Estrategias de Análisis y Exploración de Datos Como Soporte a la Operación y Supervisión de Procesos Químicos

Por

Rodolfo Vicente Tona Vásquez.

Tesis presentada como requisito parcial para el grado de
Doctor por la Universitat Politècnica de Catalunya

Dirigida por:

Dr. Antonio Espuña Camarasa

Prof. Dr. Luís Puigjaner Corbella.

Departament d'Enginyeria Química

Escola Tècnica Superior d'Enginyeria Industrial de Barcelona

Universitat Politècnica de Catalunya

Barcelona, Septiembre del 2006

Resumen

En esta tesis se presenta un conjunto de metodologías que intentan facilitar la tarea de explotación de la información contenida en los datos históricos de proceso y como reaprovecharlos de manera que se produzca un impacto positivo en la operación del proceso.

Se comienza por atacar el problema de asegurar la calidad de los datos. Se hace una revisión de los métodos de filtración univariable, en especial los basados en técnicas wavelets, ya que estos últimos se han mostrado en la literatura particularmente ventajosos para el filtrado de datos. Se establece, mediante experimentos, cuales son las funciones wavelets más apropiadas para el filtrado de diversos patrones de señales. Luego, se propone una mejora a los métodos actuales de filtración con wavelets por añadir un paso previo de estimación del nivel de descomposición que afecta a la aplicación de las wavelets. Lo anterior ayuda a una mayor autonomía en la aplicación en línea de estos métodos, a la vez que aseguran precisión en la estimación de los filtrados resultantes. Adicionalmente, se propone una estrategia que combina varias wavelets para intentar dar respuesta en aplicaciones en-línea a la pregunta de cual wavelets utilizar.

El problema de la calidad de los datos también se estudia a través del enfoque de Reconciliación de Datos (RD). Se intenta contribuir al desarrollo de estrategias para casos dinámicos y lineales, uno de los retos actuales de la RD. La propuesta desarrollada combina un paso inicial de extracción de tendencias mediante filtrado basado en wavelets con la posterior reconciliación de las tendencias con una técnica RD basada en la representación polinómica del modelo del proceso. Las propuestas se muestran mejor que las estrategias actuales, en términos de precisión de los estimados obtenidos. Adicionalmente, se propone una primera extensión del método para procesos altamente no lineales, obteniéndose resultados satisfactorios.

En el área de supervisión se presenta un análisis comparativo de diversas estrategias de monitorización basada en Análisis de Componentes Principales (ACP) y para el caso de procesos afectados frecuentemente por perturbaciones de lenta aparición. Se propone una variante basada en filtrado wavelets con ACP que logra obtener respuestas competitivas con los métodos actuales para la detección de este tipo de perturbaciones pero que, adicionalmente, reduce drásticamente la generación de alarmas falsas.

En otro bloque de trabajos de supervisión se presenta el análisis de estrategias que combina ACP con técnicas de Clustering, para la supervisión de procesos multioperacionales. En una primera parte se presenta una comparación de diversas combinaciones ACP-clustering. Esta comparación permite establecer cual de ellas brinda un mejor manejo de aspectos como la identificación de clusters de formas diversas o el tratamiento de outliers. En la comparación se añaden leves extensiones a algunas de las técnicas existentes que conducen a un mejor manejo de todos los aspectos mencionados anteriormente. Adicionalmente, se establecen alternativas de cómo usar las técnicas para casos en que se tiene poco o ningún conocimiento previo de los grupos de operación.

A continuación, se propone una integración TEM-clustering con modelos ACP multigrupos para la supervisión de procesos multi-operacionales. A diferencia de las

estrategias existentes en la literatura, se introduce el tratamiento de las transiciones durante la etapa de diseño del sistema de monitorización, como soporte al operador durante un cambio de operaciones y/o para ayudarle a reducir rápidamente el espejo de posibles causas de anormalidades tras la ocurrencia de un fallo. Finalmente, las estrategias anteriores se adaptan al análisis de procesos afectados por decaimiento en la operación. La estrategia resultante se muestra potencialmente útil tanto para profundizar en el conocimiento del proceso como para asistir en su supervisión, planificación y mantenimiento.

Summary

This thesis presents a set of new methodologies that tries to exploit the information embedded in process historical data and effectively support process analysis and supervision tasks.

The data rectification problem was considered in first place. The adequacy of some type of wavelet for univariate filtering of different signal patterns was studied. Then, a strategy to determine the best decomposition level was proposed and consequently, an initial step to improve current wavelet filtering approaches was found. The obtained results expand the applicability and reliability of existing filtering schemes with wavelets for on-line applications without losing of accuracy on signal estimation. Additionally, an alternative strategy was proposed to solve the problem of which wavelet to choose. This last strategy consist on a weighted combination of different wavelets functions with only one output.

The data rectification problem was also studied through a Data Reconciliation (DR) approach. The focus was set on DR developments for Dynamics and Linear Systems. The proposed strategy consists on first applying a trend extraction step, to identify measured process variables trends and then reconciling these trends to make them consistent with the dynamic process model studied. For the trend extraction step, filtering using wavelets was adopted. To reconcile the estimated variables trends, an extended polynomial approach was used. The comparison with existing RD approaches shows promising results in terms of accuracy and computing efficiency. Further extensions that contemplate nonlinear cases were also introduced, showing also satisfactory results.

Process Supervision problems were considered in second place. Primarily, Principal Components Analysis (PCA) based monitoring strategies for treatment of processes frequently affected by slowly appearing disturbances or small relative shifts were compared. This comparison included some new proposals combining wavelets filtering approaches and PCA. One of the proposed approaches was capable of handling the detection of small disturbances as good as other existing approaches, but dramatically reducing the problem of false alarms generation.

Multioperational process supervision strategies were also considered and studied. First, a comparison of different strategies from literature was considered. The aim was to determine the strategy that produced better results in front of issues like identification of clusters with different forms or its performance facing outliers. The considered strategies are based on the combination of PCA with clustering techniques (PCA-clustering). Not only existing approaches were studied but also some extensions of them were also considered. Finally it was shown how new modified strategies lead to improve handling of all the considered issues. In addition, cluster number estimation problem was studied and some successfully strategies were proposed to perform it.

Finally, the integration of the above PCA-clustering strategies with multigroup PCA for supervising of multioperational process was proposed and evaluated. The aim was to allow good process supervision capabilities for handling operating changes situations and to facilitate fault diagnosis tasks together with additional capabilities like data transitions treatment. Additionally, an extension of the above-integrated strategy for

analysis and supervision of process with decaying performance was evaluated. The resulting strategy was shown as potentially useful to extract useful knowledge from data and to support supervising, planning and maintenance process tasks.

Agradecimientos

Por que me podría olvidar de muchos,
por que nunca somos capaces de reconocer todo el bién que se nos ha hecho,
por que mis palabras a veces son torpes para expresar las cosas correctamente y
por que era lo que sentía que debía decir en este momento,
dirijo mi agradecimiento:

"A …. todos"

Índice

CAPÍTULO 1. INTRODUCCIÓN

1.1 Un punto de partida: Minas de datos operacionales en la Industria Química y de Procesos

La Industria Química y de Procesos (IQP) representa un sector importante tanto en la economía de los países desarrollados y de la economía global como en la economía y desarrollo de países emergentes (Grossman, 2003). Debido a esto, se ha hecho un considerable esfuerzo para mejorar el diseño y la operación de las plantas tal que los procesos operen en forma segura y eficiente, se asegure la mejor calidad de los productos, se mantenga la viabilidad económica y la competitividad del negocio, y se reduzca el impacto ambiental de sus actividades.

1.1.1 El enfoque jerárquico para la operación de plantas en la IQP

Dada la complejidad creciente de las plantas en la IQP, se ha adoptado extensivamente una visión jerárquica de las mismas que divide a la planta en un conjunto de niveles de operación (SP95, 2000; Edgar *et al*., 2001; Sequeira, 2003). Con esto se intenta que la toma de decisiones asociada al manejo global de la planta (un problema grande y complejo) se resuelva como la suma de soluciones a la toma de decisiones a distintos niveles operativos (problemas más pequeños y menos complejos).

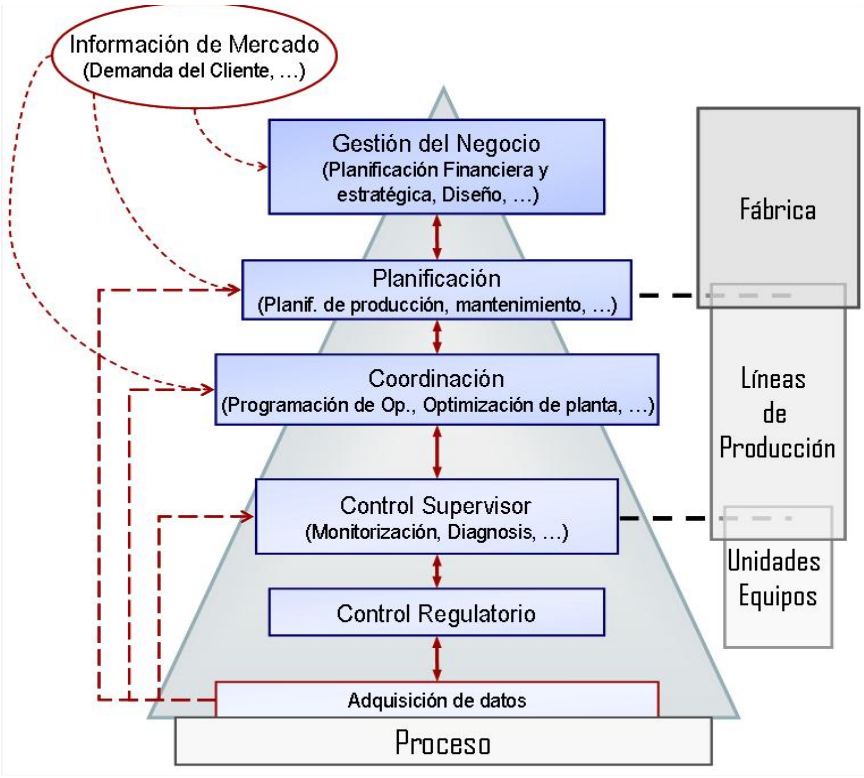

Figura 1.1. Estructura jerárquica - IQP.

A cada nivel, la toma de decisiones se soporta por una variedad de tareas operacionales (control regulatorio, rectificación de datos, monitorización y diagnostico de fallos, optimización en tiempo real, programación de operaciones, manejo de inventarios, mantenimiento, planificación de la producción, previsión de demanda y mercadeo, gestión financiera,…, etc.). El éxito en la toma de decisiones y la aplicación eficiente de cada una de estas tareas descansa en 3 factores claves (Edgar *et al*., 2001):

- Métodos apropiados para su aplicación.
- La posible integración entre más de una tarea
- Una apropiada información de soporte.

1.1.2 Minas de Datos de Proceso = Adecuada Información de Soporte

En años recientes y como consecuencia de la adopción y continuo desarrollo en tecnología de sensores, equipos de recogida de datos y sistemas informáticos de almacenaje, la cantidad de datos de operación recogidos y disponibles a cada nivel de operación se ha incrementando enormemente. Diversos autores han reconocido el potencial de conocimiento inmerso en estos datos para asistir en la operación y mejoras aplicables al proceso (Harmon y Schlosser, 1999; Wang, 2001; Stockill, 2002), pero también han resaltado la poca capacidad actual para realizar con éxito la explotación de la información de dichos datos (Wang, 2001; Stockill, 2002). Esto ha provocado la necesidad de desarrollar, adoptar y/o extender técnicas de análisis de datos que ayuden a manejar estas minas de datos para extraer información útil y con ella soportar la aplicación de diversas tareas operativas. El trabajo de esta tesis se centrará en estos problemas de obtención, manipulación y utilización de la información a partir de datos de operación.

1.2 Una solución reciente

El problema de la continua generación de inmensas cantidades de datos asociados a una actividad concreta y el reto de cómo extraer conocimiento útil de los mismos no es un problema único de la IQP. Diversas organizaciones en la banca, el comercio y la industria así como grupos de investigadores de diversos campos (desde la astronomía a la biología o la informática) se han estado enfrentando desde hace varios años al mismo reto (Fayyad *et al*., 1996; Goebel y Gruenwald, 1999; Apte *et al*., 2002), lo que les ha llevado a explorar diversas iniciativas las cuales tienden a agruparse bajo los nombres de los procesos *KDD* (*Knowledge Discovery in Databases*) y la minería de datos (Han y Kamber, 2001; Zabala, 2003).

1.2.1 ¿Que es KDD?

KDD es el acrónimo del ingles *Knowledge Discovery in Databases*, un campo de investigación y aplicación interdisciplinario que se ha hecho relevante en los últimos 12-15 años. Intenta proponer soluciones al problema de cómo extraer información de grandes cantidades de datos.

Diversos autores (Cios *et al*., 1998; Han y Kamber, 2001; Hand *et al*., 2001; Wang, 2001; Apte *et al*., 2002) han adoptado como básica la definición de *KDD* propuesta por Shapiro: "*KDD* es el proceso no trivial de identificar patrones noveles, potencialmente útiles y entendibles de los datos" (Fayyad *et al*., 1996). Por patrón se entiende algún

tipo de estructura de información que represente de forma clara y resumida uno o varios aspectos del sistema en estudio.

1.2.2 Arquitectura del proceso KDD

En la literatura *KDD* se reconoce de forma prácticamente unánime como un proceso en varias etapas (Cios *et al*., 1998; Han y Kamber, 2001; Hand *et al*., 2001; Wang, 2001; Apte *et al*., 2002). A continuación, se describe brevemente este proceso:

- Definición del objetivo del análisis.
- Selección de datos. Esta selección se hace de acuerdo a los objetivos propuestos y muchas veces se asocia a aspectos informáticos relacionados a como acceder y almacenar los datos.
- Preprocesamiento de los datos. Este paso se asocia básicamente a asegurar la calidad de los datos en el sentido de eliminar ruidos aleatorios, *outliers* (datos atípicos o errores gruesos) manejo de datos ausentes o perdidos.
- Transformación de los datos. Este paso se refiere a encontrar algún tipo de características que ayude a mejorar la eficiencia y facilidad de identificar patrones. Lo típico en esta etapa son los métodos de proyección y reducción de dimensionalidad de los datos.
- Minería de Datos (MD). Es el paso central del proceso *KDD*. Frecuentemente se le identifica por su nombre en ingles: *Data Mining*. La meta en esta etapa es identificar patrones bien definidos y significativos de cara al objetivo del análisis. Para ello se hace uso de diferentes algoritmos y técnicas de clasificación, *clustering* (agrupamiento), regresión, asociación mediante reglas, etc., muchas de ellas desarrolladas a partir del éxito de la filosofía *KDD*.
- Interpretación y validación. Se enfoca a la evaluación e interpretación de los resultados del paso MD.

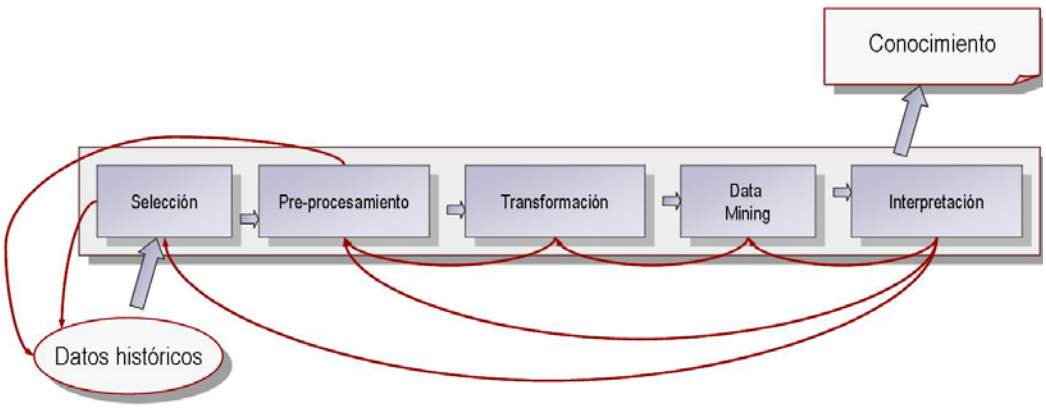

Figura 1.2. Arquitectura genérica del Proceso *KDD*.

El creciente entusiasmo generado alrededor de *KDD* y MD ha provocado un esfuerzo significativo de desarrollo de metodologías y herramientas académicas y comerciales que proporcionen los instrumentos necesarios para obtener conocimiento a partir de datos (Fayyad *et al*., 1996; Cios *et al*., 1998; Han y Kamber, 2001; Hand *et al*., 2001; Wang, 2001; White, 2003). En la literatura citada también se pueden encontrar, en cantidades mucho menores, aplicaciones de *KDD* y MD a situaciones reales y en problemas específicos. Sin embargo, son muchos los analistas tanto académicos como industriales que ven en estos trabajos mucho refinamiento teórico de las herramientas pero poca efectividad en soportar la resolución de los problemas que se plantean,

echando en falta una mayor sinergia entre dicho refinamiento y la utilidad para soportar los problemas a los que se aplican (Fayyad *et al*., 1996; Goebel y Gruenwald, 1999; Apte *et al*., 2002; Wu *et al*., 2003).

1.3 Necesidades y retos específicos en la IQP

La adopción y aplicación de estrategias *KDD* dentro de la IQP, se ha visto como una opción a evaluar para poder enfrentarse al análisis de los datos operacionales (Harmon y Schlosser, 1999; Yamashita, 2000; Wang, 2001; Stockill, 2002). Ahora, como bien se propone en el paso inicial del proceso *KDD* (ver sección 1.2.2), el análisis de datos y todos los métodos y estrategias que se desarrollen para tal fin deben responder a necesidades específicas o problemas concretos para los que la información en los datos operacionales disponibles sea un elemento clave.

Así, la adopción de propuestas como el proceso *KDD* y técnicas MD para la IQP debe atender a problemas, necesidades y retos específicos a las que ésta se enfrente, lo cual puede llegar a implicar la adopción y aplicación del proceso completo que se describe en la sección 1.2.2, o solo algunos (o incluso uno) de los pasos del proceso. A continuación se describen algunos de estos problemas asociados a la gestión y operación de plantas químicas.

1.3.1 La calidad de los datos

En general, los datos de mediciones de proceso son de baja calidad en el sentido de que vienen afectados por errores aleatorios, errores gruesos (*outliers*) de diversos tipos (por ejemplo, valores picos o saltos en el valor de la media de una variable) y valores ausentes o perdidos.

Uno de los primeros pasos de cualquier análisis, por tanto, debe tender a la llamada "rectificación de los datos", que consiste en estimar los valores verdaderos (libres de errores) de las variables del proceso $\mathbf{y}^*$, a partir de las mediciones del proceso $\mathbf{y}$ tales que:

$$\mathbf{y} = \mathbf{y}^* + \boldsymbol{\varepsilon} \tag{1.1}$$

Donde $\boldsymbol{\varepsilon}$ representa los errores que se añaden a la medición. Es una tarea clave, ya que si se logran minimizar los errores, los datos resultantes conducirán a mejores resultados en las aplicaciones que utilicen estos datos (Edgar *et al*., 2001; Abu-el-zeet *et al*., 2002) (Amand *et al*., 2001).

Aun cuando la rectificación representa un paso intermedio del proceso *KDD*, como paso de preprocesamiento de los datos (ver sección 1.2.2), en la industria química es un requisito de diversas tareas (control, monitorización, optimización de procesos, mantenimiento, etc.) por lo que en sí mismo representa un problema de gestión de información a resolver (Narasimhan y Jordache, 2000). Los desarrollos en esta área han ido surgiendo a lo largo de las 4 últimas décadas y pueden dividirse en 2 grandes bloques:

- Métodos de rectificación basados en modelos.
- Métodos de rectificación basados en filtros univariable.

1.3.1.1 Métodos de rectificación basados en modelos

Cuando se cuenta con un modelo fiable del proceso la rectificación se orienta a estimar los valores reales de las variables, tal que los estimados resultantes sean consistentes con el modelo del proceso. Los desarrollos en esta área se agrupan bajo el nombre de Reconciliación de Datos (RD). La RD en su forma más elemental propone un problema de optimización que busca minimizar el error entre las mediciones y los valores reales de las variables, teniendo como restricción el modelo del proceso.

$$\min_{\hat{\mathbf{y}}*} \left[\hat{\mathbf{y}}* - \mathbf{y}\right]^T \mathbf{Q}^{-1} \left[\hat{\mathbf{y}}* - \mathbf{y}\right] \tag{1.2}$$

Sujeto a:

$$f() = 0 \tag{1.3}$$

Donde f representa las restricciones asociadas al modelo del proceso y **Q** es la matriz de varianza-covarianza de las variables del proceso. Los valores en la diagonal de **Q** representan las varianzas de cada variable. Luego, al dividir por **Q** se intenta compensar la diferencia de escalas entre las variables tal que todas tengan un mismo efecto sobre la función objetivo (Narasimhan y Jordache, 2000). La RD tradicionalmente se ha estudiado y evaluado utilizando un modelo estacionario del proceso, lo que se conoce como RD en Estado Estacionario (RDEE), existiendo una teoría bien desarrollada y documentada que incluye aplicaciones en la industria (Narasimhan y Jordache, 2000; Romagnoli y Sánchez, 2000; Bagajewicz, 2003). No obstante, muchos investigadores del mundo académico y analistas en la industria (Liebman *et al*., 1992; Albuquerque y Biegler, 1996; Mingfang *et al*., 2000; Narasimhan y Jordache, 2000; Romagnoli y Sánchez, 2000; Bagajewicz, 2003) han insistido en el hecho de que al ser los procesos químicos intrínsecamente dinámicos, su aplicación a los mismos requiere el uso de estrategias RD basadas en modelos en estado Dinámico (RDD).

Los desarrollos RDD han ido apareciendo con cierta frecuencia en los últimos 12-14 años, con propuestas que incluyen:

- Resolución del problema de optimización mediante enfoques de programación dinámica (Liebman *et al*., 1992; Albuquerque y Biegler, 1996; Barbosa *et al*., 2000; Mingfang *et al*., 2000).
- Soluciones analíticas y numéricas basadas en representaciones polinomiales del modelo del proceso (Bagajewicz y Jiang, 1997; Bagajewicz y Jiang, 2000).
- Soluciones analíticas discretas basadas en la formulación del filtro de Kalman (Rollins y Devanathan, 1993; Narasimhan y Jordache, 2000; Abu-el-zeet *et al*., 2002).

Todos los desarrollos existentes padecen de ciertos problemas comunes como falta de métodos adecuados para la estimación de la varianza de las variables o de algoritmos que permitan una resolución numérica apropiada en tiempo de ejecución para la aplicación de las mismas en línea y en tiempo real, no habiéndose llegado hasta ahora a tener un conjunto de estrategias lo suficientemente fiables como para alcanzar su adopción progresiva en la industria (Mingfang *et al*., 2000; Bagajewicz, 2003).

