos. En cambio el valor del AVTI se incrementa con K para la técnica LS y se mantiene en valores cercanos a los obtenidos cuando no se simulan errores sistemáticos para los métodos robustos. En general, el mayor valor del MSE se obtiene para la estrategia LS, luego le sigue el correspondiente al SiM y finalmente el del SoM. Si bien esta relación de orden se mantiene para K=2, las tres medidas de desempeño son cercanas. En cambio, para mayores valores de K, el MSE de la estrategia LS es notablemente superior al de los métodos robustos.

Para el Caso de Estudio 3 (mediciones con distribución normal con la adición al azar de errores sistemáticos de magnitud constante con probabilidad ε=0.1) se observan comportamientos similares de las tres técnicas hasta R=3, luego se evidencia un mejor comportamiento de los métodos robustos y una ligera superioridad del SoM con respecto al SiM.

Con respecto al esfuerzo computacional, la relación entre el tiempo de corrida para SoM y SiM es aproximadamente 4.7 para el ejemplo lineal y 1.5 para el no lineal. Se utilizó el enfoque de regresión lineal presentado por Maronna y Arcas [6] para resolver el Ejemplo 1, y se empleó el algoritmo de Programación Cuadrática Sucesiva del paquete Matlab para abordar la solución de los problemas de optimización no lineal resultantes del Ejemplo 2.

6 Conclusiones

En este trabajo se presentaron dos procedimientos, de diferente grado de complejidad, destinados a resolver el problema de reconciliación de datos robusta en estado estacionario para procesos cuya operación se representa mediante sistemas lineales y no lineales. Los procedimientos combinan las fortalezas de los M-estimadores monótonos y redescendientes. Al comparar estos procedimientos con el método de cuadrados mínimos queda en evidencia la insensibilidad de los M-estimadores robustos a la presencia de valores atípicos.

Se analizó el desempeño de las estrategias tanto para la estimación de las variables medidas y no medidas, como para la identificación de valores atípicos. Los resultados indican que el desempeño del SoM es ligeramente superior al del SiM. Sin embargo el el SiM requiere un tiempo de cómputo notablemente menor para el caso lineal, por lo que constituye una alternativa de solución atractiva para este tipo de sistemas.

Referencias

1. Romagnoli, J., Sánchez, M.: Data Processing and Reconciliation for Chemical Process Operations. Academic Press, San Diego (2000)
2. Maronna, R., Martin, R.D., Yohai, V.: Robust Statistic: Theory and Methods. John Wiley and Sons Ltd., Chichester (2006)
3. Arora, N., Biegler, L.T.: Redescending estimators for data reconciliation and parameter estimation. Comp. & Chem. Eng. 25, 1585--1599 (2001)
4. Zhang Z., Shao, Z., Chen, X., Wang, K., Qian, J.X.: Quasi-weighted least square estimator for data reconciliation. Comp. & Chem. Eng. 34, 154--162 (2010)
5. Chen, J., Peng, Y., Munoz, J.C.: Correntropy estimator for data reconciliation. Chem. Eng. Sci.104, 1019--1027 (2013)
6 Maronna, R.A., Arcas, J.: Data reconciliation and gross error detection based on regression, Comp. & Chem. Eng. 33, 65--71 (2009)
7. Serth, R., Heenan, W.: Gross error detection and data reconciliation in steam-metering systems. AIChE J. 32: 733--741(1986)
8. Swartz, C.L.E.: Data reconciliation for generalization flowsheet applications. 197th Natl. Meet., Am. Chem. Soc., Dallas, TX (1989).