1.3.1.2 Métodos de rectificación basados en filtros univariable

Una limitante principal en el uso industrial de la RD ha sido y sigue siendo la falta de un modelo fiable o incluso la ausencia de un modelo del proceso en muchos casos reales. Frente a esto, la opción ha sido el uso de filtros univariable basados en estadísticas (por ejemplo, *EWMA* o *Exponential Weighted Moving Average*, promedio móvil, filtros por la mediana,…, etc.). Estos son los filtros más populares en la industria (Bakshi *et al*., 1997). Presentan la desventaja respecto a la RD de que los estimados que se obtienen no necesariamente son consistentes con las relaciones fundamentales que rigen al proceso. Además (a diferencia de la RD) con el tratamiento por separado de cada variable se anula la disponibilidad de información de redundancia entre las variables y, en consecuencia, es imposible estimar los valores de variables no medidas en el proceso. Por el contrario, son mucho más fáciles de implementar y evaluar lo que ha facilitado su disponibilidad en muchos sistemas informáticos de planta (por ejemplo en los sistemas de control distribuidos), además de posibilitar su uso en línea. Junto a esto, se ha visto que con estos métodos muy frecuentemente se obtenien buenos estimados. El número de métodos disponibles es muy amplio (Narasimhan y Jordache, 2000; Köhler y Lorenz, 2004).

En ingeniería química y desde hace 10-15 años se ha discutido y demostrado la marcada naturaleza multiescala de los datos[a] y la necesidad de usar estrategias de análisis que sepan explotar esta característica en los datos de proceso (Stephanopoulos y Han, 1996; Davis *et al*., 2000; Yoon y MacGregor, 2004). Por otro lado, los métodos tradicionales de filtración univariable procesan la señal bajo un enfoque a una sola escala (Nounou y Bakshi, 1999). Por este motivo, en años recientes se ha propuesto el uso de sistemas multiescala y, entre estos (ver sección 2.1.1.2.3), las funciones *wavelets*[b] para el filtrado y extracción de tendencias de las variables de proceso (Bakshi y Stephanopoulos, 1994b; Bakhtazad *et al*., 1999; Nounou y Bakshi, 1999; Jiang *et al*., 2003). Las propuestas han incluido comparaciones de la eficiencia del filtrado con filtros tradicionales en los que se ha mostrado la mejora de calidad en los datos rectificados al usar estrategias basadas en *wavelets* (Bakhtazad *et al*., 1999; Köhler y Lorenz, 2004). Pese a este éxito y el número de propuestas en la literatura, la adopción de las *wavelets* para filtrado involucra ciertos problemas prácticos que se deberían intentar resolver como por ejemplo la selección de la *wavelets* más apropiada o la determinación del nivel de descomposición que se aplica durante un análisis con *wavelets* (Nounou y Bakshi, 1999).

1.3.1.3 Otros métodos

La clasificación anterior no es absoluta. Existen otros métodos que se han ido proponiendo en la literatura para rectificación pero los mismos no han sido hasta ahora relevantes. Es el caso de los métodos basados en modelos empíricos en los cuales se pretende procesar todos los datos de las variables disponibles mediante una técnica como el Análisis de Componentes Principales (ACP) la cual funciona a la vez como filtro y como modelo de la señal. No obstante, en la misma literatura se ha visto que para un mejor aprovechamiento del modelo ACP en aplicaciones posteriores al recitifcado, como la monitorización o diagnóstico de procesos, es mejor aplicar un filtrado previo al tratamiento de los datos con ACP ((Musulin *et al*., 2006).

[a] Ver discusión sobre noción de escala en el anexo A.
[b] Ver definición de wavelets en sección 2.2.2.

En la literatura también se pueden encontrar versiones multivariables de varios filtros univariables como el *EWMA*. En estos casos la extensión multivariable consiste en una versión matricial de los filtros con lo que se logra un tratamiento simultáneo de cada variable filtrada. No obstante, los filtrados se producen de manera independiente lo que equivale a decir que en las extensiones multivariables lo que se hace es colocar varios filtros en paralelo, con iguales parámetros, y por cada uno se pasa una de las variables a filtrar. Así, en términos de precisión de los estimados, estas extensiones multivariables conducen a los mismos resultados del caso univariable del que se han derivado.

1.3.1.4 Tratamiento de Outliers

En cualquiera de los 2 enfoques anteriores, la rectificación se orienta a la eliminación de errores aleatorios. El manejo de *outliers* o errores gruesos se plantea como una tarea separada y complementaria que puede aplicarse antes, en paralelo o después de eliminar los errores aleatorios (Pearson, 2002; Chiang *et al*., 2003). En estos casos los *outliers* siempre son tomados como datos anormales y se intenta eliminarlos.

No obstante, los *outliers* pueden estar asociados a información importante del proceso (un ciclo de ensuciamiento de un equipo, caída de la demanda de un producto como efecto de la entrada de un competidor en el mercado, etc.) por lo que los tratamientos que no consideran el conocimiento que involucran pueden conducir a una perdida de oportunidad de conocer mejor y posiblemente mejorar la operación de un proceso. Así, en esta tesis la detección y tratamiento de *outliers* no se discute dentro del marco del rectificado de los datos sino que se aborda indirectamente y dentro de la perspectiva de los problemas que se discuten en las secciones 1.3.2.

1.3.2 La supervisión de procesos

El término se utiliza con frecuencia para designar la tarea de observar continuamente el proceso o las variables del proceso en busca de detectar anormalidades que puedan representar un problema operativo o de calidad. Con los años esta tarea se ha ido refinando, dando lugar a nuevas subtareas como la detección de fallos, el diagnóstico de fallos y el análisis de procesos y donde todas juntas trabajan por alcanzar el objetivo de supervisar el proceso.
Dependiendo del horizonte de tiempo en el que se trabaja, la supervisión se puede aplicar a 2 niveles:

- A corto plazo: A este nivel, las variables de proceso se observan continuamente con la meta de detectar cualquier desviación respecto del estado normal del proceso y de reaccionar lo más rápidamente para asegurar la operación normal de la planta. El término **monitorización** se utiliza más frecuentemente para referirse a este nivel con un énfasis en la detección e identificación de fallos.
- A largo plazo: A este nivel se analiza el comportamiento del proceso a largo plazo y a través de los datos históricos con la meta de identificar causas de bajo rendimiento y oportunidades de mejora. Los términos **Análisis del proceso** o **Mejora del Proceso** se utilizan con cierta frecuencia en la literatura para designar este tipo de supervisión (Wang, 2001; MacGregor, 2004).

1.3.2.1 Clasificación de métodos existentes

Los desarrollos en esta área, durante los últimos 15 años, son abundantes y el número de aplicaciones en la industria también ha ido creciendo (Ferrer, 2002; Venkatasubramanian *et al*., 2003a; Venkatasubramanian *et al*., 2003b;

Venkatasubramanian *et al*., 2003c; MacGregor, 2004). Aún cuando no existe una única clasificación, los métodos existentes se tienden a agrupar como sigue (Yoon y MacGregor, 2000; Chen y Liao, 2002; Venkatasubramanian *et al*., 2003c): (1) Estrategias basadas en modelos cuantitativos, (2) Estrategias basadas en modelos cualitativos y (3) Estrategias basadas en los datos históricos del proceso.

1.3.2.1.1 Estrategias basadas en modelos cuantitativos

En estas propuestas, el punto de partida es siempre el mismo: obtener un modelo matemático explicito del proceso. Luego, durante la operación del proceso analizado se generan predicciones del modelo y se comparan con las mediciones actuales. Finalmente, se evalúa el estado del proceso y/o la ocurrencia de un fallo específico a través de los vectores residuales que se obtienen de la comparación anterior. Los modelos que se utilizan pueden ser de tipo causal como estimadores de estado, ecuaciones de paridad, etc., que en su mayoría resultan ser modelos "caja negra" lineales discretos (Gertler, 1998), o modelos basados en principios fundamentales (balances, cinéticas,…, etc.). En años recientes, se ha intentado dar mayor prioridad al uso de modelos de principios fundamentales debido a que éstos son más transparentes en el conocimiento que aportan y debido a la disponibilidad de más simuladores de proceso junto a la posibilidad de obtener soluciones rápidas (si la hay) de sistemas matemáticos complejos por medio de la mayor capacidad de los ordenadores actuales (Venkatasubramanian *et al*., 2003c).

Varios autores argumentan que si se dispone de un modelo bastante fiable del proceso, este tipo de estrategias son las más apropiadas para monitorizar (Gertler, 1998; Chen y Liao, 2002; Srinivasan *et al*., 2005). No obstante, la gran dificultad para obtener un modelo fiable e incluso la imposibilidad de resolver tales modelos en muchas situaciones reales ha sido el obstáculo principal para la adopción de este tipo de estrategias en procesos químicos, lo que también ha repercutido en una disminución del esfuerzo de investigación de este tipo de estrategias (Venkatasubramanian *et al*., 2003c).

1.3.2.1.2 Estrategias basadas en modelos cualitativos

En este grupo se enmarcan una serie de metodologías que mantienen un principio de trabajo similar al de las metodologías discutidas en la sección anterior, es decir, comparan una predicción de un modelo frente a la situación actual del proceso y a través de esta comparación se detectan y diagnostican posibles anormalidades en el proceso. Sin embargo, existe una gran diferencia en cuanto al tipo de modelos y la forma en que estos se utilizan. Para la construcción de los modelos, el conocimiento que se tiene del proceso se estructura en forma de relaciones causales con ayuda de técnicas tipo grafos o estructuras de árboles (Vedam y Venkatasubramanian, 1999; Wang, 2001), o en forma de abstracciones jerárquicas que permitan hacer inducciones sobre el comportamiento del sistema (Quantrille y lin, 1991; Stephanopoulos y Han, 1996; Venkatasubramanian *et al*., 2003a). El modelo resultante generalmente es un sistema experto que proyecta no una sino varias predicciones instantáneas de operaciones normales y anormales ante las que se compara el estado actual. Esto teóricamente permite obtener un razonamiento claro sobre el comportamiento del proceso y una fácil explicación de problemas para las tareas de diagnosis.

Por ahora, el uso de estas estrategias se ha visto fuertemente limitado por el hecho de que los desarrollos de los modelos que las soportan y/o de los sistemas expertos en que quedan finalmente expresados son muy costosos en tiempos de desarrollo. Además, las

complejas relaciones de muchos procesos químicos pueden hacerse imposibles de representar mediante jerarquías, árboles o algún otro tipo de representación cualitativa no solo por lo complejo de las relaciones y del número de ellas que puedan surgir, sino también por la falta de información en muchas situaciones reales.

1.3.2.1.3 Estrategias basadas en datos históricos del proceso

Estas propuestas descansan en la extracción de información de los datos medidos. Han atraído mucho la atención como alterativas para atacar el problema de las minas de datos con propósitos de monitorización. Adicionalmente, se han constituido como la alternativa potencial al problema de obtener modelos rigurosos para monitorizar. Las propuestas en esta área se pueden agrupar en 3 tipos:

Análisis de Tendencias Cualitativas

Se basan en el modelado del comportamiento temporal de un proceso mediante las señales de sus variables a lo largo del tiempo y todas las estrategias responden a un esquema de análisis similar al propuesto originalmente en los trabajos de Stephanopoulos y colaboradores (Cheung y Stephanopoulos, 1992a; Cheung y Stephanopoulos, 1992b; Bakshi y Stephanopoulos, 1994a; Bakshi y Stephanopoulos, 1994b): Se parte de la definición inicial de un conjunto de patrones o primitivas de las señales. Luego, cada señal que se recibe de un proceso se segmenta de acuerdo a la semejanza con las primitivas predefinidas y, finalmente, se relacionan los patrones obtenidos, con las condiciones del proceso mediante algún mecanismo de reglas o asociaciones causales. Algunas de las propuestas más recientes proponen procedimientos más elaborados y en varias etapas que intentan explotar más efectivamente la información en las tendencias constituyéndose (sin ser planteados como tal) en verdaderos procesos de análisis tipo *KDD* (Dash *et al*., 2003; Sun *et al*., 2003; Srinivasan y Qian, 2005).

Dado el atractivo que ofrecen en cuanto a facilidad de explicación de un problema en el proceso para propósitos de diagnostico de fallos, el número de propuestas en la literatura es significativo incluyendo varias aplicaciones industriales (Kivikunnas, 1999; Davis *et al*., 2000; Venkatasubramanian *et al*., 2003b). No obstante, se ven difíciles de usar para procesos con muchas variables, y su implementación es difícil de llevar a cabo, especialmente la parte en la que se relacionan los patrones identificados con las condiciones del proceso.

Estrategias basadas en Redes Neuronales

Las Redes Neuronales Artificiales (RNA) han gozado de gran popularidad para aplicaciones como monitorización y diagnostico de fallos[c] (Chen y Liao, 2002; Venkatasubramanian *et al*., 2003b). El principal atractivo que ofrecen las RNA y estrategias derivadas consiste en la capacidad que tienen para manejar no linealidades y dinámicas diversas en los datos. El esquema de estrategias de monitorización con RNA es en muchos casos como sigue: Se toman datos históricos del proceso y con ellos se entrena una red neuronal. El conjunto de datos de entrenamiento incluye tanto datos de operación normal como datos de operaciones anormales para casos de fallos conocidos. Después del entrenamiento el modelo resultante se utiliza en línea. Con los nuevos

[c] El nº de trabajos ha disminuido en los últimos 3-4 años como se puede apreciar en las últimas ediciones de los congresos ESCAPE, sin acompañarse de reportes de aplicaciones en la industria.

datos del proceso se clasifica el estado del mismo como normal o anormal con lo que se cubre la detección. En algunos casos, la salida de la red se alimenta a otro sistema para asistir en el diagnóstico del fallo que ha ocurrido (Ruiz *et al*., 2000; Musulin *et al*., 2006).

A pesar del interés despertado y de las numerosas propuestas para detección y diagnostico de fallos esta aproximación presenta ciertas dificultades. En primer lugar, el conocimiento expresado en el modelo que se obtiene con la red es poco transparente para ser interpretado por un operador no experto, a menos que se integren sus salidas con un sistema experto. Aún más crítico es el hecho de que este conocimiento tiende a ser muy preciso con los datos usados para entrenar la red, pero poco generalizable fuera de dicho conjunto de datos. Finalmente, durante la etapa de diseño de la RNA, la selección de la estructura de la red o el ajuste de sus parámetros tienden a consumir muchos recursos en tiempo de desarrollo y trabajo del analista (Chen y Liao, 2002; Venkatasubramanian *et al*., 2003b).

Estrategias basadas en Técnicas Estadísticas Multivariables

Las Técnicas Estadísticas Multivariables (TEM) se han propuesto y estudiado de forma intensiva durante los últimos 16 años (Wise y Gallagher, 1996; Kano *et al*., 2002; Wold *et al*., 2002; Kourti, 2003; Qin, 2003; MacGregor, 2004). El interés que han despertado radica en:

- Su capacidad para manejar conjuntos de datos de dimensiones altas y/o con variables de proceso altamente correlacionadas, resumiendo toda la información de cientos de variables en unas pocas cantidades.
- Su capacidad de inferencia, útil para la identificación y, en menor grado, para el diagnostico de fallos.

Las propuestas originales se basaron en el uso del Análisis de Componentes Principales (ACP) y Mínimos Cuadrados Parciales, MCP (en ingles *Partial Least Squares*) para monitorización de procesos continuos y lineales (Kresta *et al*., 1991; Piovoso *et al*., 1992). Posteriormente, se intentó extender y mejorar la propuesta inicial para un adecuado manejo de la dinámica (Ku *et al*., 1995; Li *et al*., 2000), para el manejo de no linealidades (Kramer, 1994), para su aplicación en procesos discontinuos (Nomikos y MacGregor, 1994; Nomikos y MacGregor, 1995), para manejar datos multiescala (Bakshi, 1998) y más.

Adicionalmente se han reportado numerosas aplicaciones exitosas en la industria siendo hoy por hoy los métodos más prometedores como soluciones de monitorización para el futuro (Ferrer, 2002; MacGregor, 2004), aun cuando se sigue trabajando en mejorar sus capacidades de diagnosis.

1.3.2.2 Algunos retos de la monitorización

Las estrategias de monitorización basadas en TEM constituyen hoy por hoy las propuestas más prometedoras en el campo de la monitorización. No obstante, y a pesar del número de propuestas, los retos en el área de monitorización siguen siendo muchos como se puede ver en las discusiones de Davis et al., (2000), Qin (2003) y Venkatasubramanian et al., (2003b). Dada la extensión de los retos planteados a continuación solo se discuten algunos de ellos con respecto al uso de las técnicas estadísticas multivariables.

1.3.2.2.1 Manejo de anormalidades de larga aparición o larga duración

A pesar de que en las primeras propuestas y aplicaciones ya se puso de relieve el potencial de monitorización con ACP y técnicas similares, se ha reconocido que la monitorización mediante el ACP clásico adolece de ciertas incapacidades para manejar correctamente algunas características del proceso o de los fallos que se puedan presentar. Una de estas incapacidades se refiere al hecho de que el ACP no es capaz de detectar rápida ni exactamente la aparición de ciertos fallos de proceso con pequeñas perturbaciones (perturbaciones largas) como por ejemplo un suave decaimiento en las condiciones de operación (Wold, 1994; Wachs y Lewin, 1999; Chen *et al*., 2001). Esto se debe a que en estos casos existe una dependencia significativa entre el comportamiento del punto de trabajo actual del proceso y los puntos de trabajo recientes, llamada memoria del proceso y, no obstante, el ACP asume independencia de los puntos de trabajo (Wold, 1994; Chen *et al*., 2001).

Como una alternativa para mejorar la respuesta del ACP frente a este tipo de anormalidades, se ha propuesto el modelar los efectos de memoria mediante métodos que de algún modo filtren las señales de las variables individuales integrando éstos con el ACP. Así, se han propuesto el *EWMA*-ACP (Wold, 1994) y la combinación de una estadística llamada *s-summed* (ver sección 4.1.3) con el ACP (Wachs y Lewin, 1999). Ambos métodos, muestran una mayor resolución y un menor tiempo de respuesta en la detección de las perturbaciones discutidas cuando se comparan con el ACP tradicional. No obstante, en ambos casos se observa la ausencia de definición de límites de control apropiados para usar con estadísticas derivadas de estos métodos. Adicionalmente, no se indica nada sobre cómo usar la información de las variables que aporta el ACP (por ejemplo para crear y usar un gráfico de control) con lo que se 'pierden' ciertas posibilidades de monitorización que se tienen con el ACP tradicional. Alternativamente, se han propuesto versiones de los métodos anteriores llamadas *MEWMA*-ACP y *MSSUMED*-ACP (Chen *et al*., 2001), los cuales se muestran comparativamente superiores a los anteriores, particularmente el *MEWMA*-ACP.

Por otro lado, se han propuesto estrategias en las que las variables originales se filtran mediante un Filtro Híbrido basado en la Mediana (FHM) para luego procesarlos con *wavelets* (Kosanovich y Piovoso, 1997). Sobre los coeficientes aportados por la *wavelets* se aplica el ACP para monitorizar a diferentes escalas. Otra propuesta similar llamada ACP Multiescala desarrolla aún más el uso combinado de *wavelets* con ACP a fin de explotar las capacidades multiescala que aporta el análisis *wavelets* para intentar mejorar el diagnóstico y la detección (Bakshi, 1998). Se definen gráficas para el *SPE* (Squared Prediction Error) y T^2 adecuadas para análisis a diferentes escalas. También, en las diferentes propuestas se incluyen casos de análisis donde se obtiene la detección de "perturbaciones de aparición lenta" con bastante retraso respecto del punto real de aparición, por lo que la discusión final de los resultados se orienta hacia la detección de otro tipo de perturbaciones sin concluir sobre la capacidad real del ACP Multiescala para el manejo de este tipo de perturbaciones. El reto de una adecuada monitorización de tales perturbaciones sigue en espera de soluciones apropiadas o de evaluaciones adicionales de las propuestas actuales.

1.3.2.2.2 Monitorización de procesos multiproducto continuos y discontinuos

Los desarrollos y aplicaciones de monitorización reportadas básicamente se han enfocado a problemas de operación donde se produce un solo producto (un grado, una receta, una sola condición de operación), y con modelos separados para diferentes

productos. Sin embargo, los requerimientos de mercados cambiantes y la demanda diversificada de productos está empujando cada vez más a procesos de fabricación flexible que permitan la producción de múltiples productos o grados de un mismo producto. Esto último ha conducido a la necesidad de sistemas de monitorización multiproducto (Martin *et al*., 2002).
Para dar respuesta a esta necesidad, recientemente se han propuesto varios enfoques *KDD* que combinan ACP (o alguna otra TEM como el MCP) con técnicas de clustering o el *clustering* de forma individual para ayudar a diseñar sistemas de monitorización de procesos multiproducto (Chen y Liu, 1999; Hwang y Han, 1999; Li y Wang, 1999; Chen *et al*., 2002; Srinivasan *et al*., 2004). El número de propuestas es significativo, con un énfasis especial en el uso de ACP con diversas variantes de técnicas *clustering* basadas en lógica difusa (Næs y Mevik, 1999; Teppola y Minkkinen, 1999; Teppola *et al*., 1999; Rosen, 2001; Tona *et al*., 2001; Choi *et al*., 2003; Yoo *et al*., 2003).

Una revisión exhaustiva de las estrategias propuestas provoca interrogantes en cuanto a cual de ellas brinda un mejor manejo combinado de aspectos como la identificación de *clusters* de formas diversas, el tratamiento de *outliers*, el manejo adecuado de la información aportada,…, etc. Aspectos todos que pueden influir dramáticamente en el diseño del sistema de monitorización y en su posterior puesta en practica en una situación en línea. Esto plantea la necesidad de estudios comparativos que den respuesta a estos interrogantes.

Finalmente, la mayoría de las propuestas se han orientado al caso de procesos operando en régimen continuo dejando en incógnita si las mismas propuestas podrían servir para procesos similares pero operando en semicontinuo o discontinuo.

1.4 Objetivos y alcance de la tesis

Después de describir los problemas anteriores relacionados al uso de estrategias basadas en datos para asistir diversas tareas operacionales y atender a distintos problemas de gestión de planta en la IQP, se introducen los objetivos de esta tesis.

El objetivo principal es "Contribuir al desarrollo de estrategias de análisis de datos operacionales que soporten de manera eficiente la toma de decisiones asociadas a la operación y la supervisión de procesos".

Para dar respuesta efectiva a los retos planteados, el objetivo general se desglosa en una serie de objetivos específicos que se listan a continuación:

- Estudiar y proponer mejoras a las técnicas de rectificación de datos para asegurar la calidad de la información recogida del proceso.

- Contribuir al desarrollo de estrategias de reconciliación de datos para sistemas dinámicos.

- Estudiar y proponer mejoras a las estrategias actuales de monitorización para el caso de procesos afectados por perturbaciones de larga duración o aparición lenta.

- Contribuir al desarrollo de estrategias de monitorización para procesos multiproducto o sujetos a cambios frecuentes en las condiciones de operación.

Para ello:

- En el capítulo 2, se estudian y comparan diversas estrategias de filtración univariable incluyendo la propuesta de estrategias noveles que expanden la aplicabilidad de las técnicas actuales.

- En el capítulo 3, se propone y valora una estrategia nueva para la reconciliación de datos en plantas lineales dinámicas.

- En el capítulo 4, se estudian y comparan diversas estrategias para la monitorización de procesos afectados por perturbaciones de larga duración, incluyendo la propuesta de estrategias noveles que mejoran la aplicabilidad de las estrategias actuales.

- En los capítulos 5 y 6, se comparan y se proponen mejoras a las estrategias de monitorización de procesos multiproducto o que están sujetos a cambios frecuentes en las condiciones de operación.

Nomenclatura

f	Restricciones asociadas al modelo del proceso utilizado en RD.
Q	Matriz de varianza-covarianza de las variables del proceso.
y	Vector de mediciones de las variables de proceso.
y*	Vector de valores reales de las variables de proceso.

LETRAS GRIEGAS

ε	errores que se añaden a la medición de una variable de proceso.

ACRÓNIMOS

ACP	Análisis de Componentes Principales.
EWMA	*Exponential Weighted Moving Average*.
FHM	Filtro Híbrido basado en la Mediana.
IQP	La Industria Química y de Proceso.
KDD	*Knowledge Discovery in Databases*.
MCP	Mínimos Cuadrados Parciales o *Partial Least Squares* (*PLS*).
MD	Minería de Datos.
RD	Reconciliación de Datos.
RDD	Reconciliación de Datos basada en modelos en estado dinámico.
RDEE	Reconciliación de Datos basada en modelos en estado estacionario.
RNA	Redes Neuronales Artificiales.
TEM	Técnicas Estadísticas Multivariables.

CAPÍTULO 2. MEJORANDO LA FILTRACIÓN BASADA EN WAVELETS

RESUMEN
La verificación de la calidad de los datos de proceso es una tarea clave en la industria química y de procesos. En efecto, el uso de datos incorrectos (contaminados por ruidos) puede conducir a conclusiones erróneas en las tareas operacionales que los utilicen, mientras que el uso de datos verificados o rectificados minimizará el riesgo de errores en su aplicación. En este capítulo se estudian y proponen estrategias de filtración de datos usando *wavelets*. En una primera parte se presenta un estudio basado en simulaciones para verificar ciertas ventajas y desventajas de dichas estrategias así como para proponer un uso adecuado de las mismas. A continuación, se analiza una estrategia para estimar el nivel de descomposición asociado a la aplicación de las *wavelets* y su integración a la filtración. En otra propuesta, se busca tomar ventaja conjunta de diversas *wavelets* para obtener un mismo estimado, tal que la estrategia resultante sea aplicable a un mayor número de patrones de señales. Las propuestas se comparan entre si y con otras metodologías clásicas o de filtrado multiescala tanto para casos de aplicaciones fuera de línea como en línea. Para ello, se utilizan señales típicas de la literatura. Los resultados muestran que las metodologías propuestas permiten extender la aplicabilidad de las actuales estrategias de filtrado con *wavelets* sobre un rango amplio de señales con comportamiento diverso, garantizando con ello estimados de calidad de las variables del proceso y una mayor autonomía de las estrategias.

2.1 Introducción

Las mediciones de las variables de proceso aportan información clave para un amplio número de tareas en ingeniería. En general, dichas mediciones vienen contaminadas con ruido proveniente de diversas fuentes (error de los sensores, naturaleza del proceso,…, etc.). El uso de estos datos contaminados puede provocar errores en los resultados de la tarea para los que se estén utilizando. Por lo tanto, se hace necesaria la verificación o rectificación de los datos de medición para obtener una información de calidad. A continuación se presenta una revisión y desarrollos de la filtración univariable.

2.1.1 Rectificación basada en filtros univariable

En este caso la rectificación se aplica mediante filtros basados en una función matemática o estadística que procesan solo una señal a la vez. El problema básico que se plantea es como sigue: Sea la señal $\mathbf{y}(k)$ compuesta por una realización de datos sucesivos y afectados por ruido gausiano o aleatorio tal que $\mathbf{y}(k) = (y(1), y(2),...)$. Se acepta ampliamente que $\mathbf{y}(k)$ puede expresarse como sigue (Bakhtazad *et al*., 1999; Craigmile y Percival, 2002):

$$\mathbf{y}(k) = \mathbf{y}^*(k) + \varepsilon(k) \tag{2.1}$$

Donde $\mathbf{y}^*(k)$ es la tendencia verdadera (no contaminada con ruidos) de la señal original y $\varepsilon(k)$ es el error aleatorio que se añade a la medición (ruido). Esta descripción supone la no

presencia de errores gruesos afectando a $\mathbf{y}(k)$. La meta de la filtración es estimar un valor apropiado de $\mathbf{y}^*(k)$ ($\hat{\mathbf{y}}^*(k)$) mediante la eliminación de la mayor parte del ruido.

En la práctica, los filtros univariable son los que más se utilizan (sección 1.3.1.2). Una característica de los filtros univariable tradicionales es que los mismos manejan la señal procesada bajo una representación a una escala sencilla (Bakshi, 1999). En oposición a esto, numerosos autores han señalado que los datos provenientes de Industrias Químicas y de Procesos (IQP) son de naturaleza multiescala, esto es, la señal que se mide es el resultado de efectos combinados a distintas escalas. Un ejemplo típico es el que se muestra en la figura 2.1 (Cheung y Stephanopoulos, 1992b). La señal de proceso en cuestión es la resultante de la superposición, sobre el valor real de la variable, de los efectos combinados de: un fallo del sensor, una perturbación senoidal, un fallo puntual de un equipo, una degradación en la operación de otro equipo y el ruido.

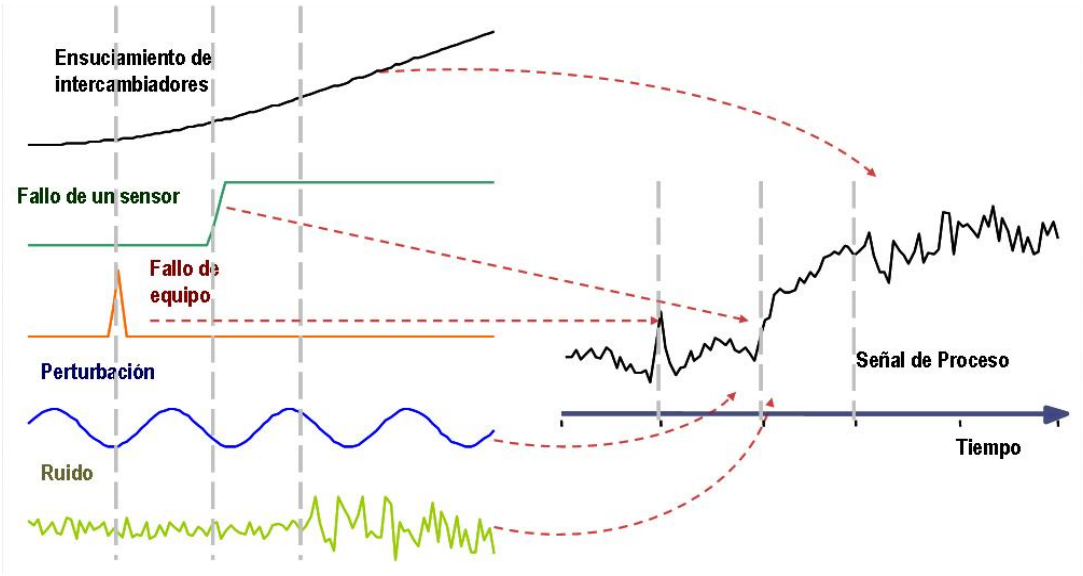

Figura 2.1. Ilustración de una señal de proceso.

Debido a dichas características multiescala se ha explorado y propuesto ampliamente el uso de *wavelets* para diversas tareas de análisis de datos de la IQP, incluyendo filtración univariable (Bakshi, 1999; Davis *et al*., 2000). En las secciones que siguen se discute con mayor extensión sobre las ventajas y desventajas de diversas técnicas de filtración univariable.

2.1.1.1 Métodos de Filtración a una escala

Han sido los filtros más estudiados y usados en investigación y en la IQP, en las últimas 3 décadas.

2.1.1.1.1 Filtros lineales

Estos filtros se caracterizan por rectificar la señal mediante suma ponderada de mediciones pasadas sobre una ventana de longitud finita o infinita tal como sigue (Bakshi, 1999):

$$\hat{\mathbf{y}}^*(k) = \sum_{i=0}^{nc-1} b_i . \mathbf{y}(k-i) \qquad (2.2)$$

Donde *nc* es la longitud de la ventana de datos y los b_i conforman un conjunto de coeficientes ponderados, que definen las características del filtro, cumpliendo siempre la siguiente restricción:

$$\sum_i b_i = 1 \tag{2.3}$$

Dependiendo de la longitud de los filtros lineales se pueden subdividir en:

- Filtros de Respuesta al Impulso Finita: Se les conoce más comúnmente según las siglas de su nombre en ingles *FIR* (*Finite Impulse Response*). Se trata de un tipo de filtros digitales en el que, como su nombre lo indica, si la entrada es una señal impulso, la salida tendrá un número finito de b_i no nulos. Luego, la longitud de la ventana sobre la que se define el filtro es finita. El filtro de media móvil (MM) es el caso más simple y el más popular. En el caso del MM todos los b_i tienen un mismo valor.
- Filtros de Respuesta al Impulso Infinita: También se les conoce según las siglas de su nombre en ingles *IIR* (*Infinite Impulse Response*). En este caso, el nombre indica que si la entrada es una señal impulso, la salida de este tipo de filtros tendrá un número infinito de b_i no nulos. Luego, cualquier punto de datos rectificado se representa como una suma ponderada de infinitas mediciones anteriores.

  El filtro de Suavizado Exponencial Simple (SES), mejor conocido como *EWMA* (siglas de *Exponentially Weighted Moving Average*), es un caso típico de este tipo de filtro en el que los datos rectificados son el resultado de un promedio ponderado exponencialmente, según la siguiente expresión reducida:

  $$\hat{\mathbf{y}}^*(k) = \alpha \cdot y(k) + (1-\alpha) \cdot \hat{\mathbf{y}}^*(k-1) \tag{2.4a}$$

  o equivalentemente:

  $$\hat{\mathbf{y}}^*(k) = (1-\alpha)^k \cdot y(0) + (1-\alpha)^{k-1} \cdot y(1) + (1-\alpha)^{k-2} \cdot y(2) + ... + \alpha \cdot y(k) \tag{2.4b}$$

  Donde α es una constante de suavización que recoge el efecto de los pesos asignados a cada observación. Aunque la ecuación (2.4a) es la que se adopta en la práctica, la ecuación (2.4b) ayuda a visualizar al *EWMA* como un promedio móvil ponderado de todas las mediciones pasadas y actuales, observándose que las contribuciones de las mediciones más antiguas sobre el filtrado actual no se han eliminado completamente, sino que decaen exponencialmente según se obtienen nuevas muestras. Esto permite ver la naturaleza infinita del filtro.

Como se indica en la literatura (Nounou y Bakshi, 1999), estos filtros procesan los datos a una sola escala. En consecuencia, para datos donde las características del ruido estén dispersas sobre más de una escala, la rectificación será incompleta ya que solo se eliminarán las características de los datos contaminados sobre la escala a la que trabaje el filtro seleccionado.

2.1.1.1.2 Filtro por la mediana y filtros híbridos

El filtro basado en la mediana (FM) toma la mediana de las mediciones dentro de una ventana como el valor central filtrado. Este filtro es robusto frente a la presencia de errores gruesos y responde bien frente a cambios súbitos en la señal, siendo por ello muy atractivo en la práctica.

Opcionalmente, la capacidad de FM para retener cambios súbitos se puede mejorar al combinarlo con filtros FIR, resultando en lo que se conoce como filtros híbridos basados en mediana-FIR o FHM (en ingles *FIR-Median Hybrid Filtres* o *FMH*). La forma general de un FHM es como sigue (Kosanovich et al*,* 1997):

$$\hat{\mathbf{y}}^*(k) = median\Big(FIR\big(y(k-1), y(k-2), ..., y(k-nt)\big), y(k), FIR\big(y(k+1), y(k+2), ..., y(k+nt)\big)\Big) \qquad (2.5)$$

Donde *nt* es la mitad de la ventana (*nc*/2). La longitud de la ventana se selecciona en función de la longitud de los errores gruesos que puedan afectar al sistema, lo que conduce a distorsiones en las características de la señal cuando se toman longitudes cortas (Bakshi, 1999). Adicionalmente, estos filtros requieren valores del futuro para computar el rectificado del valor actual, por lo que su aplicación en línea queda restringida a casos donde se permita un retraso entre el tiempo en que se produce la medición y el tiempo en que se computa el filtrado (Bakshi, 1999). Por último, el tratamiento de las señales sigue siendo a una escala.

2.1.1.2 Métodos de Filtración Multiescala

2.1.1.2.1 Las técnicas wavelets

Las *wavelets* representan un conjunto de funciones matemáticas especialmente diseñadas para análisis multiresolución o multiescala (ver anexo A). Las diversas familias de funciones se han definido de manera tal que se dispone de bases apropiadas para el espacio de las aproximaciones con las funciones de escala [$\phi_{l,\upsilon}(k)$, $\upsilon \in \mathbf{Z}$ in $\mathbf{V}_j$] y para el espacio de los detalles con las funciones *wavelets* [$\psi_{l,\upsilon}(k)$ $l=1,...,L$, $\upsilon \in \mathbf{Z}$ in $\mathbf{W}_j$]. En las definiciones anteriores los parámetros l, v y L representan un factor de escala, un factor de traslación y la escala de mayor resolución (llamada también nivel de descomposición) respectivamente. Cuando se analiza una señal $y(k)$ con *wavelets*, la aplicación de las funciones bases utilizadas (tanto las del espacio de aproximación como las del espacio de los detalles) recogen la información original de $y(k)$ tanto en frecuencia como en tiempo y la transportan a la nueva representación que aportan las *wavelets*. Variando el parámetro l se extrae la información en frecuencia a distintas escalas de $y(k)$ y variando v se extrae la localización temporal de la información en frecuencias a distintos valores de l.

Para propósitos de aplicación con datos reales y medidos, los parámetros l y v siempre se han definido como discretizados sobre una escala diádica, con lo que las funciones de aproximación y detalle quedan expresadas como sigue:

$$\phi_{L,\upsilon}(k) = 2^{-L}\phi\left(2^{-L}k - \upsilon\right), \upsilon \in \mathbb{Z} \qquad (2.6)$$

$$\psi_{L,\upsilon}(k) = 2^{-l}\psi\left(2^{-l}k - \upsilon\right), l = 1,...,L; \upsilon \in \mathbb{Z} \qquad (2.7)$$

Utilizando las funciones anteriores, una señal $\mathbf{y}(k)$ se puede descomponer como sigue (representación multiescala):

$$\mathbf{y}(k) = \sum_{\upsilon \in Z} a_{L,\upsilon} \cdot \phi_{L,\upsilon}(k) + \sum_{l=1}^{L} \sum_{\upsilon \in Z} d_{l,\upsilon} \cdot \psi_{l,\upsilon}(k) \qquad (2.8)$$

Donde $a_{L,\upsilon}$ son los coeficientes ajustados para la aproximación y $d_{l,\upsilon}$ son los coeficientes para los detalles. La descomposición anterior permite expresar la señal original mediante un nuevo conjunto de componentes llamados la señal de aproximación, $\mathbf{A}_L(k)$, y los componentes de detalles, $\mathbf{D}_l(k)$, con (l=1,...,L):

$$\mathbf{A}_L(k) = \sum_{v \in Z} a_{L,v} \cdot \phi_{L,v}(k) \tag{2.9}$$

$$\mathbf{D}_l(k) = \sum_{v \in Z} d_{l,v} \cdot \psi_{l,v}(k) \tag{2.10}$$

$\mathbf{A}_L(k)$ y $\mathbf{D}_l(k)$ se construyen a partir de los coeficientes $a_{L,v}$ y $d_{l,v}$ respectivamente representando información a distintas escalas. Luego, la ecuación 2.8 se rescribe como sigue:

$$\mathbf{y}(k) = \mathbf{A}_L(k) + \sum_{l=1}^{L} \mathbf{D}_l(k) \tag{2.11}$$

$\mathbf{A}_L(k)$ retiene principalmente información de baja frecuencia de la señal original a la escala L y se puede utilizar como una aproximación de la señal original sin ruidos $\hat{\mathbf{y}}^*(k)$. Similarmente, $\mathbf{D}_l(k)$ retiene principalmente información de alta frecuencia de la señal original que guarda relación directa al ruido que contienen los datos a una escala dada. Luego, seleccionando un L apropiado y los coeficientes más representativos de las características de la señal, se puede llegar a obtener una buena estimación $\hat{\mathbf{y}}^*(k)$ (Addison, 2002).

2.1.1.2.2 Rectificación utilizando wavelets

Las *wavelets* se han utilizado con éxito en aplicaciones de filtrado de datos en tratamiento de imágenes, telecomunicaciones, …, etc., (Addison, 2002) explotando para ello la capacidad de descomponer una señal que se vio en la sección 2.1.1.2.1.

La idea principal del filtrado con *wavelets* se basa en transformar los datos a la base *wavelets*. En esta nueva base los coeficientes estimados que tienen valores más altos (principalmente los coeficientes de $\mathbf{A}_L$) representan la información útil de la señal mientras que los coeficientes con valores más pequeños (principalmente los coeficientes de los $\mathbf{D}_l$) se asociarán al ruido. Mediante una modificación apropiada de los coeficientes de $\mathbf{D}_l$, se puede lograr eliminar todo el ruido mientras se retienen las características principales de la señal lo que resulta en una muy buena estimación de $\mathbf{y}^*(k)$. Donoho y colaboradores (Donoho, 1992; Donoho y Johnstone, 1994; Donoho y Johnstone, 1995) fueron los primeros que desarrollaron una metodología de filtrado, llamada *waveshrink*, que involucra 3 pasos (ver la figura 2.2):

- Descomponer la señal original usando *wavelets* (Bloque **W**).
- Eliminar los coeficientes *wavelets* bajo un cierto valor umbral o de referencia β (a este paso se le llama en la literatura inglesa *thresholding* o *shrinkage*).
- Reconstruir la señal procesada por tomar el inverso de la *wavelets* usada, sobre los coeficientes que quedan tras el *thresholding* (Bloque $\mathbf{W}^{-1}$).

Figura 2.2. Método *waveshrink* de filtración usando *wavelets*.

Para el paso de *thresholding*, Donoho y Johnstone (Donoho y Johnstone, 1994; Donoho y Johnstone, 1995) plantearon 2 métodos:

> *Hard thresholding*: Mediante este enfoque se anulan (se hacen igual a cero) los coeficientes con valores por debajo de β, cómo sigue:

$$d_j^H = \begin{cases} d_j & if \left|d_j\right| > \beta \\ 0 & if \left|d_j\right| \le \beta \end{cases} \tag{2.12}$$

Soft thresholding: Este enfoque es una extensión del anterior. No solo se anulan los coeficientes con valores por debajo de β, sino que también se hace una corrección a todos los coeficientes que no se eliminan respecto a β, para intentar que la señal resultante quede más suavizada.

$$d_j^S = \begin{cases} sign(d_j)\cdot(\left|d_j - \beta\right|) & if \left|d_j\right| > \beta \\ 0 & if \left|d_j\right| \le \beta \end{cases} \tag{2.13}$$

En principio, *Hard* es mejor para reproducir discontinuidades y picos que son parte de la señal mientras que *Soft* produce señales más suavizadas y que visualmente pueden lucir mejor (Addison, 2002). La selección se propone comúnmente como dependiente del patrón de la señal a analizar.

La determinación del umbral es uno de los elementos claves en la aplicación de la metodología anterior. Donoho (Donoho, 1992) planteó computar el β como sigue (regla "*Universal shrinkage*" o "*visushrink*"):

$$\beta = \sigma_l \sqrt{2\log_2(n)} \tag{2.14}$$

Donde σ_l denota la desviación estándar del ruido sobre la escala l (del detalle a la escala l) y n es el número de observaciones en $\mathbf{y}(k)$. Debido a que algunos (pocos) detalles podrían representar características importantes de la señal, se utiliza una estimación robusta del σ_l como sigue:

$$\sigma_l = \frac{median\left(\left|d_{l,v}\right| / v \in \mathbb{Z}^+\right)}{0.6745} \tag{2.15}$$

Se han propuesto diversas variantes al método descrito anteriormente que incluyen alternativas para estimar β como el Minimax o la regla de Estimación Insesgada del Riesgo de Stein (Bakhtazad *et al*., 1999), o variantes al método global de *thresholding* como los métodos "*wiener thresholding*" (Ghael *et al*., 1997) y "*Coefficients Denoising*" (Doymaz *et al*., 2001). No obstante, *waveshrink* se mantiene como el más utilizado. La metodología *waveshrink* se ha testeado en varios trabajos con éxito (Nounou y Bakshi, 1999; Taswell, 2000; Köhler y Lorenz, 2004) y se cuenta con versiones implementadas dentro de diversos paquetes informáticos comerciales (por ejemplo, Matlab, Mathematica, SAS).

2.1.1.2.3 Otros métodos multiescala

Las *wavelets* se han utilizado con éxito en aplicaciones de filtrado de datos en tratamiento de imágenes, telecomunicaciones, …, etc., (Addison, 2002) explotando para ello la capacidad de descomponer una señal que se vio en la sección 2.1.1.2.1.

El tratamiento de la información multiescala ha sido abordado en décadas previas mediante otras técnicas distintas a *wavelets*. El grupo más importante entre ellas lo constituyen las basadas en el uso de la transformada de *fourier*. Su definición matemática es similar a las *wavelets* en cuanto que transforman la señal a procesar mediante una combinación de funciones bases pero en este caso las funciones bases siempre se construyen a partir de senos y cosenos. Desde la aparición de las *wavelets* a finales de la década de 1980 se ha discutido sobre las ventajas teóricas que aportan respecto a *fourier* como el tratamiento simultaneo de la información en el tiempo y la frecuencia, el manejo adecuado de señales no estacionarias (señales cuyas características en frecuencia varían con el tiempo) o una mayor rapidez de resolución en tiempo computacional para muestras con muchas observaciones (Li *et al*., 2002; Misiti *et al*., 2004). En la literatura también se ha mostrado experimentalmente la ventaja del uso de las *wavelets* sobre *fourier* en aplicaciones de filtrado (Rowe y Abbott, 1995; Bakhtazad *et al*., 1999; Taswell, 2000; Köhler y Lorenz, 2004).

2.1.2 Algunos retos

En general, se ha mostrado que los métodos de filtración con *wavelets* funcionan mejor que otros métodos clásicos (Bakhtazad *et al*., 1999; Nounou y Bakshi, 1999; Köhler y Lorenz, 2004). Pese a esto, continúan planteándose algunos problemas prácticos asociados a la adopción de un enfoque de filtración con *wavelets* y para los que la literatura actual da soluciones subjetivas o simplemente se ignoran.

Uno de estos problemas tiene que ver con la selección del nivel de descomposición L. El filtrado con *wavelets* es dependiente del valor de L (Nounou y Bakshi, 1999). La eliminación de coeficientes a un valor muy alto de L puede conducir a la eliminación de características importantes de la señal, mientras que la eliminación a valores muy bajos de L, podría no eliminar suficiente ruido de la señal. Por lo tanto, una correcta selección del valor de L podría ser potencialmente útil para asegurar estimados de filtración óptimos. En un trabajo precedente (Roy *et al*., 1999) se planteó una estrategia de filtración que incluye una estimación de L. Dicha estrategia es como sigue:

- Primero se toma la derivada de la señal medida y(k).
- Sobre la derivada se aplica la descomposición *wavelets* según la ecuación (2.8), a distintos valores de l tal que se obtienen los detalles a distintas escalas l.
- Para cada vector de detalles se calcula su energía[a] denotada por P_{L}. Luego, detectan el mínimo global de la curva de los P_{L}, el cual indica el valor del L.
- Finalmente, la filtración consiste en fijar en cero todos los coeficientes d entre las escalas 1 y el L determinado, tomar el $\mathbf{A}_L(k)$ obtenido sobre los coeficientes que no se fijaron a cero como la señal rectificada, y luego, reconstruir la señal original sin ruido por la integración de la señal $\mathbf{A}_L(k)$.

En el procedimiento anterior, al tener que integrar (para reconstruir la señal original filtrada) el filtrado final se hace dependiente de los valores de las condiciones iniciales de dicha integración. En un análisis fuera de línea esto no presentaría problema ya que se tiene el tiempo necesario para probar distintos valores y ver cual es el que produce la mejor curva integrada. Sin embargo, si se utiliza en línea, esto no sería práctico ya que el tiempo de trabajo es limitado, además de que si se tiene un número considerable de variables a filtrar la limitación de tiempo crecerá. Por otro lado, por experimentación se vio que con algunas señales y tras la integración, los extremos de la señal filtrada recuperada quedan

[a] La energía es una medida que intenta expresar la cantidad de información contenida en una señal.

distorsionados respecto a las curvas reales **y***(*k*). Luego, en la metodología propuesta por Roy et al. (1999), se tendría que probar algún tipo de extensión que intente compensar la distorsión en los extremos. Finalmente, la eliminación directa de todos los coeficientes *d* entre las escalas 1 y *L* (fijándolos a cero), puede conllevar la eliminación de características propias de **y***(*k*) (ver discusión de la sección 2.1.1.2.2) por lo que el procedimiento propuesto por Roy et al. (1999), conducirá en muchos casos a filtrados erróneos.

Otro de los problemas se asocia a la selección de la función *wavelets*. En la literatura se han propuesto muchísimas familias de funciones *wavelets* (Addison, 2002; Misiti *et al*., 2004). De entre ellas, la familia conocida como *wavelets Daubechies* o *dbN* es la que se ha visto más eficiente para aplicaciones de filtrado dentro (Flehmig *et al*., 1998; Bakhtazad *et al*., 1999; Sun *et al*., 2003; Dash *et al*., 2004) y fuera (Ghael *et al*., 1997; Bakhtazad *et al*., 1999; Li *et al*., 2002; Köhler y Lorenz, 2004) de la IQP. La razón de esto radica en las muy buenas capacidades de este tipo de *wavelets* para representar comportamientos polinómicos y/o no lineales de diversas señales y, en consecuencia, la utilidad para estimar la señal fundamental de mediciones ruidosas (Rowe y Abbott, 1995; Flehmig *et al*., 1998; Addison, 2002). Las diferentes funciones *dbN* se generan a partir del aumento del parámetro *N* que indica el orden de las mismas, lo que hace que la función resultante se haga más suavizada a medida que N se hace mayor (Rowe y Abbott, 1995). Esto conduce a que, en teoría, las *dbN* con orden bajo ($N \rightarrow 1$) sean mejores para tratar series con patrones estacionarios o discontinuidades fuertes, mientras que con ordenes altos ($N \rightarrow \infty$) sean más apropiadas para tratar señales suaves (ver discusión en Addison, 2002). Sin embargo, en la literatura sobre filtrado frecuentemente ocurre que:

- La *dbN* seleccionada varía de un autor a otro y, en ocasiones sobre unos mismos datos de prueba (Flehmig *et al*., 1998; Nounou y Bakshi, 1999; Dash *et al*., 2004).
- La selección anterior se justifica argumentando que tras experimentar con los datos utilizados se vio que una *dbN* con un orden *N* se encontró muy apropiada. Dichas "experimentaciones" no están documentadas.

Esto plantea que, en la práctica, la selección del orden *N* asociado a las diferentes *wavelets* (*N* = 1,2,…) puede afectar la calidad de la rectificación de la señal (Nounou y Bakshi, 1999).

Un tercer problema tiene que ver con la aplicación en línea de los métodos de filtración con *wavelets*. Nounou y Bakshi (Nounou y Bakshi, 1999) abordaron el problema y propusieron 2 esquemas de solución:

- El *OLMS* (*On Line Multiscale Filtering*) en el que se aplica el filtrado con *wavelets* sobre una ventana móvil de longitud diádica (esto es, potencias de 2) dando lugar a filtrados en cada instante de los valores nuevos que se producen.
- *BCTI* (*Boundary Corrected Translation Invariant*) que es similar al *OLMS* pero que añade una corrección de varios instantes de tiempo sobre los valores filtrados y que es aplicable a situaciones donde no se requiere que se obtenga el filtrado inmediatamente después que se reciben nuevos datos.

A pesar del buen rendimiento de estas estrategias (*OLMS* y *BCTI*), no se considera la toma de decisión asociada a la ocurrencia de eventos tipo salto en los que el patrón de la señal cambia y para los cuales podría ser apropiado redefinir el tamaño de ventana utilizada, *nc*, o modificar el valor del nivel de descomposición *L* actual. Esta última observación es particularmente importante para muchas aplicaciones de proceso en las que el requerimiento de respuestas instantáneas (monitorización, detección de fallos) obliga a un esquema *OLMS*.

Todos los tópicos anteriores son abordados a través de una serie de estudios y metodologías que se describen en las secciones que siguen. En cuanto a la aplicación en línea de las estrategias de filtrado y dado el requerimiento crítico de respuestas instantáneas en la aplicación de muchas tareas operacionales como control regulador o detección de fallos, solo se evaluará el caso *OLMS*, discutido en los párrafos anteriores.

2.2 Análisis experimental de la aplicación de wavelets Daubechies en filtrado de datos.

En este apartado, se presenta un análisis empírico del rendimiento de las *wavelets dbN* en aplicaciones de filtrado. El análisis se centra en medir la habilidad de filtrado de *wavelets dbN* y bajo esquemas *waveshrink* con diversos valores en el parámetro *L*. Adicionalmente, debido a que en muchas situaciones de operación, control y monitorización de plantas se requieren datos de calidad tan pronto como se producen, el estudio hace hincapié en aplicaciones de filtración en línea.

2.2.1 Experimentos de filtración basados en wavelets

Se llevaron a cabo una serie de experimentos (simulaciones) de filtración. Se utilizan *wavelets dbN* como funciones de filtrado. En los experimentos se intenta estudiar el rendimiento de la filtración sujeta a distintas combinaciones *dbN* - *L*.

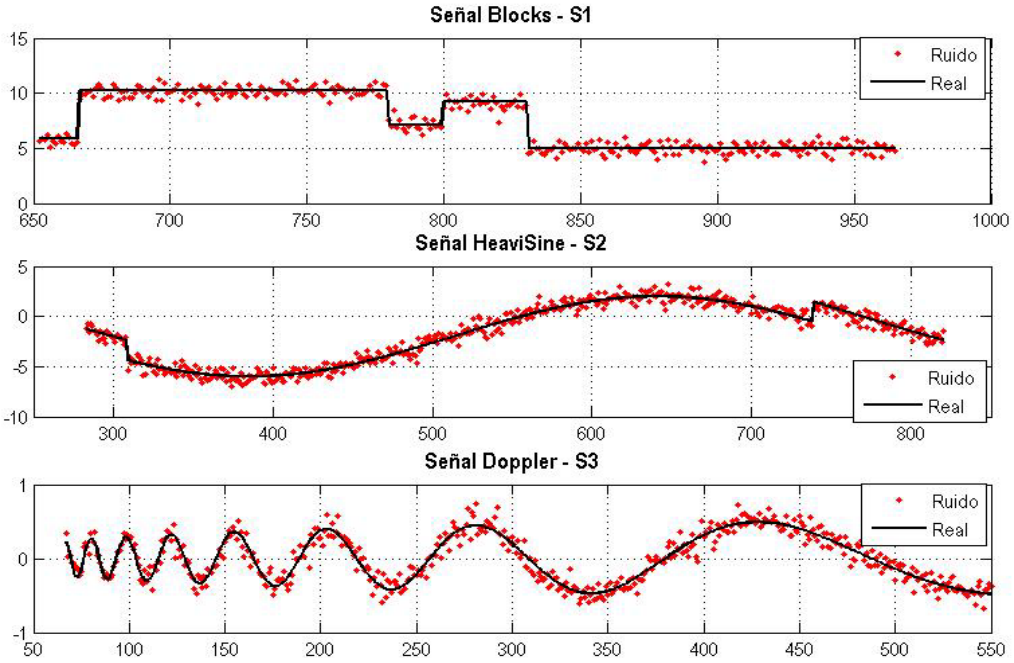

Figura 2.3. Señales de Referencia para los Experimentos

Los experimentos se organizaron como sigue:

- Se utilizaron señales típicas de la literatura (Doymaz *et al*., 2001): La señal *Blocks*, la *HeaviSine* y la *Doppler*. Se escogen estas señales por la diversidad de patrones y características de curva que contienen (Taswell, 2000). Las señales originalmente contienen 1024 observaciones. En los experimentos, las señales se utilizaron en los siguientes intervalos: (1) La señal *Blocks* o *S1* desde 620 a 965; (3) La señal *HeaviSine* o *S2* desde 251 hasta 820; (4) La señal *Doppler* o *S3* desde 35 hasta 550 (ver figura 2.3).

- Todas las señales se contaminaron con ruido gausiano aleatorio con media 0 y varianza 0.5.
- Se aplicaron *Daubechies* de orden 1 hasta orden 9 (*db1* hasta *db9*) a cada señal. Se utilizo hasta un orden 9 por que se vio que en toda la literatura revisada solo utilizan *dbN* con $N \leq 8$.
- Por cada *dbN* se hicieron diferentes experimentos, en cada uno de los cuales se utilizó un *L* constante. Los valores de *L* que se utilizaron fueron de 2 a 9.
- Cada combinación "*dbN-L*" se evaluó bajo el esquema de filtración en línea *OLMS* (ver sección 2.1.2). Así, en cada instante se crea una ventana móvil de longitud diádica (con las 32 últimas observaciones en cada instante) sobre la que se aplica la filtración. Esto implica 2 cosas: (1) Los filtrados a lo largo del tiempo están formados por el último valor de cada ventana filtrada en cada instante; (2) De la primera ventana que se filtra en el instante cero solo se guarda el último valor y el resto se pierde en el siguiente instante.
- Cada combinación "*dbN-L-OLMS*" se probó utilizando *waveshrink* bajo *Soft thresholding* y bajo *Hard thresholding*.
- El valor de β se fijó según la regla *visushrink* (ver sección 2.1.1.2.2).

2.2.2 Rendimiento local de los filtrados con wavelets dbN

Para medir con detalle los resultados de las simulaciones del filtrado en línea se adoptó el siguiente procedimiento de análisis (ver figura 2.4):

1. Se toma el vector de una señal filtrada $\hat{\mathbf{y}}^*(k)$, obtenida en una simulación, y el correspondiente vector de la señal original sin ruido, $\mathbf{y}^*(k)$.
2. Se construyen 2 versiones de la ventana *w1*: Una con los primeros 32 valores filtrados $\hat{\mathbf{y}}^*_{W1}$ y otra con la señal original sin ruido $\mathbf{y}^*_{W1}$.
3. Se calcula el Error Cuadrático Medio o *ECM* o *w1* (ECM_{W1}) como sigue:

$$ECM_{W1} = \frac{1}{nc_w}\sum_{t=1}^{nc_w}\left(\hat{\mathbf{y}}^*_{W1}(k) - \mathbf{y}^*_{W1}(k)\right)^2 \qquad (2.16)$$

   Donde nc_w es la longitud de la ventana *w1*.
4. Se repiten (2) y (3) sobre una nueva *wi* superpuestas a la mitad con la *wi* anterior.
5. Se repiten (2), (3) y (4) hasta que se procese toda la señal.

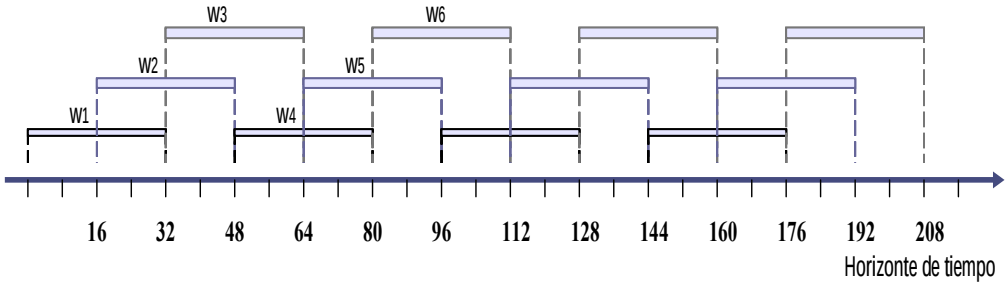

Figura 2.4. Ventanas móviles para el calculo del *ECM* local

Habiendo procesado todos los filtrados obtenidos (a partir de cada simulación disponible) para todas las señales (*S1*, *S2* y *S3*) según el procedimiento anterior, se obtiene el siguiente grupo de información:

- Un gráfico de la señal original con el número de intervalos o ventanas (w_i) consideradas.

- Un gráfico con el ECM mínimo por cada w_i y bajo las combinaciones: *OLMS* con *thresholding-Soft* (*OLMS-S*) y *OLMS* con *thresholding-Hard* (*OLMS-H*).
- Una tabla con la combinación "*L – dbN*" utilizada para obtener los ECM más pequeños por cada w_i y asociados a los casos *OLMS-S* (S) y *OLMS-H* (H).

A continuación se presenta el análisis de resultados para cada señal. Se hace un énfasis especial en las aplicaciones en Línea y, en consecuencia, se discuten principalmente los grupos asociados a los casos *OLMS*.

2.2.2.1 Análisis para la Señal S1

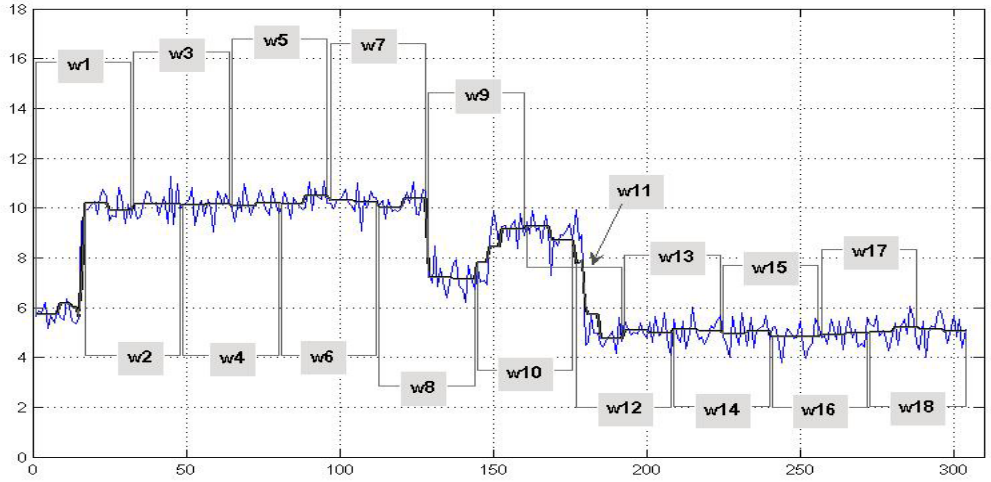

Figura 2.5. Intervalos considerados para *S1*.

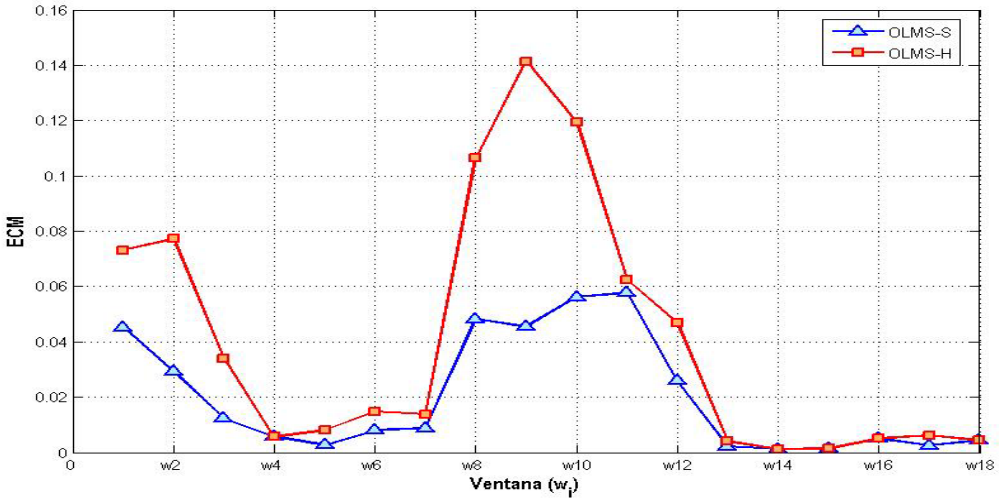

Figura 2.6. *ECM* mínimo en cada intervalo para *S1*

Tabla 2.1. Valores de *L* y *dbN* que conducen a *ECM* mínimo por periodo (*S1*).

Nº w_i		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Nivel	**S**	3	3	4	5	5	5	9	3	3	3	3	3	3	4	5	5	3	5
	H	2	4	5	5	5	8	8	2	2	4	4	4	4	4	5	5	5	5
dbN	**S**	1	1	1	1	1	1	9	1	1	1	1	1	1	1	1	1	1	1
	H	1	1	1	1	5	9	1	7	7	4	1	1	1	1	1	1	1	1

Analizando la tabla 2.1, se observa que el valor óptimo de *L* varía aunque muestra ciertos patrones. En los intervalos que incluyen los saltos, el *L* que produce el menor ECM está entre 2-4 (aunque más inestable para *Hard* que para *Soft*), mientras que en el resto de los intervalos (estacionarios) el *L* tiende a 4-5. En cuanto a la *dbN*, la *db1* es la mejor opción para zonas con saltos y zonas con nivel constante con algunas excepciones bajo la opción *OLMS-H*.

2.2.2.2 Análisis para la Señal S2

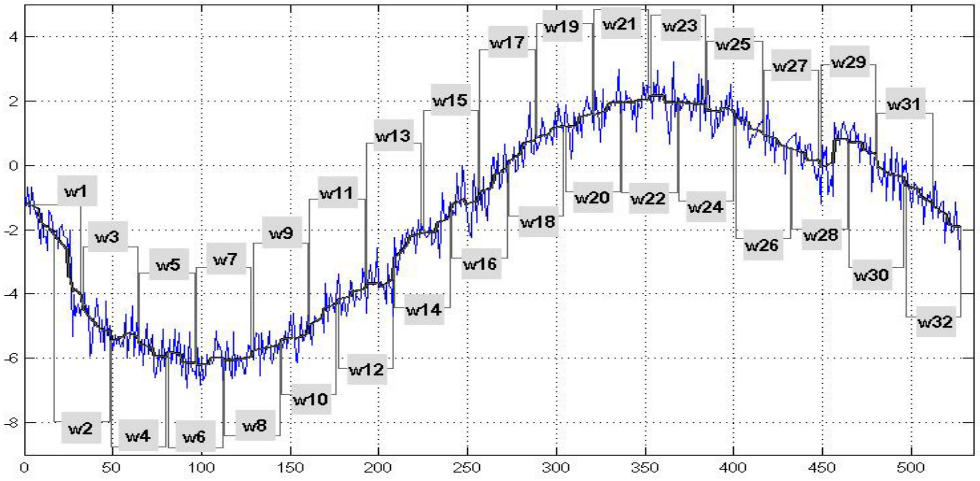

Figura 2.7. Intervalos considerados para *S2*.

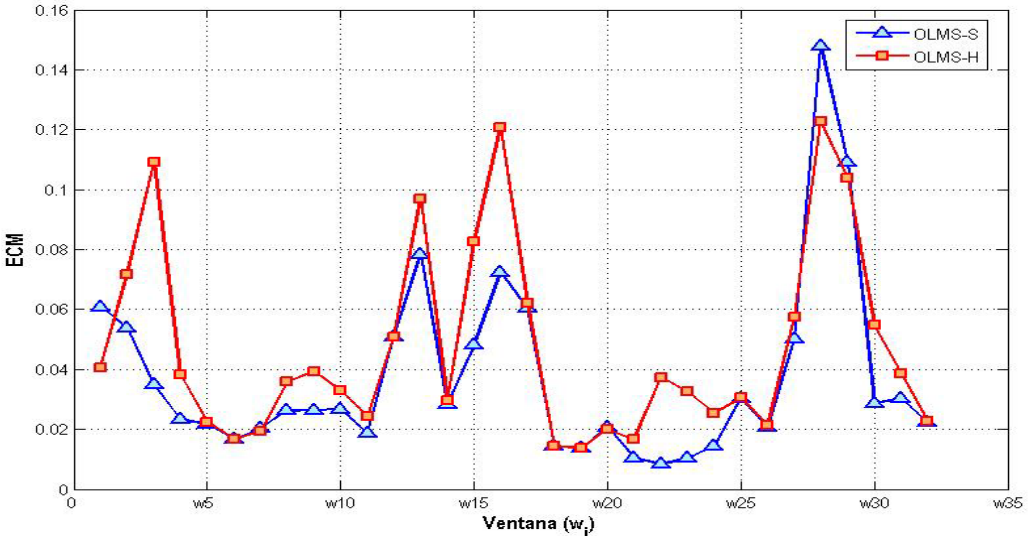

Figura 2.8. *ECM* mínimo en cada intervalo para *S2*.

Tabla 2.2. Valores de *L* y *dbN* que conduce a ECM mínimo por periodo (*S2*).

N° w_i		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Nivel	S	4	3	3	3	4	4	5	3	3	3	3	3	5	5	4	4	4	3	4	4	4	4	5	4	4	4	4	3	2	3	4	4
	H	4	3	3	4	4	4	5	4	5	4	5	3	3	5	5	9	4	3	4	4	4	5	6	5	4	5	3	3	3	5	5	4
dbN	S	7	6	1	1	4	1	1	1	1	1	1	1	8	9	4	4	4	1	1	1	1	1	1	1	4	4	7	4	1	1	4	4
	H	7	8	8	4	4	1	1	1	2	4	4	1	1	9	9	4	4	1	1	1	1	8	8	8	4	9	1	1	1	9	9	4

En este caso, el esquema *OLMS-S* es mejor que el *OLMS-H* (ver figura 2.8). En cuanto al valor de *L* (ver tabla 2.2) en los períodos que incluyen el salto hacia el final (w_i 1 y 28), el valor de *L* es más alto (*L*=4 y *L*=3) que en los períodos donde el salto se produce hacia el

principio (L=3 en intervalo 2 y L=2 en intervalo 29). En cuanto al *dbN*, se observan algunos patrones interesantes:

Zonas sin discontinuidades: En los periodos donde la curva tiene pendiente cada vez menor e incluso nula (aprox. de w_i 3 a 11, y de w_i 19 a 24 en figura 2.7) *db1* tiende a ser la que contribuye al ECM mínimo, mientras que en los periodos con pendientes más pronunciadas (aprox. de w_i 13 al 18 y aprox. de w_i 26 en adelante. Figura 2.7) *db4* y, en menor grado *db8* y *db9*, son las que contribuye al ECM mínimo. Esto concuerda con la teoría de que en curvas con pendientes fuertes es mejor usar *dbN* con valores de *N* grandes (ver sección 2.1.2).

Zonas con discontinuidades: En periodos que incluyen saltos no siempre se sigue el patrón que ocurría en *S1*, es decir, no solo la *db1* llega a ser la más apropiada (w_i 29 en figura 2.7) sino también la *db4* (w_i 28), la *db6* (w_i 2) y la *db7* y la *db9* (w_i 1). Esto se debe al patrón descrito en zonas sin discontinuidades. En efecto, debido a que en los puntos vecinos a los saltos la curva es suave (figura 2.7), se tiende a producir una compensación entre la discontinuidad y el patrón más suave por lo que no siempre *db1* es la más hábil.

2.2.2.3 Análisis para la Señal S3

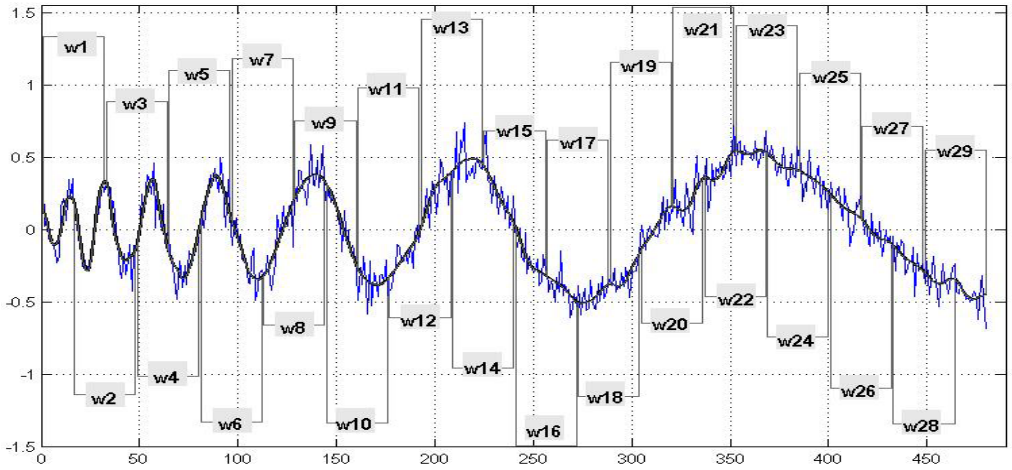

Figura 2.9. Intervalos considerados para *S3*.

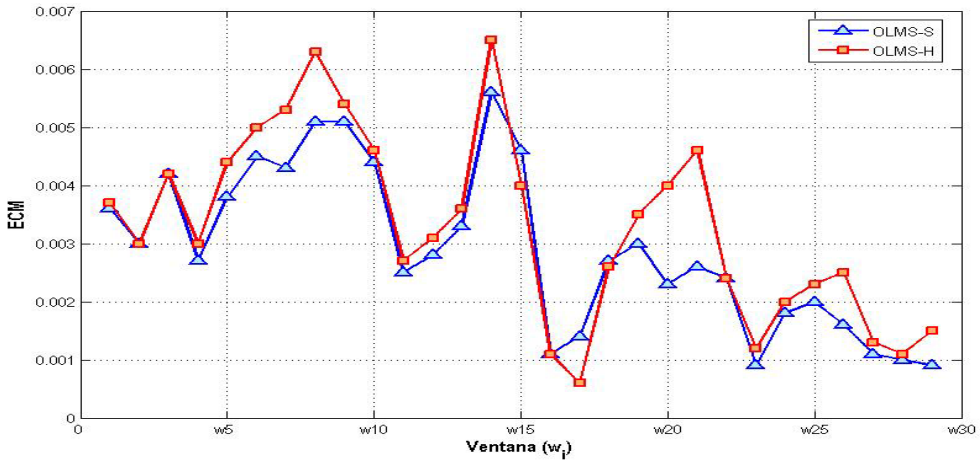

Figura 2.10. *ECM* mínimo en cada intervalo para *S3*.

Tabla 2.3. Valores de *L* y *dbN* que conduce a ECM mínimo por periodo (*S3*).

Nº w_i		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Nivel	**S**	2	2	2	3	3	3	3	4	2	3	4	4	4	4	2	4	4	4	4	4	5	2	6	4	5	3	4	4	4
	H	2	2	2	3	3	3	8	3	3	3	4	3	4	4	2	4	7	4	4	9	2	3	6	3	4	4	4	5	5
dbN	**S**	7	7	8	8	4	2	8	8	4	6	8	2	8	8	1	4	4	8	8	4	2	1	9	8	2	3	7	4	4
	H	7	7	8	8	4	4	2	8	2	2	8	8	8	8	1	4	4	8	2	4	1	1	8	1	2	2	8	2	2

En este caso, los ECM tienden a decrecer a medida que la amplitud de la curva crece (ver figura 2.10). Este patrón decreciente también marca, aunque un tanto difusos, ciertos patrones en el *L* y las *dbN*. Para el caso de *L* (ver tabla 2.3), en zonas con pendientes muy pronunciadas (amplitudes pequeñas) los valores tienden a ser muy bajos: 2 y 3. Luego, a medida que las pendientes se suavizan más, *L* tiende a tomar valores cada vez más altos aunque con una cierta primacía de *L*=4.

Para el caso de *dbN*, sucede algo parecido al de *L*, es decir, en los primeros 14 intervalos predominan *dbN* con *N* altos (7 ocurrencias de *db8*, 2 de *db7* y ninguna de *db1* ni *db3*), mientras que en los últimos 15 valores aparecen más *dbN* con valores de *N* bajos (5 ocurrencias de *db4*, 2 de *db1*, 2 de *db2* y 1 de *db3*).

2.2.3 Comentarios globales sobre los experimentos

Los análisis anteriores plantean como conclusiones:

- La estrategia *OLMS-S* es más apropiada para diferentes patrones de curvas que *OLMS-H*.
- Los valores de *L* que pueden conducir a estimados con bajos ECM dependen un tanto del patrón de la curva. Así, en presencia de discontinuidades o pendientes fuertes como las vistas en las curvas *S1* y *S2*, Los valores de *L* más apropiados deben estar entre 2 y 3 (y ocasionalmente 4), mientras que en curvas con pendientes suaves o sin pendientes (estacionarias) los valores de *L* deberían tender a valores entre 4 y 5 (y ocasionalmente 6).
- La *db1* puede ser muy útil para asegurar estimaciones con bajos ECM para tipos de patrones como los observados en las curvas *S1* y gran parte de *S2*. Otras *dbN* como la *db4* y la *db8*, pueden llegar a ser también útiles en casos con patrones como los de las curvas *S2* y *S3*.

Las anteriores conclusiones coinciden en gran parte con la teoría de *dbN*. No obstante, permiten ver como se puede obtener un mejor aprovechamiento de *dbN* en la práctica. Se podría haber explorado el uso de otros tipos de *wavelets* como las *Symlet* o las *Coiflets* también desarrolladas por *Daubechies* para mejorar la simetría de las *dbN*. En trabajos recientes (Trygg y Wold, 1998; Teppola y Minkkinen, 1999) se utilizan *wavelets* tipo *Symlet* para modelar datos multivariables. Los autores coinciden en que el uso de *Daubechies* y *Symlet*, sobre distintos problemas, no lleva a diferencias significativas en la calidad de las tendencias de variables modeladas a diferentes escalas. No obstante, en tiempo de cálculo, *dbN* se mostró más rápida que la *Symlet* (Trygg y Wold, 1998), lo cual realza la utilidad de medir el rendimiento de las *dbN* para filtrado.

2.3 Análisis del filtrado usando wavelets y con selección del nivel de descomposición óptimo

En esta sección, se explora una estrategia para estimar el nivel óptimo de descomposición de una señal mediante *wavelets*. Luego, se estudia la integración de esta estrategia dentro del esquema *waveshrink*. La estrategia final integrada se identifica como *levashrink* a fin de poder diferenciarla del *waveshrink* dentro de este capítulo.

2.3.1 Identificación del nivel óptimo de descomposición

De la teoría expuesta en secciones precedentes, se puede establecer cómo obtener la descomposición de una señal medida $\mathbf{y}(k)$ usando *wavelets* (ver ecuaciones 2.8 y 2.13). Se podrían obtener descomposiciones a infinitas resoluciones variando el valor de l entre 1 y $+\infty$. En la práctica una señal se mide a una resolución finita y su escala solo se puede variar sobre un rango finito. Luego, la posibilidad de descomposición con *wavelets* queda restringida entre la escala de mínima resolución, esto es, la escala a la cual se miden los datos (l=1), y la escala de máxima resolución identificada como $l = L_{\mathrm{m}}$. Por otro lado y tomando en cuenta lo anterior Mallat (Mallat, 1989) propuso (a través de la derivación de la ecuación de energía representada por (2.17)) que, en un caso ideal, la información de alta frecuencia que forma parte de $\mathbf{y}(k)$ solo se asociaría a los detalles que se obtienen mediante la descomposición con *wavelets* entre $l = 1$ y $l = L_{\mathrm{m}}$.

$$\left|\mathbf{y}(k)\right|^2 = \sum_{v\in Z}\left|a_{L,v}\cdot\phi_{L,v}(k)\right|^2 + \sum_{l=1}^{L}\left|\sum_{v\in Z} d_{l,v}\cdot\psi_{l,v}(k)\right|^2 \tag{2.17}$$

o equivalentemente:

$$\left|\mathbf{y}(k)\right|^2 = \left|A_L\right|^2 + \sum_{l=1}^{L}\left|D_l\right|^2 \tag{2.18}$$

La energía en este caso es una medida que intenta expresar la cantidad de información contenida en una señal o en parte de la misma. Las ecuaciones (2.17) y/o (2.18) plantean que, idealmente, para descomposiciones hasta niveles $L<L_{\mathrm{m}}$, solo se retiene información de alta frecuencia (comúnmente representando el ruido) en los detalles D_{l} mientras que la información de baja frecuencia (características de la señal) se retendrá en la aproximación obtenida A_{L}. También, para valores de $L>L_{\mathrm{m}}$ es posible que alguna información de baja frecuencia se retenga en los detalles y no solo en las aproximaciones. Por lo tanto, se puede llegar a obtener una buena aproximación a la señal verdadera $\mathbf{y}^*(k)$ por aplicar la descomposición de $\mathbf{y}(k)$ hasta la escala L_{m}. A pesar de lo anterior, en la teoría no se establece como identificar L_{m}.

A través de múltiples experimentos sobre distintas señales (ver sección 2.5) se evaluó el siguiente procedimiento:

1. Se obtienen aproximaciones de $\mathbf{y}^*(k)$ ($\mathbf{A}_L(k)$) a diferentes valores de L y según la ecuación 2.9.

2. Se computa la energía de los detalles (como la potencia P_{L}) eliminados a distintos valores de L como sigue:

$$P_L(\mathbf{D}_{l=1,\ldots,L}) = \sum_{l=1}^{L}\left|\mathbf{D}_l(k)\right|^2 = \left|\mathbf{y}(k) - \mathbf{A}_L(k)\right|^2 \tag{2.19}$$

3. Se calcula la variación de potencia entre sucesivas escalas como sigue:

$$\Delta P_L = P_L(\mathbf{D}) - P_{L-1}(\mathbf{D}) \tag{2.20}$$

Como resultado de este procedimiento, se obtiene un conjunto de valores de ΔP_L. Sobre este conjunto se observó que al ir incrementando la escala, los valores de ΔP_L decrecen rápidamente hasta alcanzarse un primer mínimo, seguido por un incremento en los valores de ΔP_L. De acuerdo a la ecuación (2.17), este incremento se puede explicar diciendo que hasta el primer mínimo, los coeficientes de detalle eliminados representan básicamente información de alta frecuencia y sus valores se incrementan suavemente con la escala. Pero después de este primer mínimo los detalles también contendrían información de alta frecuencia que es mayor, en magnitud de energía, que la información de baja frecuencia. Esto último provoca un nuevo y significativo incremento en ΔP_L. Luego, la detección del primer mínimo en ΔP_L puede usarse para identificar L_{m}. Así se propone determinar L_{m} como sigue:

4. Se obtienen aproximaciones de $\mathbf{y}^*(k)$ ($\mathbf{A}_L(k)$) a distintos valores de L (ec. 2.9).

5. Se calculan los P_L de los detalles eliminados a distintos valores de L (ec. 2.19).

6. Se calcula ΔP_L entre sucesivas escalas (ec. 2.20).

7. Se identifica el primer mínimo en ΔP_L y su L asociado. Este último, se fija como el L_{m}.

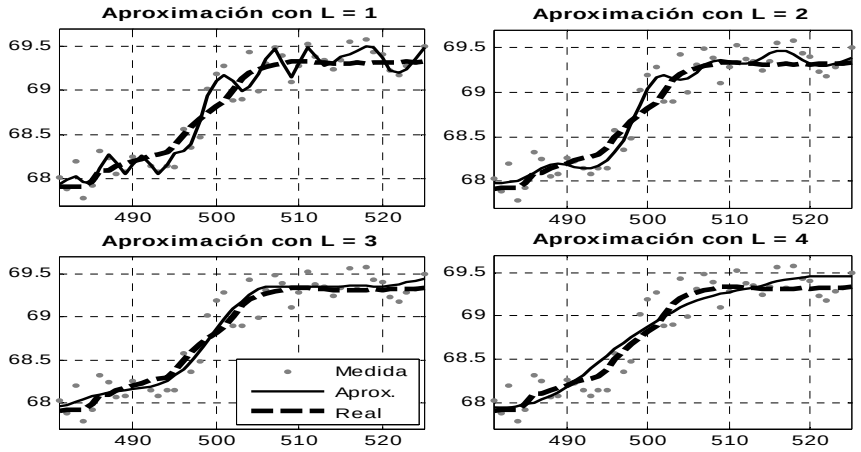

Figura 2.11. Obtención del $\mathbf{A}_{\mathrm{L}}$ para la curva *Jumps* (entre L=1 y L=4).

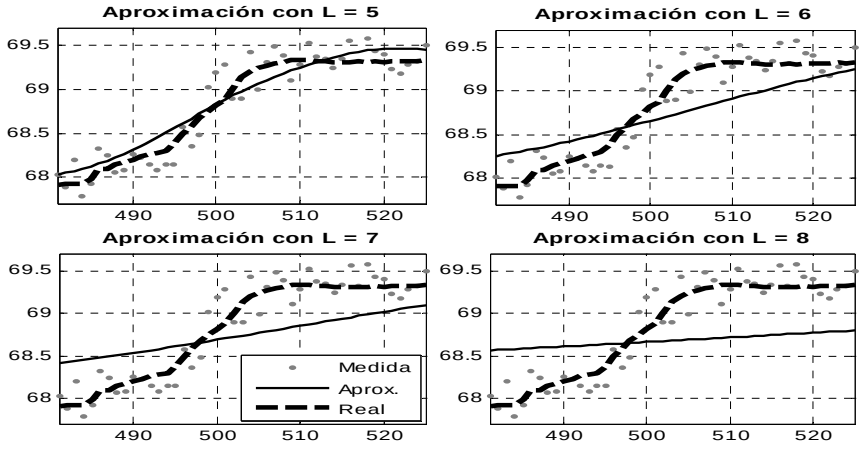

Figura 2.12. Obtención del $\mathbf{A}_{\mathrm{L}}$ para la curva *Jumps* (entre L=5 y L=8).

El procedimiento anterior se ilustra mediante un ejemplo:

- Se tienen datos de un experimento cualquiera (afectados por ruido) de una curva $\mathbf{y}(k)$ a la que se le llama *Jumps* (ver la curva "Medida" en las figuras 2.11 y 2.12). De esta misma curva, se conoce la versión original sin ruido $\mathbf{y}^*(k)$ (curva "Real" en las figuras 2.11 y 2.12).
- Variando L desde 1 hasta 8 y utilizando una *db8*, se calculan varias aproximaciones $\mathbf{A}_L(k)$ (ver figuras 2.11 y 2.12).
- Se calculan las potencias P_L y las ΔP_L. Al hacer un gráfico de los valores sucesivos de ΔP_L se observan uno o más mínimos (ver figura 2.13).
- Se identifica el primer mínimo (L =3) y se etiqueta como L_m. Luego, se observa que el $\mathbf{A}_L(k)$ asociado al L_m, muestra un comportamiento muy cercano a $\mathbf{y}^*(k)$ (ver la curva "Aprox." en el gráfico inferior derecho de la figura 2.11).

De esta manera, $\mathbf{A}_{L=L\mathrm{m}}(k)$ representa una aproximación bastante buena de la señal original $\mathbf{y}^*(k)$. También se observa que, para valores de $l > L_\mathrm{m}$, la aproximación que se obtiene se distorsiona respecto a la tendencia local, lo que provoca una progresiva eliminación de características propias de $\mathbf{y}^*(k)$. Esto se refleja en los ΔP_L que crecen considerablemente después de l=3 (Ver gráfico superior de la figura 2.13).

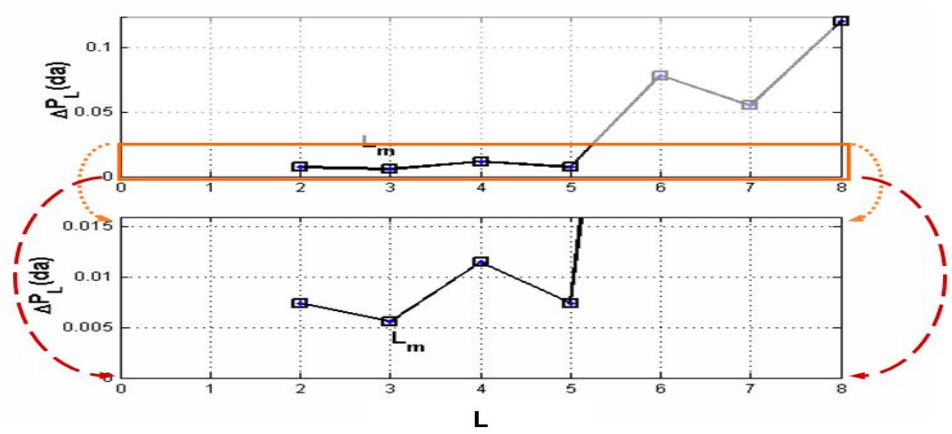

Figura 2.13. Determinación del nivel optimo L_m.

Finalmente, en la figura (2.14) se muestra un esquema del procedimiento propuesto para la detección de L_m.

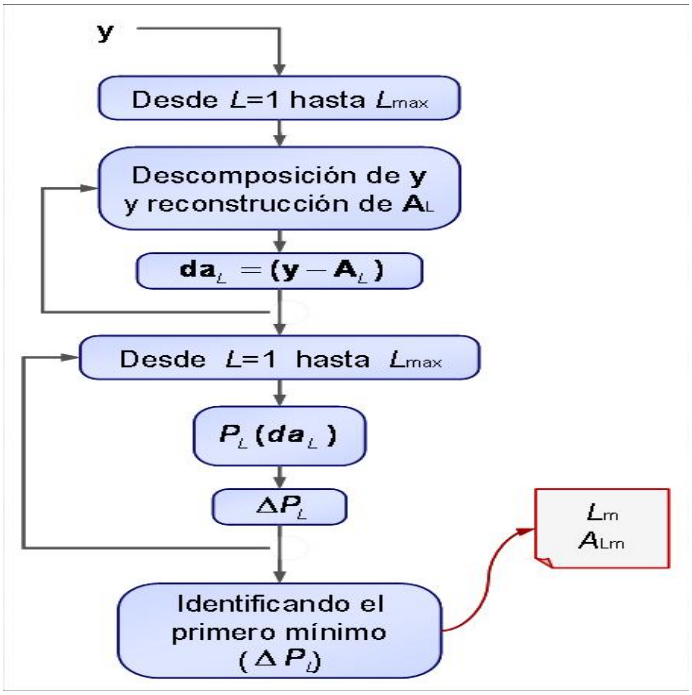

Figura 2.14. Procedimiento para identificar L_m

2.3.2 La Estrategia levashrink

Si consideramos el problema de estimar $\mathbf{y}^*(k)$ a partir de sus lecturas observadas $\mathbf{y}(k)$, el procedimiento anterior asegura una buena aproximación de $\mathbf{y}^*(k)$ pero no asegura la señal con mínimo o ningún ruido. Alguna pequeña porción de información de baja frecuencia de la señal se podría perder junto a los detalles eliminados entre $l = 1$ y $l = L_m$. Por lo tanto, se deberían identificar y retener los coeficientes de detalle con información de baja frecuencia para asegurar la mejor estimación de $\mathbf{y}^*(k)$. Se propone que el procedimiento de identificar L_m se podría utilizar como un paso inicial y, luego, se podría aplicar un filtrado, como el *waveshrink*, sobre la descomposición de $\mathbf{y}(k)$ hasta la escala L_m. Como consecuencia, el procedimiento descrito en la sección anterior (ver sección 2.3.1) podría redefinirse como sigue:

1. Se aplica la descomposición *wavelets* de $\mathbf{y}(k)$, según la ecuación (2.8), para obtener los coeficientes de escala, $\mathbf{a}_{L,v}$, y los coeficientes *wavelets*, $\mathbf{d}_{l,v}$. Se descompone hasta una escala L tomando valores enteros entre 8 y 9.
2. Se aplican los pasos del método para identificar L_m (bloque L_m).
3. Se aplica un *thresholding* como en el *waveshrink* (ver ecuaciones 2.12 y 2.13) sobre todos los coeficientes *wavelets* desde la escala 1 hasta la escala L_m. Los coeficientes que quedan después de este paso se identifican como $d^*_{l,v}$.
4. Se recupera la señal filtrada como:

$$\hat{\mathbf{y}}(k) = \sum_{v \in Z} a_{Lm,v} \cdot \phi_{Lm,v}(k) + \sum_{l=1}^{Lm} \sum_{v \in Z} d^*_{l,v} \cdot \psi_{l,v}(k) \qquad (2.21)$$

En la figura 2.15 se ilustra el esquema integrado (que llamamos *levashrink* para diferenciarlo del *waveshrink*).

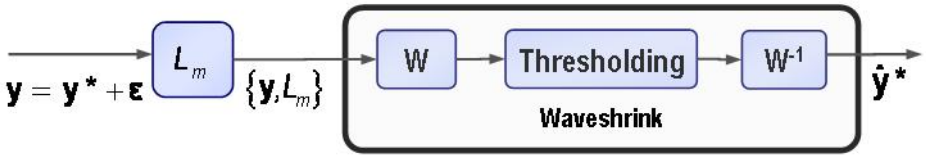

Figura 2.15. Esquema *levashrink*.

2.4 Rectificación Combinada

Como se señaló en la sección 2.1.2, la selección de la *wavelets* puede afectar la calidad del filtrado. Basándose en el uso de las *wavelets dbN*, Nounou y Bakshi (1999) propusieron que la *db1* podría ser más apropiada para tratar señales con patrones estacionarios o cuadrados (por ejemplo, una señal con discontinuidades en forma de escalón), mientras que para señales más suavizadas recomendaron el uso de *db2*. El análisis de la sección 2.2 sirvió para verificar que la *db1* es particularmente apropiada para señales estacionarias y/o con discontinuidades fuertes, aunque también se mostró adecuada para procesar señales con pendientes leves crecientes o decrecientes. Asimismo, se vio que las *db4* y *db8* son útiles para obtener rectificados de calidad a partir de curvas suaves con pendientes diversas (desde leves a muy bruscas)[a]. Todo lo anterior evidencia que no hay una única *dbN* que posibilite buenos rectificados sobre curvas con distintos patrones y, en la práctica, la selección de la *dbN* apropiada seguirá implicando un trabajo previo de ensayo sobre las curvas a analizar que, en el caso de aplicaciones en línea, no siempre es factible de realizar ni en tiempo ni en costos.

En esta sección se estudia una estrategia que combine las capacidades de distintas *dbN* con los objetivos de: ahorrar el paso de selección de la *wavelets*, asegurar un rango amplio de aplicación y que a la vez produzca estimados de calidad. La idea base para proponer la presente estrategia se toma de lo que se conoce como Previsión Combinada (PreviCom). La PreviCom, fue propuesta (Bates y Granger, 1969) como una alternativa para intentar mejorar las previsiones obtenidas individualmente por diferentes métodos de series temporales. En las décadas que siguieron a esta propuesta y hasta hoy se han hecho múltiples estudios sobre la validez y la efectividad de este concepto, llegándose a la conclusión de que la combinación es una alternativa sencilla y válida de obtener posibles mejoras en predicción (Armstrong, 1989; Makridakis *et al*., 1998; Chatfield, 2002; Yang, 2004). La idea, ha sido trasladada a otras áreas como por ejemplo clasificación de datos (Cho y Kim, 1995) y modelado de *Soft-sensors* (Zhong y Yu, 2000). El esquema más simple de una estrategia PreviCom se ilustra en la figura 2.16.

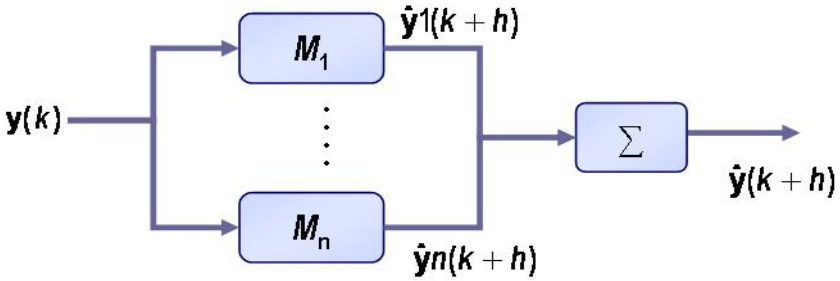

Figura 2.16. Esquema General de Previsión Combinada.

[a] En una sección posterior (sección 2.7), también se verá que las conclusiones sobre las *dbN* se mantienen cuando se utilizan el esquema de filtración propuesto en este capítulo (*levashrink*).

La serie de datos $\mathbf{y}(k)$ se modela por más de una técnica (*Mi*), obteniéndose diferentes modelos. Con cada uno de ellos se produce una estimación de $\mathbf{y}_t$ en los h instantes futuros ($\hat{\mathbf{y}}\mathrm{i}(k+h)$). Finalmente, se hace un promedio con todos los $\hat{\mathbf{y}}\mathrm{i}(k+h)$ resultando en el $\hat{\mathbf{y}}(k+h)$ final. El esquema anterior basado en un promedio simple es el caso más sencillo y de los más usados. En la literatura sobre el tema, se pueden encontrar muchas otras alternativas de combinación (Delurgio, 1998; Terui y Van Dijk, 2002; Yang, 2004).

Volviendo al caso de la rectificación, se quiere obtener un estimado $\hat{\mathbf{y}}^*(k)$ de la señal original $\mathbf{y}^*(k)$, a partir de una realización $\mathbf{y}(k)$ derivada de $\mathbf{y}^*(k)$ pero con ruido añadido. En este caso, las *dbN* representaran las *M*i de la figura 2.16, y las respuestas de cada *M*i serán distintos estimados de $\mathbf{y}(k)$ ($\hat{\mathbf{y}}_i^*(k)$). Luego, el estimado final $\hat{\mathbf{y}}^*(k)$ se obtendrá como un promedio ponderado de los $\hat{\mathbf{y}}_i^*(\mathrm{k})$:

$$\hat{\mathbf{y}}^*(k) = \sum_{i=1}^{F} pw_i \cdot \hat{\mathbf{y}}_i^*(k) \qquad (2.22)$$

En la ecuación anterior, F indica el número de filtros utilizados y pw_i representa los pesos asignados a cada estimado. Como una primera aproximación se podría asignar a cada peso un mismo valor igual a *1/F*. De esta manera, el estimado resultante se obtiene por promedio simple. No obstante, dado que se ha visto que ciertas *dbN* son más apropiadas para ciertos patrones de señal, el valor de los pesos debería ser mayor para aquellos estimados que representen mejor el patrón actual de la curva. Para intentar deducir un peso que se aproxime a lo anterior se propone un ejemplo. Se tiene una señal medida afectada por ruido a la cual llamamos *Xorg*. Se conoce su versión original sin ruido *Xreal*. De *Xorg* se obtienen 2 rectificados que se etiquetan como *X1* y *X2*. Las 4 señales se muestran en la figura 2.17. Analizando los rectificados, se observa que ambas trazan una aproximación cercana a *Xreal*. No obstante, *X1* tiende a seguir levemente los valores que aparecen como extremos en *Xorg*, aunque en realidad todos los valores hallan sido afectados con el mismo tipo de ruido. Esto se puede ver claramente en el intervalo 6-11 (ver figura 2.17) donde se producen 2 valores (en *X1)* más bajos que sus precedentes y en el intervalo 66-71 donde se produce un valor más alto que sus precedentes. Por tanto, los errores entre *X1* (la aproximación que mantiene una dispersión más semejante a la señal medida) y *Xorg* siempre serán más pequeños que los errores entre *X2* (la aproximación que traza una tendencia menos oscilante respecto a la señal medida) y *Xorg*. Luego, si se computa la Diferencia Cuadrática Media (DCM) entre *Xorg* y cada una de las aproximaciones (*X1* y *X2*), se verá que la DCM será mayor para la señal que fluctúa menos respecto a *Xorg* (*X2*) y será menor para la señal que fluctúa más respecto a *Xorg* (*X1*). Si computamos un peso que sea función directa de las DCM entonces se dará más peso a *X2* y menos peso a *X1* y esto responde claramente a dar más peso a la aproximación que más se asemeja a *Xreal*.

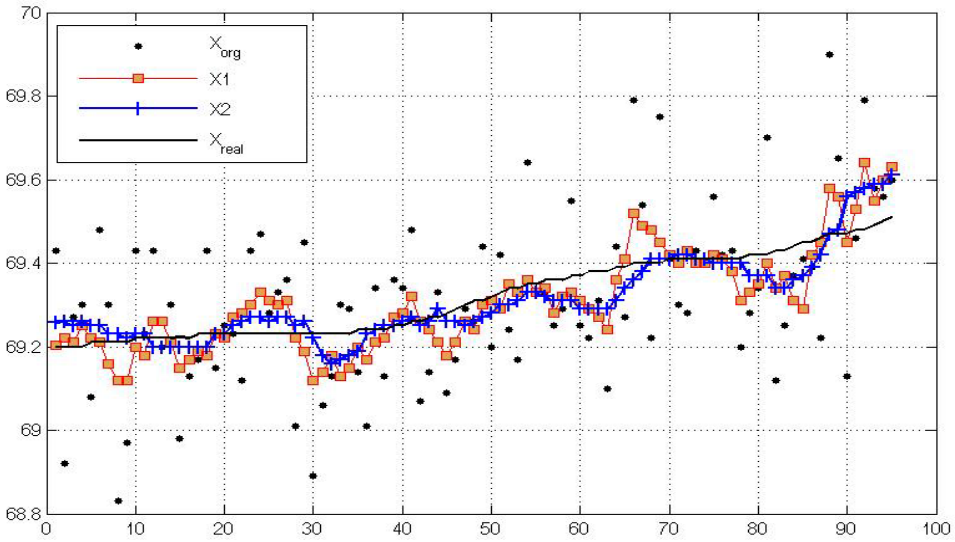

Figura 2.17. Esquema General de Previsión Combinada.

A partir de lo anterior se propone el siguiente procedimiento para estimar los pesos:

1. Se calcula la DCM entre la señal con ruido **y**, y cada uno de los estimados $\hat{\mathbf{y}}_i^*(k)$:

$$DCM_i = \frac{1}{n}\sum_{t=1}^{n}\left(\hat{\mathbf{y}}_i^*(k) - \mathbf{y}(k)\right)^2 \tag{2.23}$$

2. Se calculan los pesos como sigue:

$$pw_i = \frac{DCM_i}{\sum_{i=1}^{F} DCM_i} \tag{2.24}$$

3. Finalmente, se obtiene $\hat{\mathbf{y}}^*(k)$, según la ecuación (2.22).

2.5 Evaluación de la estrategia levashrink

En esta sección se presenta la evaluación de la metodología propuesta en la sección 2.3, incluyendo la comparación con otras metodologías de filtrado existentes. Para esta evaluación se organizaron una serie de experimentos similares a los descritos en la sección 2.2. A continuación se describen los aspectos básicos de la metodología propuesta:

- Se utilizan las mismas señales de la sección 2.2: (1) *Blocks* o *S1* desde 620 a 965; (2) *HeaviSine* o *S2* desde 251 hasta 820; (3) *Doppler* o *S3* desde 35 hasta 550. Todas las señales se contaminaron con ruido normal aleatorio con media 0 y varianza 0.5.

- Los métodos a comparar fueron los siguientes:
  - Filtro de media móvil (MM). Por experimentación preliminar se estableció que el tamaño de la ventana de datos optima (que produce el menor error de estimación) para cada una de las señales de estudio debía ser de 5 valores para *S2* y de 3 valores para *S1* y *S3*.
  - Filtro *EWMA*. También, por experimentación preliminar, se estableció que los valores óptimos del parámetro de suavización α (ver ecuación 2.4) para cada una de las señales de estudio debía ser de 0.5 para *S1*, 0.2 para *S2* y 0.4 para *S3*.

  - Método *waveshrink*. Se aplicó con ambos *Soft thresholding* y *Hard thresholding*. El parámetro β, para el *thresholding*, se fijó según las ecuaciones 2.19 y 2.20.
  - Método *levashrink*. Se aplicó con ambos *Soft thresholding* y *Hard thresholding*. El parámetro β, se fijó según las ecuaciones 2.19 y 2.20.

- Para los métodos basados en *wavelets* (*waveshrink* y *levashrink*):
  - Se adoptaron las siguientes *wavelets* Daubechies: Se escogieron la *db1*, *db2*, *db3*, *db4* y *db8*[b]. Adicionalmente, se seleccionó para todas las *wavelets* la corrección en los extremos según el método de simetrizado (Misiti *et al*., 2004). La corrección en los extremos permite evitar imprecisiones en los valores extremos, de los rectificados obtenidos por *wavelets*, que son los más importantes para aplicaciones en línea (Nounou y Bakshi, 1999).
  - Se aplicaron bajo un esquema en línea *OLMS*.
  - Para el caso del método *waveshrink* y de acuerdo a los valores que se utilizan en la literatura revisada, los valores de L_m que se utilizan, combinados con las distintas *dbN*, son L_m =4 y L_m=5.

Teniendo en cuenta todo lo anterior se llevaron a cabo las simulaciones de rectificación para cada señal. El número de simulaciones por cada método fue de 3 para el MM, 3 para el *EWMA*, 30 para el *waveshrink* y 30 para el *levashrink*. A continuación, se pasó a evaluar el error de estimación de cada filtro a lo largo de todo el horizonte de simulación. Para hacer esto, se tomó el vector de cada señal filtrada, $\hat{\mathbf{y}}(k)$, obtenida en cada simulación y el correspondiente vector de la señal original sin ruido, $\mathbf{y}(k)$, y se computó el error cuadrático medio entre ellas. Los ECM resultantes se resumen en las tablas 2.4, 2.5, 2.6 y 2.7 y en las figuras 2.18, 2.19, 2.20 y 2.21, donde se puede visualizar el rendimiento de las estrategias basadas en *wavelets*. Solo se muestran los gráficos de las curvas *S1* y *S2* y sobre un intervalo de datos menor al número original de datos procesados para cada curva (desde 1 hasta 110 para *S1*; desde 71 hasta 230 para *S2*). La razón de mostrar solo un pequeño intervalo de cada curva fue la de ayudar a una más fácil visualización de los resultados y a que el comportamiento de estos intervalos es más o menos fiel al comportamiento de los estimados sobre el resto de los intervalos no mostrados. Para las tablas, y en el caso de *waveshrink*, solo se muestra la combinación "*dbN* – L_m" con cada método de *thresholding* que condujo al mínimo ECM. De manera similar, para el caso *levashrink* solo se muestra el ECM de la mejor combinación de *dbN* con cada método de *thresholding*.

Tabla 2.4. ECM para los métodos *waveshrink*.

SEÑAL	Método thresholding	*dbN* utilizada	Nivel Aplicado	ECM
1	Soft	db3	L=5	0.1223
	Hard	db1	L=4	0.0632
2	Soft	db4	L=4	0.0586
	Hard	db1	L=4	0.0780
3	Soft	db3	L=4	0.0057
	Hard	db8	L=4	0.0044

Tabla 2.5. ECM para los métodos *levashrink*.

[b] Esta selección se basó en los resultados del análisis de la sección Análisis experimental de la aplicación de wavelets Daubechies en filtrado de datos.2.2 y en las *wavelets* que más se utilizan en la literatura revisada.

SEÑAL	Método thresholding	*dbN* utilizada	ECM
1	Soft	db4	0.1028
	Hard	db1	0.0724
2	Soft	db1	0.0660
	Hard	db1	0.0747
3	Soft	db4,db8	0.0043
	Hard	db8	0.0047

Tabla 2.6. ECM para los métodos *EWMA*.

SEÑAL	Valor de α	ECM
1	0.5	0.1378
2	0.2	0.0656
3	0.4	0.1301

Tabla 2.7. ECM para los métodos MM.

SEÑAL	Ventana	ECM
1	3	0.2171
2	5	0.0826
3	3	0.1300

En primer lugar se observa que, en general, los estimados mediante filtros *EWMA* y MM son menos precisos en términos de ECM que los estimados obtenidos con cualquiera de los métodos basados en *wavelets*. Esto concuerda con resultados previos de la literatura (Nounou y Bakshi, 1999; Köhler y Lorenz, 2004).

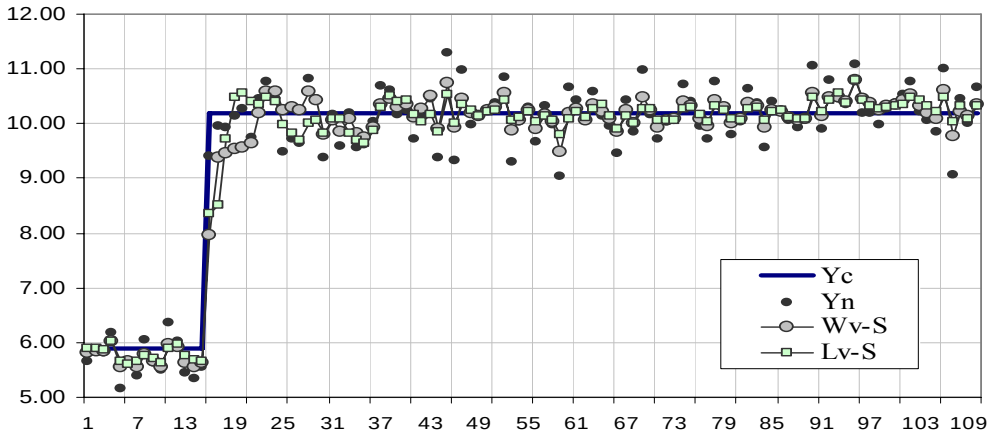

Figura 2.18. Comparación de métodos usando *Soft thresholding* (curva *S1*).

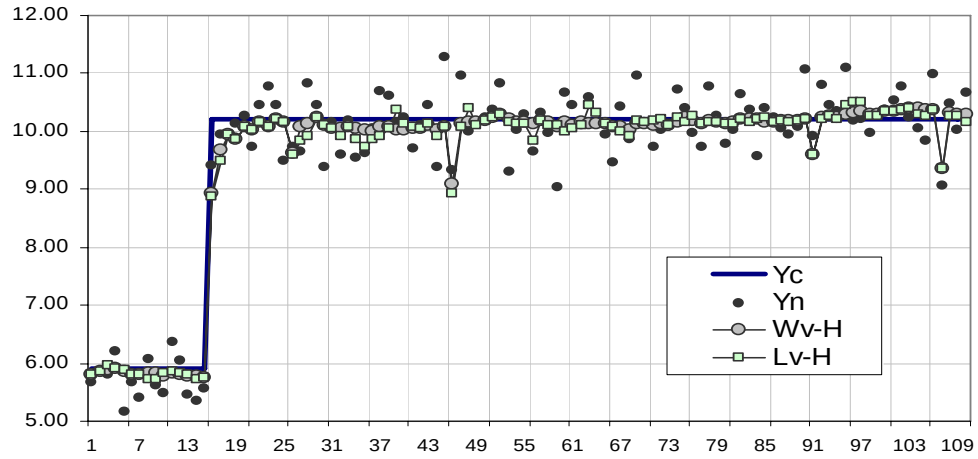

Figura 2.19. Comparación de métodos usando *Hard thresholding* (curva *S1*).

A continuación se consideran los resultados de las tablas 2.4 y 2.5 correspondientes a los casos *waveshrink* y *levashrink*, respectivamente.

Si solo se consideran las alternativas aplicadas con *thresholding* tipo *Soft* se observa que, globalmente, la estrategia *levashrink* conduce a mejores estimaciones que el método *waveshrink*. En efecto, para las curvas *S1* y *S3* el ECM obtenido es menor con el método *levashrink* que con *waveshrink*. También, en las figuras 2.18 y 2.20 se ven las comparaciones entre los filtrados de *S1* en el intervalo 1 a 110 y entre los filtrados de *S2* en el intervalo 71 a 230. Se observa que, para el caso de *S1* el filtrado con *levashrink* usando *Soft* (etiquetado como Lv-S) es levemente más preciso en el punto donde se produce la discontinuidad y, además, tiende a ser menos fluctuante en diversos puntos donde el estimado usando *waveshrink* (etiquetado como Wv-S) se presenta con pequeños picos. Esto significa que el *Lm* estimado continuamente se adapta al patrón de la señal De esta manera, se ve que estimando L_m, instante a instante, puede conducir a estimados de mejor calidad cuando la curva presenta patrones estacionarios (media más o menos constante) con cambios intercalados por discontinuidades. Si se explora el gráfico de *S2*, en el intervalo de 71 a 110 se ve que, aunque el ECM de *waveshrink* es menor (ver tablas 2.4 y 2.5) la diferencia entre ambos estimados es visualmente equivalente. Así, para casos de señales con curvas de pendientes suaves el método *levashrink* es equivalente a *waveshrink* con la ventaja comparativa de no tener que explorar con diversos valores de L_m al momento de la implementación inicial o ante cambios que se produzcan en el patrón de la señal. Ventaja que se hace más valiosa para situaciones en las que se debe trabajar con muchísimas señales.

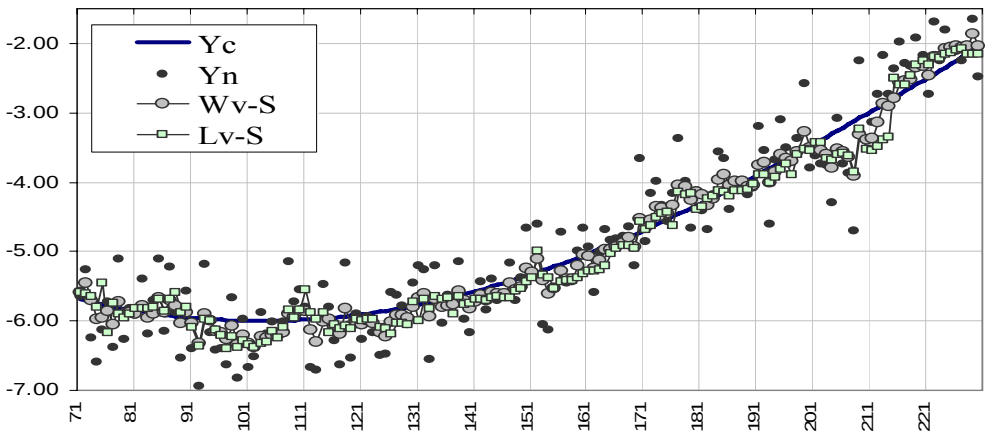

Figura 2.20. Comparación de métodos usando *Soft thresholding* (curva *S2*).

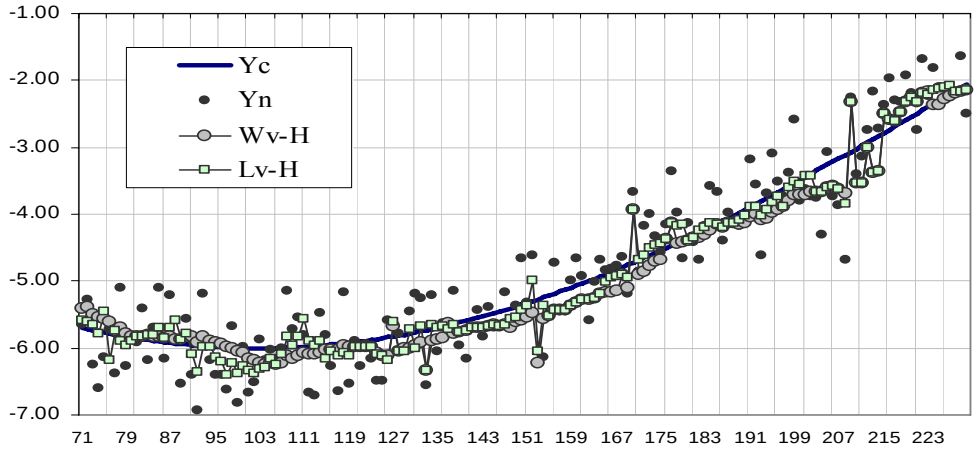

Figura 2.21. Comparación de métodos usando *Hard thresholding* (curva *S2*).

Si se considera la aplicación de los métodos usando *Hard thresholding*, se ve que, globalmente, la estrategia *waveshrink* conduce a estimaciones con ECM más bajos. No obstante, en este caso se observa que las diferencias entre los ECM de *waveshrink* y *levasrhink* son menores que las mismas diferencias para los casos de aplicación usando *Soft thresholding*. Así, explorando en detalle las figuras 2.19 y 2.21 vemos que los comportamientos de ambos estimados (Wv-H para *waveshrink* y Lv-H para *levashrink*) son muy similares. Por tanto, la aplicación de Lv-H es equivalente a aplicar Wv-H, pero con la ventaja de la estimación continua de L_m.

Por otro lado, si se comparan entre si a las figuras 2.18 y 2.19 y a las figuras 2.20 y 2.21, se ve que los estimados de la curva, en muchísimos intervalos, son más suaves usando *Hard* que usando *Soft*. Esta observación también se resaltó en un trabajo precedente (Nounou et al, 1999) y, consecuentemente, en ese trabajo se propuso como más apropiado el uso de esquemas de filtración basados en *wavelets* que utilicen *Hard thresholding*. No obstante, si se sigue mirando el detalle de las figuras 2.18 a 2.21, se observa que en varios puntos del estimado basado en *Hard* se producen picos bastantes pronunciados que distorsionan el estimado resultante. En vista de ello, se concluye que los esquemas basados en *wavelets* conducirán a mejores estimados si se utiliza el *Soft* como método para el *thresholding*.

Los resultados anteriores indican que la estrategia de estimar un L_m óptimo integrado con un método de rectificación basado en *wavelets* es equivalente y, en algunos casos, levemente mejor que aplicar la rectificación basada en *wavelets* cuando se usa un L_m que se fija mediante ensayo y error o por experiencia. Si se traslada la conclusión anterior a una situación de análisis de datos fuera de línea (por ejemplo, utilizar los métodos para rectificar datos obtenidos en un experimento de laboratorio pasado) la ventaja real de *levashrink* es ciertamente pequeña. No obstante, si el análisis involucra el tratamiento de muchísimas curvas, la estrategia de estimar L_m automáticamente puede conducir a una sustancial ganancia en tiempo. Si se traslada lo anterior al caso de aplicaciones en línea, la estrategia *levashrink* cobra aun más relevancia debido a que como se ha visto, en cada instante la decisión de seleccionar un mejor L_m puede mejorar la estimación continuamente. En el capítulo 4 se explora el impacto de esta mejora cuando se integra la estrategia *levashrink* a un sistema de monitorización.

Es importante destacar que los resultados mostrados en la literatura normalmente se refieren a datos analizados fuera de línea y que por tanto se procesan en un solo lote y no continuamente como fue el caso de los análisis anteriores. Así, los análisis presentados cobran un mayor valor de cara a la posibilidad de ser implementados en situaciones reales donde se requiere un estimado de la señal tan pronto como se produce y que es el caso de las aplicaciones mostradas bajo el esquema de aplicación *OLMS*.

2.6 Evaluación de la estrategia de Rectificación Combinada

En este apartado se evalúa la aplicación de la estrategia de rectificación por combinación de *dbN*. Para ello se hicieron una serie de simulaciones como sigue:

- Se evaluaron 2 combinaciones: una utilizando 2 *dbN* y otra utilizando 3 *dbN*.
- Al igual que en la evaluación de *levashrink*, se utilizaron las siguientes *wavelets*: *db1*, *db2*, *db3*, *db4* y *db8*.
- Para todos los casos, uno de los filtros fue siempre la *db1*. Esto obedece a los resultados mostrados en las secciones 2.4 y 2.7, donde se vio que la *db1* era adecuada para varios tipos de señales.
- Para los casos de 2 filtros, los pares evaluados fueron *db1-db2* (12), *db1-db3* (13), *db1-db4* (14) y *db1-db8* (18). Para los casos de 3 filtros, los tríos evaluados fueron *db1-db2-db8* (128), *db1-db3-db8* (138) y *db1-db4-db8* (148).
- Los filtros, individualmente, se aplicaron según la estrategia *levashrink* y bajo esquemas *OLMS* con *Soft thresholding*.
- Las señales rectificadas fueron las mismas que las del caso *levashrink*.
- La estimación combinada se aplicó según se propuso en la sección 2.4.

Teniendo en cuenta todo lo anterior se llevaron a cabo simulaciones de cada señal. A continuación, se pasó a evaluar el error de estimación de cada combinado a lo largo de todo el horizonte de simulación. Para hacer esto, se tomó el vector de cada señal filtrada, $\hat{\mathbf{y}}^*(k)$, obtenida en cada simulación y el correspondiente vector de la señal original sin ruido, $\mathbf{y}^*(k)$, y se computó el ECM entre ellas. Los resultados (ECM resultantes, junto con los ECM de los filtros usados individualmente) se resumen en la tabla 2.8 y en las figuras 2.22, 2.23 y 2.24.

Tabla 2.8. ECM para los métodos con combinación de *dbN*.

	db1	db2	db3	db4	db8	12	13	14	18	128	138	148
S1	0.1479	0.1159	0.1620	0.1068	0.1181	0.1239	0.1047	0.1125	0.1200	0.1146	0.0978	0.1091
S2	0.0664	0.1037	0.1334	0.0814	0.0931	0.0674	0.0696	0.0604	0.0653	0.0694	0.0694	0.0641
S3	0.0063	0.0050	0.0061	0.0043	0.0043	0.0051	0.0050	0.0050	0.0049	0.0046	0.0045	0.0045

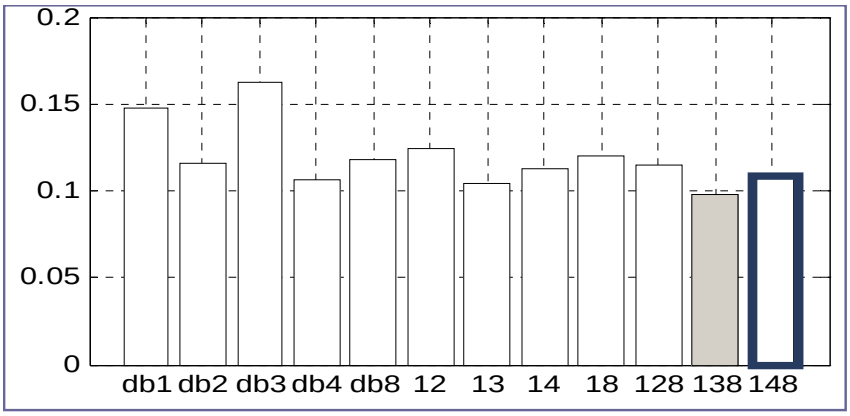

Figura 2.22. ECM de filtrados por combinación (curva *S1*).

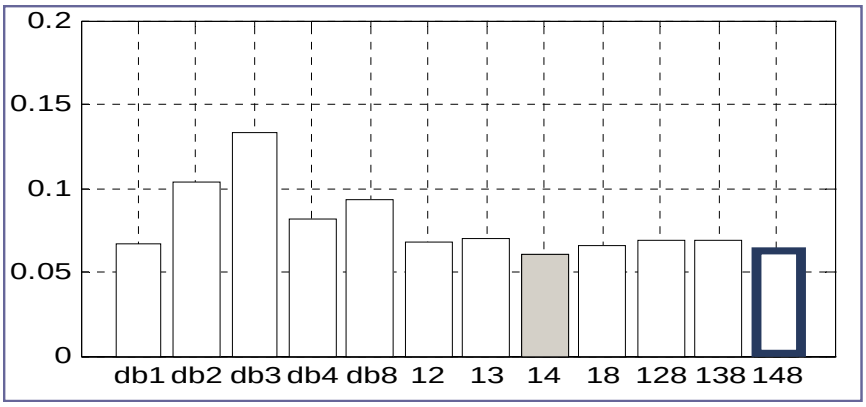

Figura 2.23. ECM de filtrados por combinación (curva *S2*).

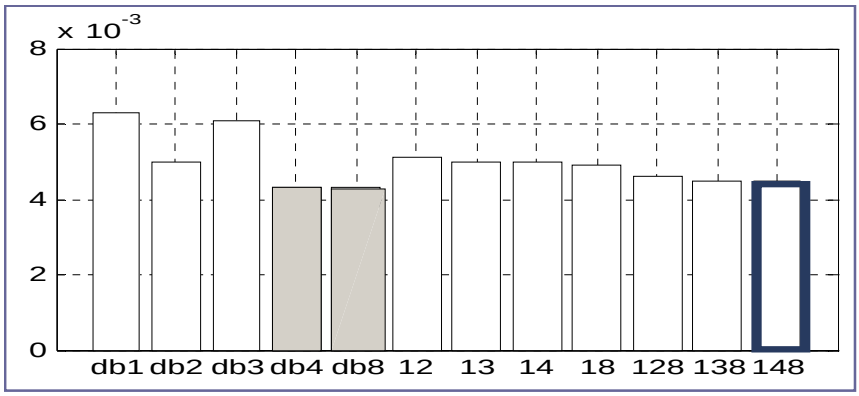

Figura 2.24. ECM de filtrados por combinación (curva *S3*).

Se observa que los métodos combinados dan lugar a mejores estimaciones para los casos de las curvas *S1* y *S2*, mientras que en *S3* la mejor estimación corresponde al uso de filtros individuales. También se observa que, pese a que no hay una alternativa que mejore las estimaciones para todas las señales, la opción de filtrado que combina "*db1-db4-db8*" muestra ECM significativamente bajos en todos los casos, tanto cuando se compara a los ECM de los filtros individuales como cuando se compara a los ECM de las diferentes combinaciones (ver las figuras 2.22, 2.23 y 2.24). De este modo, y como una primera alternativa de filtrado combinado, la combinación "*db1-db4-db8*" acompañada de la aplicación del *levashrink* con *Soft thresholding*, podría ser utilizada como una alternativa para filtrar datos en línea con la ventaja de evitarse la selección de la *dbN* y la selección del parámetro *Lm*, dando lugar a una

estrategia de rectificación aplicable a un rango amplio de patrones de curvas y autónoma y fiable para propósitos de aplicaciones de filtración en línea.

2.7 Conclusiones

En este capítulo se han presentado y discutido una serie de estrategias de filtración de datos que buscan mejorar el rendimiento de los métodos existentes de rectificación de datos. Los resultados que se han obtenido a partir de pruebas sobre señales estándar que combinan diferentes patrones de señales son coherentes. Esto permite generalizar las conclusiones, en la medida que es posible plantear generalizaciones a problemas de esta naturaleza.

En una primera propuesta, se discute una estrategia que aplica un paso previo de determinación del nivel optimo de descomposición de una señal mediante *wavelets* y, luego, utiliza esta información como entrada a un método de filtrado basado en *wavelets*. Se vio que esta estrategia, aplicada en combinación con el método *waveshrink*, conduce a estimaciones de señales equivalentes y, en algunos casos levemente mejores, a las estimaciones obtenidas mediante *waveshrink* pero fijando el nivel de descomposición manualmente y por ensayo y error. Asimismo, de acuerdo a la forma en que se realizaron los experimentos, los resultados obtenidos aportaron evidencia de la validez del método para aplicaciones en línea donde se requieren estimaciones de las señales tan pronto como se produce un valor medido en el proceso (monitorización y detección de fallos, control *feedback*, optimización en tiempo real). Adicionalmente, se mostró que el método produce mejores estimados que los obtenidos con el uso de filtros tradicionales como el *EWMA* o el filtro de media móvil.

En una segunda propuesta, se ha planteado explotar de forma combinada las ventajas de diferentes *dbN* para rectificación. Se propuso combinar los filtrados individuales de diferentes *dbN*, con objeto de asegurar una estrategia de rectificación adaptable a un número diverso de patrones de señales que a la vez mantuviera el mínimo de calidad en la estimación. Se llegó a la conclusión de que se puede proponer una estrategia de rectificación, basada en una adecuada combinación de *db1-db4* y *db8*, que evita la tarea de evaluar diferentes *dbN* en el momento de iniciar la implementación de una estrategia de filtración basada en *wavelets* (o incluso tras un cambio observado en el comportamiento del proceso) y que a la vez asegura estimados que compiten en calidad con estimaciones basadas en el uso de una sola *dbN*. Cabe destacar que todos estos resultados apuntan a que todavía existe un gran potencial de mejora en los métodos de filtrado si se llegan a explotar adecuadamente variantes de la estrategia de combinación propuesta (combinación de diferentes funciones *wavelets*, diferentes patrones, diferentes filtros,…, etc.).

NOMENCLATURA

$\mathbf{A}_L(k)$	Componente de aproximación a una escala $l=L$.
$a_{L,v}$	Coeficiente de escala o aproximación a la resolución L y traslación v.
b	Pesos que definen las características de los filtros lineales.
$\mathbf{D}_l(k)$	Componente de aproximación a una escala $l \leq L$.
$d_{l,v}$	Coeficientes de detalles a una resolución l y traslación v.
$d^*_{l,v}$	Coeficientes de detalles que quedan tras el *thresholding*.
F	Número de filtros utilizados para la rectificación combinada.
k	Tiempo de muestreo.
L	Escala de mayor resolución o nivel de descomposición.

L_m Nivel de descomposición optimo.
n Número de observaciones en $\mathbf{y}(k)$.
nc Longitud de la ventana de datos para filtración .
nc_w Longitud de la ventana de datos para cálculo de ECM Local
nt La mitad de la longitud de la ventana del filtro mediano (nc/2).
P_L Potencia contenida en los detalles hasta la escala L.
ΔP Variación de Potencia entre escalas.
pw_i Pesos asignados a cada estimado individual en la rectificación combinada.
$\boldsymbol{V}_l$ Subespacio de aproximación a la escala l.
$\mathbf{W}_l$ Subespacio de los detalles a la escala l.
wi Ventana de datos para el cálculo del ECM local
$\mathbf{y}(k)$ Variable de proceso o señal medida (mediciones afectadas por ruido).
$\hat{\mathbf{y}}^*(k)$ Estimado de la variable de proceso real.
$\mathbf{y}^*(k)$ Variable de proceso o señal real (no contaminada con ruidos).
$\hat{\mathbf{y}}i(k+h)$ Estimado de $\mathbf{y}$ parcial en h instantes futuros para la rectificación combinada.
$\hat{\mathbf{y}}(k+h)$ Estimado final en h instantes futuros para la rectificación combinada.
$\mathbb{R}$ Conjunto de los números reales.
$\mathbb{Z}$ Conjunto de los números enteros.

LETRAS GRIEGAS
α Constante de suavización del *EWMA*.
β Valor umbral para el paso de *thresholding* del método *waveshrink*.
$\varepsilon(k)$ Error aleatorio que se añade a $\mathbf{y}(k)$ (ruido).
σ_l Desviación estándar del ruido a la escala l.
$\psi_{l,v}(k)$ Función *wavelets* a una resolución l y traslación v.
$\phi_{l,v}(k)$ Función de escala a una resolución l y traslación v.

SUPERÍNDICES
^{H} *Hard thresholding*.
^{S} *Soft thresholding*.

SUBÍNDICES
j j-ésimo coeficiente.
l Resolución o escala.
v Factor de traslación.
h Horizonte de tiempo para previsión combinada.

ACRÓNIMOS
dbN Funciones *wavelets* pertenecientes a la familia Daubechies.
DCM Diferencia cuadrática media definida dentro del esquema Rectificación Combinada.
ECM Error cuadrático medio.
OLMS *On Line Multiscale Filtering* o Filtración multiescala en línea.
BCTI *Boundary Corrected Translation Invariant* o Filtración multiescala con corrección de los límites invariante en la traslación.
EWMA *Exponentially Weighted Moving Average* o filtro de suavización exponencial.
PreviCom Previsión Combinada.
FIR *Finite Impulse Response Filters*.
IIR Infinite Impulse Response Filters.

FHM Filtro Hibrido basado en la Mediana.

CAPÍTULO 3. RECONCILIACIÓN DE DATOS DE SISTEMAS DINÁMICOS LINEALES INTEGRADA CON FILTRACIÓN BASADA EN WAVELETS

RESUMEN

En el capítulo precedente se analizaron estrategias de filtrado para casos donde no se dispone de un modelo del proceso. Cuando se dispone de un modelo, la reconciliación de datos ofrece una alternativa de rectificación que, además de aportar filtrados de calidad, aseguran la consistencia con el modelo del proceso y la posibilidad de un análisis conjunto de las variables potencialmente útil para monitorización, diseño y análisis de sensores, etc. En este capítulo se exploran estrategias de rectificación que combinan la capacidad de filtrado mediante *wavelets* con la consistencia de los estimados aportada por una técnica de Reconciliación. Las estrategias se desarrollan para casos de plantas con modelos lineales y para el caso de la reconciliación dinámica que es uno de los retos principales de investigación y aplicación en el área de la reconciliación de datos. Las estrategias se proponen como métodos en 2 etapas: un paso preliminar de estimación de las tendencias de las variables del proceso y un paso final de reconciliación de las tendencias. Para la estimación inicial de las tendencias se adopta alguna de las estrategias descritas en el capítulo precedente. Adicionalmente, se proponen y evalúan varios esquemas de horizonte móvil que permiten una actualización recursiva de la varianza y la aplicabilidad de la estrategia propuesta para casos en línea. Se propone el caso de simulación de la operación de un reactor continuo sobre el que se evalúa el rendimiento de las propuestas frente a otras estrategias existentes. Adicionalmente, se considera la extensión de las estrategias para casos no lineales. Los resultados muestran que las estrategias propuestas son competitivas frente a los métodos actuales, en términos de precisión de las estimaciones, reducción de la variabilidad y factibilidad de uso en aplicaciones en línea.

3.1 Introducción

Pese a la eficiencia de los filtros uní variables (ver capítulo 2), los esfuerzos de investigación y aplicación de estrategias de rectificación de datos en la Industria Química y de Procesos (IQP) han puesto un gran énfasis en el uso de estrategias, basadas en modelos, conocidas bajo el nombre de Reconciliación de Datos (RD). La razón de este esfuerzo estriba en el postulado de que, si se dispone de un modelo fiable de la planta, los rectificados a obtener deberían ser consistentes (cumplir) con los balances de masa, energía, etc., y los filtros uní variables no garantizan esta consistencia. En este capítulo la discusión y desarrollos se orientarán hacia las estrategias RD aunque, como se verá más adelante, se intentará tomar ventaja de la filtración unívariable mediante técnicas *wavelets* para mejorar la calidad, en términos de precisión, de los reconciliados obtenidos con algunas de las estrategias actuales de reconciliación.

La RD se enfoca a la eliminación de errores aleatorios, bajo la suposición de que los datos no se ven afectados por errores gruesos (Narasimhan y Jordache, 2000; Romagnoli y Sánchez, 2000), y se propone como la tarea de estimar los valores verdaderos de las variables de proceso (o ajustar los datos medidos), de forma que cumplan con las leyes naturales que rigen sobre el proceso, es decir, los balances de masa, balances de energía, …, etc. Los ajustes se hacen tomando ventaja de la redundancia temporal y funcional de las mediciones. Se habla de redundancia funcional cuando el número de mediciones disponibles es mayor al número de

mediciones libres de errores que se requieren para calcular todos los parámetros y variables de un sistema. Por otro lado, se habla de redundancia temporal, cuando se dispone de mediciones pasadas de manera que se puedan utilizar para propósitos de estimación a través del correspondiente modelo dinámico del proceso.

En su forma más común y clásica, la RD se formula como un problema de optimización de mínimos cuadrados ponderados y con restricciones (Narasimhan y Jordache, 2000; Romagnoli y Sánchez, 2000). Las restricciones las forman el conjunto de ecuaciones del modelo del proceso, más posibles restricciones de desigualdad asociadas a límites de los valores de las variables de proceso. Los pesos generalmente vienen representados por el inverso de las varianzas de los errores de medición. La solución del problema de optimización proporciona los estimados óptimos de las mediciones de las variables del proceso. La RD en estado estacionario se ha estudiado durante varias décadas (ver sección 1.3.1.1).

Pese a que en muchos casos los procesos se operan para intentar mantener ciertas condiciones nominales de estado estacionario, los procesos químicos son de naturaleza dinámica. Aun en los casos donde las condiciones del proceso son habitualmente muy cercanas al estado estacionario (*quasi steady state* o *QSS*), se tienden a experimentar frecuentes variaciones alrededor de estas condiciones deseadas. En consecuencia, el uso de modelos dinámicos permitirá una mejor representación del comportamiento del proceso que el uso de modelos en estado estacionario. Bajo esta óptica, diversos autores han propuesto (Liebman *et al*., 1992; Bagajewicz y Jiang, 1997) que el uso de procedimientos RD para casos dinámicos, esto es, un modelo dinámico como restricción, conduciría a mejores estimaciones, aun para casos de procesos reales operando en un *QSS*.

La reconciliación de datos dinámicos (RDD) adopta una formulación similar a la RD en estado estacionario (ver sección 1.3.1.1), esto es, como un problema de optimización donde el objetivo es minimizar la desviación entre los valores medidos de las variables de proceso $\mathbf{y}(k)$ y sus valores estimados $\hat{\mathbf{y}}(k)$. Esta desviación se pondera en la función objetivo por el inverso de la varianza de los errores de medición. Asimismo, la minimización esta sujeta al modelo dinámico, según se expresa en las ecuaciones 3.1, 3.2, 3.3 y 3.4.

$$min \sum_{k=0}^{c} [\hat{\mathbf{y}}(k) - \mathbf{y}(k)]^T \mathbf{Q}^{-1} [\hat{\mathbf{y}}(k) - \mathbf{y}(k)] \tag{3.1}$$

Sujeto a:

$$\frac{d\mathbf{y}(k)}{dk} = \mathbf{f}\left(\hat{\mathbf{y}}(k), \hat{\mathbf{u}}(k), \hat{\theta}(k)\right) \tag{3.2}$$

$$0 = \mathbf{h}\left(\hat{\mathbf{y}}(k), \hat{\mathbf{u}}(k), \hat{\theta}(k)\right) \tag{3.3}$$

$$0 < \mathbf{g}\left(\hat{\mathbf{y}}(k), \hat{\mathbf{u}}(k), \hat{\theta}(k)\right) \tag{3.4}$$

Donde el parámetro *c* representa el tiempo actual, *k* representa el tiempo de muestreo, **Q** representa la matriz covarianza conteniendo las varianzas de todas las variables, **f**() representa un conjunto de ecuaciones diferenciales, **h**() representa un conjunto de ecuaciones algebraicas, **g**() representa un conjunto de restricciones de desigualdad[a], $\hat{\mathbf{u}}$ son los estimados de las variables de entrada y $\hat{\theta}$ representa estimados de los parámetros del modelo. A pesar de

[a] Estas pueden incluir, además de relaciones asociadas al modelo, límites de los valores de las variables.

que la RDD proporciona estimados precisos, la solución del problema de optimización involucra varias dificultades importantes:

- Para el manejo de las ecuaciones diferenciales, en la literatura actual se recurre a transformarlas a un conjunto equivalente de ecuaciones algebraicas. Esto se alcanza a través del uso de métodos de discretización tales como la aproximación de Euler (Rollins y Devanathan, 1993), Runge-Kutta implícito (Albuquerque y Biegler, 1996) o colocación ortogonal (Liebman *et al*., 1992). La eficiencia de estos métodos basados en discretización descansa en la selección de un tiempo de integración suficientemente alto. En la práctica, los tiempos de muestreo requeridos para la recolección de datos medidos es usualmente mayor al tiempo de integración, lo que puede llegar a limitar el uso de estas técnicas de discretización para propósitos de RDD (Albuquerque y Biegler, 1996). Los métodos de discretización pueden proveer ecuaciones asociadas a las salidas que produzcan datos a prácticamente cualquier intervalo. Esto último podría se ventajoso para ganar en redundancia pero como resaltan Bagajewicz y Jiang también podría crear fluctuaciones inexistentes en los valores reconciliados que podrían acarrear inconformidades en la gestión operativa y económica diaria de la planta (Bagajewicz y Jiang, 1997). Adicionalmente, la discretización puede llegar a incrementar el número de variables y ecuaciones en el problema de optimización y esto, según el estado del arte de los sistemas informáticos actuales, podría ser perjudicial en términos de esfuerzo computacional y para aplicaciones en tiempo real.
- La información sobre las varianzas del proceso es un prerrequisito importante para la aplicación de la RDD. En efecto, como se puede ver en la ecuación 3.1, la función objetivo esta ponderada por el inverso de la matriz covarianza **Q**. La evaluación de las varianzas de las variables en sistemas dinámicos es un problema relativamente complejo ya que dichas varianzas se ven afectadas por variabilidades tanto del proceso como de las mediciones. Es común encontrar que estas varianzas se asuman constantes (Liebman *et al*., 1992). No obstante, en muchas situaciones reales esta aproximación podría no ser valida con el consecuente efecto sobre la precisión de los reconciliados. Todo esto desemboca en la necesidad de explorar estrategias más apropiadas para el manejo de la varianza.
- Si el modelo del proceso presenta no linealidades, la obtención de la solución del problema RDD es mucho más compleja que en el caso lineal. En la literatura existente se ha establecido claramente que los problemas RD no lineales son más difíciles de resolver que los asociados a problemas de RD lineal (Narasimhan y Jordache, 2000). Para la resolución de la optimización asociada a las propuestas RD no lineal que se encuentran en la literatura, se han utilizado métodos como la linealización sucesiva o la Programación No Lineal (PNL), entre otros (Liebman *et al*., 1992; Tjoa y Biegler, 1992; Barbosa *et al*., 2000). En el caso de la linealización sucesiva, la idea base es la linealización de las restricciones no lineales como series de expansión de Taylor alrededor de los estimados actuales. La estrategia es relativamente sencilla y rápida en términos computacionales. No obstante, no permite un fácil manejo de límites en variables como por ejemplo los asociados a minimizar los efectos de errores gruesos (Liebman *et al*., 1992). En el caso del uso de PNL, ésta permite el manejo de funciones objetivos no lineales generales y el manejo explicito de restricciones no lineales (igualdades o desigualdades) así como límites asociados a las variables. También, Liebman y Edgar (1992) demostraron que, comparada a la linealización sucesiva, una solución con PNL era considerablemente superior en cuanto a calidad de los estimados. Asimismo, existen varios *solvers* que permiten la resolución de este tipo de problemas (por ejemplo, GAMS/MINOS). En términos de precisión, los

resultados en la literatura usando PNL son bastante buenos. Pese a esto, el esfuerzo computacional asociado, aún bajo el estado del arte actual en ordenadores, es considerablemente alto por lo que su utilización para aplicaciones en línea queda seriamente limitada.

Las propuestas que se discuten en las secciones que siguen intentan dar respuestas alternativas a los problemas descritos anteriormente.

3.2 Estrategia Polinomial Ampliada (EPA)

La Estrategia Polinomial Ampliada (EPA) es una extensión de la estrategia integral (EIn), presentada por Bagajewicz y Jiang (1997). No obstante, en EPA se intentan resolver varios problemas asociados a la EIn, y también a las estrategias RDD en general, como la solución numérica y la identificación del grado de los polinomios asociados a la formulación en EIn, el calculo dinámico de la varianza y la posibilidad de uso de la estrategia resultante en línea. La estrategia se orienta a la reconciliación de datos de sistemas lineales y dinámicos como los asociados a la información del procesamiento de materiales (flujos másicos de entrada salida, volúmenes en tanques, etc.) o similares. En las secciones que siguen se describe en detalle la estrategia EPA.

3.2.1 Representación Polinómica del Modelo de Proceso

El balance dinámico asociado a los materiales de un proceso se puede representar, de forma genérica, mediante un sistema de ecuaciones diferenciales y algebraicas como sigue:

$$\frac{d\hat{\mathbf{x}}(k)}{dk} = \mathbf{C1}.\hat{\mathbf{y}}(k) \tag{3.5}$$

$$\mathbf{C2}.\hat{\mathbf{y}}(k) = 0 \tag{3.6}$$

Luego, las ecuaciones anteriores se pueden combinar, lo que da lugar al siguiente modelo:

$$\frac{d\hat{\mathbf{x}}(k)}{dt} = \mathbf{C}.\hat{\mathbf{y}}(k) \tag{3.7}$$

con

$$\mathbf{C} = \begin{bmatrix} \mathbf{C1} \\ \mathbf{C2} \end{bmatrix} \tag{3.8}$$

Donde $\hat{\mathbf{x}}(k)$ representa a los estimados de las variables de estado (habitualmente masas) y $\hat{\mathbf{y}}(k)$ representa los estimados de las variables de entrada/salida (habitualmente flujos másicos). ***C***, ***C1*** y ***C2*** son matrices del sistema. Adicionalmente, si se representa al término derivativo de las variables de estado como una variable ficticia de proceso llamada $\hat{\mathbf{w}}(k)$, el modelo en la ecuación (3.7) se puede rescribir como:

$$\hat{\mathbf{w}}(k) = \mathbf{C}.\hat{\mathbf{y}}(k) \tag{3.9}$$

El modelo resultante (ecuación 3.9) sustituye al modelo inicial representado por las ecuaciones (3.2) y (3.3) en la formulación de la RDD (ver sección 3.1). La i-ésima variable $\mathbf{y}_i(k)$ y la j-ésima variable $\mathbf{w}_j(k)$ se pueden ajustar con el tiempo mediante un polinomio de grado *p* como sigue:

$$\hat{\mathbf{y}}_i(k) = \sum_{r=0}^{p_i} \left(\hat{\alpha}_{i,r} \cdot k^r \right), \quad \forall i = 1,...,I \tag{3.10}$$

$$\hat{\mathbf{w}}_j(k) = \sum_{r=0}^{p_j} \left(\hat{\xi}_{j,r} \cdot k^r \right), \quad \forall j = 1,...,J \tag{3.11}$$

Donde I y p_i representan el número de variables de proceso y el grado del polinomio ajustado a la i-ésima variable de proceso respectivamente. También, J y p_j representan el número de variables de estado y el grado del polinomio ajustado a la j-ésima variable de estado respectivamente. Si se sustituyen las ecuaciones (3.10) y (3.11) en la ecuación (3.9) y se opera sobre la ecuación resultante, se puede llegar a una nueva expresión para el modelo de la ecuación (3.9) como sigue:

$$\hat{\xi}_{j,r} = \sum_{i=0}^{I} C_{j,i} \cdot \hat{\alpha}_{i,r}, \quad \forall j \tag{3.12}$$

o:

$$\boldsymbol{\Xi}^{JxR} = \mathbf{C} \cdot \boldsymbol{\Lambda}^{JxR} \tag{3.13}$$

Donde cada fila j en la matriz $\boldsymbol{\Xi}$ contiene los coeficientes de la variable $\hat{\mathbf{w}}_j(k)$ en orden descendente de tiempo, mientras que cada fila i en la matriz $\boldsymbol{\Lambda}$ contiene los coeficientes de la variable $\hat{\mathbf{y}}_i(k)$ en orden de tiempo descendente. Para determinar el número de columnas de las matrices $\boldsymbol{\Xi}$ y $\boldsymbol{\Lambda}$ se selecciona el grado del polinomio más alto entre las variables involucradas ($p^{\max}$) y luego, se añaden términos nulos a cada polinomio de cada variable de modo de igualarse todos a $p^{\max}$. También, la representación obtenida a partir de polinomios permite combinar las redundancias funcional y temporal de las mediciones. Finalmente, la sustitución de variables afecta a la función objetivo (ver secciones 3.2.5 y 3.3).

3.2.2 Reconciliación sobre un horizonte móvil

Idealmente, un esquema RD utilizaría todas las mediciones del proceso desde el arranque del mismo k_0 y hasta el tiempo de operación actual para intentar aprovechar al máximo la redundancia temporal c o k_c. Desafortunadamente, tales esquemas en aplicaciones reales y en línea conducirían en el tiempo a problemas de sobrecarga de datos, lo que afectaría la capacidad de manejo de y la velocidad de procesamiento de cualquier *solver* actual. Luego, para aplicaciones reales y en línea es mejor usar un enfoque de ventana (horizonte de datos) móvil.

Este último tipo de enfoques ya se ha usado en trabajos precedentes de reconciliación dinámica (Liebman *et al*., 1992; Jang *et al*., 1996; Robertson *et al*., 1996; Mingfang *et al*., 2000). En alguno de estos trabajos (Mingfang *et al*., 2000) cada variable de procesos se reconcilia *nc* veces, tantas como las muestras correspondientes de cada variable intervengan en reconciliaciones sucesivas. No obstante, en los trabajos anteriores no queda claro si la ventana, a cada instante k_c (c), se debería construir con las *nc*-1 estimaciones recientes más la nueva medición (Liebman *et al*., 1992) o con solo las *nc* mediciones pasadas sin reconciliar (Mingfang *et al*., 2000). En una fase inicial del trabajo de este capítulo (Benqlilou *et al*., 2001), EPA se aplicó utilizando solo las *nc* mediciones pasadas. Aquí, se consideran 2 alternativas para la aplicación del EPA. Ambas se describen a continuación.

- Reconciliación uno a uno (caso OBOR): Bajo este esquema, en el primer instante k_c se construye una ventana con las mediciones disponibles más recientes, desde $k_{c\text{-}nc+1}$

hasta k_c. Sobre esta ventana se aplica EPA. El vector de reconciliados que se obtiene (desde k_{c-nc+1} hasta k_c) se guarda como el conjunto reconciliado actual. En el siguiente instante k_c, se actualiza la ventana de datos eliminando el último valor de la ventana, y añadiendo el nuevo valor medido en k_c. Sobre esta ventana actualizada se aplica nuevamente EPA. Los sucesivos reconciliados que se guardan en cada instante corresponden al correspondiente k_c. En la figura 3.1 se ilustra este esquema.

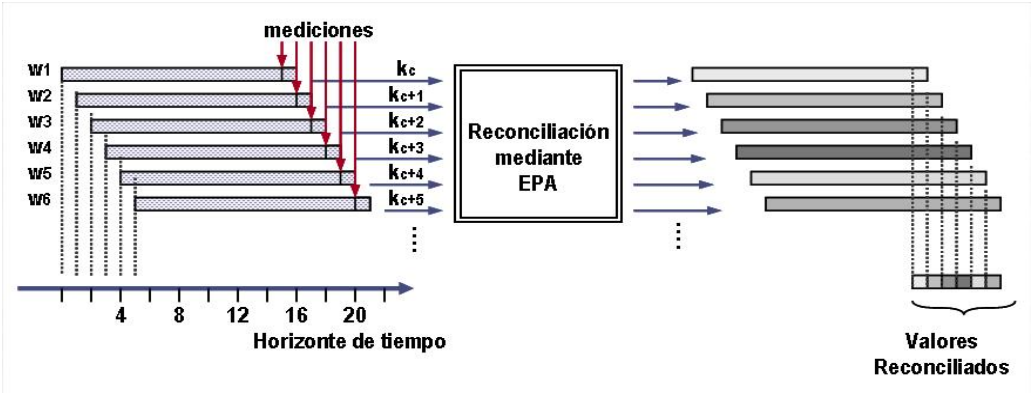

Figura 3.1. Esquema OBOR para Reconciliación con EPA.

- Reconciliación Corregida (caso CR): La aplicación inicial de este caso es similar a la del caso OBOR. Esto es, en el primer instante k_c, se construye una ventana con las mediciones disponibles más recientes desde k_{c-nc+1} hasta k_c. Sobre esta ventana se aplica EPA. El vector de reconciliados obtenidos (desde k_{c-nc+1} hasta k_c) se guarda como el conjunto reconciliados actual. En los siguientes instantes k_c, se actualiza la ventana de datos de la siguiente manera: se toman los *nc-1* últimos reconciliados y se le añade el nuevo valor medido del proceso. Sobre esta ventana se aplica EPA. El vector de reconciliados obtenidos (desde k_{c-nc+1} hasta k_c) se guarda como el conjunto reconciliados actual lo que significa que el reconciliado se corrige durante *nc* reconciliaciones sucesivas. En la figura 3.2 se ilustra este esquema.

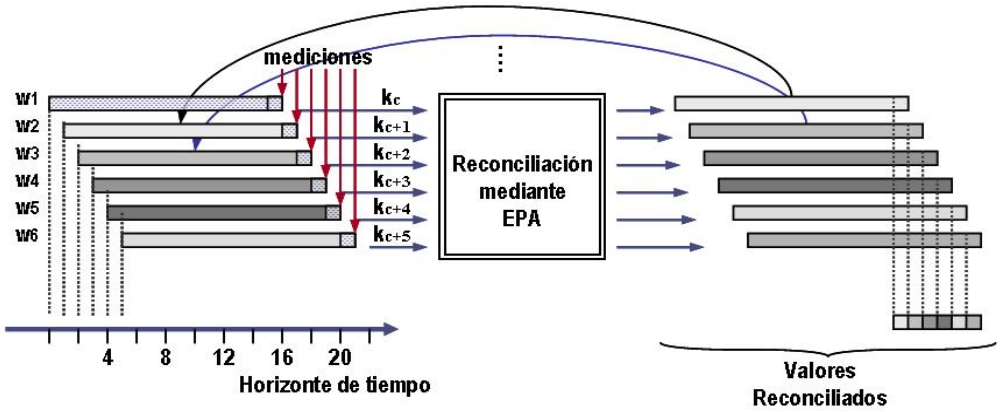

Figura 3.2. Esquema CR para Reconciliación con EPA.

La estrategia EPA bajo ambos esquemas se evalúa en una sección posterior (sección 3.4.1) tal que se pueda ver el impacto de uno y otro sobre la reconciliación.

3.2.3 Grado del polinomio para las variables del proceso

La identificación del grado del polinomio que ajusta a cada variable del proceso es el paso previo para obtener la representación según la ecuación (3.13). Una primera alternativa para acometer la identificación sería como sigue: Dada una ventana de datos de una longitud adecuada, la tarea de identificación se debe aplicar en cada instante y mediante un procedimiento recursivo que consiste en seleccionar un grado inicial 1 para p, ajustar los datos según p y luego incrementar el valor de p hasta que el error entre la salida del polinomio y la variable de proceso sea menor a un valor umbral o de referencia. El umbral se fija de modo que, dentro de la ventana utilizada, la dinámica de las variables quede bien representada.

La anterior estrategia de identificación, aplicada en cada nuevo instante de muestreo t_c del proceso puede resultar en un esfuerzo computacional considerable. Una manera de reducir este esfuerzo, que a su vez produzca el mínimo efecto sobre la calidad de los resultados, sería como sigue. Se aplica el análisis de identificación descrito, en el primer instante en que se comienza la reconciliación. Los resultados (p_i y p_j identificados) se utilizan en las reconciliaciones que siguen. Cuando ocurra un cambio importante (por ejemplo, un salto en una o varias señales debido a un cambio del punto de consigna de un controlador), se hace un nuevo análisis de identificación de modo de adaptar (si es necesario) los p_i y p_j según los cambios experimentados.

La estrategia anterior, brinda un modo sencillo de adaptar los polinomios a la dinámica reinante en cada instante y, además, se puede implementar con facilidad. En una fase inicial de trabajo (Benqlilou *et al*., 2001), se adoptó este mismo procedimiento para la aplicación de EPA en casos dinámicos. Finalmente, si se conoce que el proceso no es propenso a sufrir cambios bruscos (dinámicas suaves), se podría seleccionar una ventana de datos suficientemente pequeña y, luego, se podrían fijar, para estas ventanas, valores constantes y pequeños (2 ó 3 y mucho menos 4) de p_i y p_j. El conjunto de p_i obtenido según cualquiera de los procedimientos anteriores, se introducen en la ecuación (3.13). Luego, los coeficientes determinados en cada instante se utilizan como la solución inicial para la optimización numérica del problema RDD.

3.2.4 Cálculo de las varianzas

El cálculo de la matriz **Q** asociada a las varianzas de los errores de medición es una de las principales dificultades que deben resolverse al afrontar sistemas dinámicos. Esto se debe a que a la variabilidad del proceso se añade la variabilidad de las mediciones y ambas pueden cambiar continuamente.

Antes de explicar la forma en que se calculará esta varianza, se discuten una serie de condiciones a tomar en cuenta para el cálculo de **Q**.

- Se asume que un error puede verse como la suma de un número considerable de errores más pequeños. De acuerdo al *Teorema de Límite Central* y bajo condiciones generalmente aceptables, la distribución de tales sumas se puede aproximar mediante una distribución normal. Luego, en este trabajo se asume que las variabilidades en los errores de medición siguen una distribución normal.
- Dado que no se conoce ni el valor esperado ni la media muestral de las variables medidas, el cálculo de las dispersiones se basará en los valores estimados de cada variable.

- Se asume que las mediciones son independientes, de modo que los elementos no pertenecientes a la diagonal de **Q** se anulan (se igualan a cero).

Se podrían utilizar todas las mediciones y estimaciones pasadas para estimar la varianza. No obstante, si la varianza no es constante a través del tiempo, esto podría provocar errores en la estimación de **Q**. Esto es por que, en términos de variabilidad, las mediciones actuales se asemejarán más a las mediciones más recientes. Por lo tanto, la adopción del enfoque de horizonte móvil explicado en secciones precedentes será útil para calcular la varianza, ya que en la ventana solo principalmente se atrapa el efecto de variabilidades de mediciones pasadas recientes y de mediciones actuales.

Basado en las condiciones anteriores, se construye una estimación recursiva de la matriz **Q**, que se actualiza continuamente por utilizar las *nc* últimas mediciones y estimaciones según se muestra en la siguiente ecuación:

$$\hat{\sigma}_{m_i} = \frac{1}{nc-1} \sum_{k=c-nc+1}^{c} \left(\hat{m}_i(k) - m_i(k)\right)^2 \tag{3.14}$$

y

$$\hat{\mathbf{Q}} = \{\hat{\sigma}_{m_i}\} \tag{3.15}$$

Donde m_i puede ser una variable de tipo **w** (derivada de variable de estado) o de tipo **y** (variable de entrada/salida).

3.2.5 Reformulación del problema RDD y Resolución en línea

La reformulación del problema RDD (ecuación 3.1) se puede expresar en su forma final mediante las ecuaciones:

$$\min \sum_{k=c-n+1}^{c} \left[\frac{\hat{\mathbf{y}}(k) - \mathbf{y}(k)}{\hat{\sigma}_y}\right]^2 + \left[\frac{\hat{\mathbf{w}}(k) - \mathbf{w}(k)}{\hat{\sigma}_w}\right]^2 \tag{3.16}$$

Sujeto a:

$$\hat{\xi}_{j,r} = \sum_{i=0}^{I} C_{j,i} \cdot \hat{\alpha}_{i,r}, \quad \forall j \tag{3.12}$$

En la formulación anterior, los estimados $\hat{\mathbf{y}}_i(k)$ y $\hat{\mathbf{w}}_j(k)$ se calculan, en cada iteración de la optimización, mediante las ecuaciones (3.10) y (3.11), y los coeficientes polinomiales se van corrigiendo de una iteración a otra para alcanzar el mínimo de la función objetivo. Luego, los coeficientes reconciliados, $\hat{\xi}_{j,r}$ y $\hat{\alpha}_{j,r}$, se usan para reconstruir las variables de proceso reconciliadas utilizando las ecuaciones (3.17) y (3.18):

$$\hat{x}_j(k) = x_j(0) + \sum_{p=0}^{p_j} \frac{\hat{\xi}_{j,r} \cdot k^{p+1}}{p+1}, \quad \forall j = 1...J \tag{3.17}$$

$$\hat{y}_i(k) = \sum_{p=0}^{p_i} \hat{\alpha}_{i,r} \cdot k^p, \quad \forall i = 1...I \tag{3.18}$$

Como **w** representa un término derivativo, la correspondiente variable de estado se debe encontrar por integración de **w**. La ecuación (3.17) representa la integral de $\mathbf{w}_j$ según la

ecuación (3.11). El término $x_j(0)$, es la constante de integración y su valor corresponde a las condiciones iniciales de la integración.

Se debe notar que la formulación mediante las ecuaciones (3.16) y (3.12) también permite la introducción de restricciones de desigualdad como por ejemplo límites para evitar polinomios sobre-especificados en alguna de las variables. En los ejemplos que se muestran en la sección 3.4, el problema se resuelve numéricamente utilizando un algoritmo de gradiente conjugado pre-condicionado (Optimization-Toolbox, 2003) aunque se podrían utilizar otros métodos para la resolución numérica.

3.3 Integrando la Filtración-Extracción de Tendencias mediante wavelets con la RDD basada en EPA

En esta sección, se presenta la estrategia RDD que se propone en este capítulo. En esencia, la estrategia es un proceso de rectificación en 2 etapas que integra una estrategia de filtrado mediante *wavelets* con EPA (WEPA). En la primera etapa, se comienza por filtrar el ruido de las mediciones, utilizando para ello la estrategia *levashrink* descrita en el capítulo 2. Aún cuando estas tendencias son buenas estimaciones de los valores reales (ver secciones 2.1.1.2 y 2.3), podrían ser inconsistentes con las restricciones del modelo del proceso. Por ello, en la segunda etapa, las tendencias son reconciliadas mediante EPA de modo de asegurar la consistencia de las mismas. A continuación se describe el algoritmo de la estrategia WEPA.

3.3.1 Consideraciones Adicionales sobre el Horizonte Móvil

En la metodología WEPA el reconciliado se obtiene a partir de las tendencias extraídas y no de las mediciones. Por tanto, los esquemas de horizonte móvil descritos en la sección 3.2.2 se vuelven a considerar para el caso WEPA.

- Reconciliación uno a uno para WEPA (caso OBOR): El esquema se mantiene de modo similar al descrito en la sección 3.2.2. En cada instante se construyen ventanas con la medición del tiempo actual y las *nc*-1 mediciones más recientes. Pero en el caso de WEPA, las ventanas anteriores se filtran y, luego, se continúa la aplicación del WEPA según se describe en la sección 3.3.2. En la figura 3.3 se muestra un esquema del caso OBOR para WEPA.
- Reconciliación Corregida para WEPA (caso CR): La aplicación inicial de este caso es similar a la del caso CR de la sección 3.2.2. En el primer instante k_c (c), se construye la ventana con las mediciones disponibles más recientes desde $k_{c\text{-}nc+1}$ hasta k_c. Sobre esta ventana se sigue la aplicación del WEPA según se describe en la sección 3.3.2. El vector de reconciliados obtenidos (desde $k_{c\text{-}nc+1}$ hasta k_c) se guarda. En los siguientes instantes k_c, se actualiza la ventana de datos de la siguiente manera:
    - Se toman las últimas *nc-1* mediciones y se le añade la registrada en k_c.
    - Se filtra la ventana anterior con *wavelets*.
    - Se toman los últimos *nc-1* reconciliados y se añade, del filtrado del paso anterior, el valor filtrado en k_c.

  Sobre esta ventana se sigue la aplicación de WEPA. El vector de reconciliados obtenidos (desde $k_{c\text{-}nc+1}$ hasta k_c) se guarda como el conjunto de reconciliados actual. En la figura 3.4 se ilustra WEPA según el caso CR.

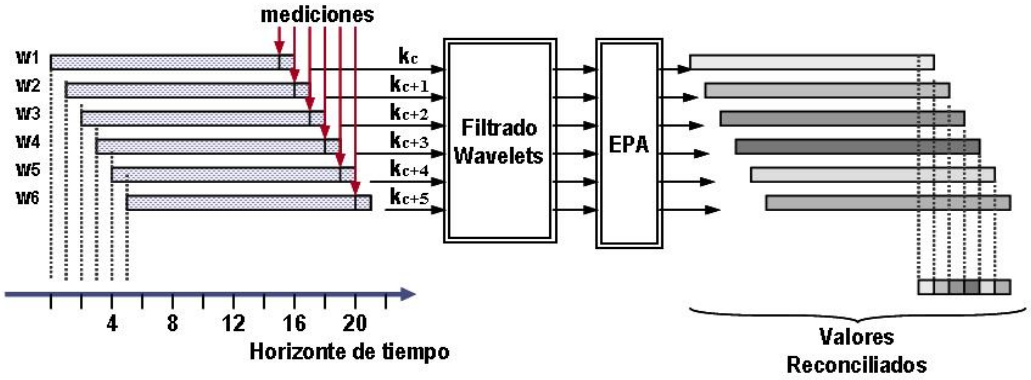

Figura 3.3. Esquema OBOR para Reconciliación WEPA.

Asimismo, se añade una observación relacionada a la longitud de la ventana que afecta por igual a los 2 esquemas anteriores (OBOR y CR). Para aplicaciones reales, en la literatura se han propuesto y usado versiones discretas de las *wavelets* (ver capítulo 2). La discretización asociada se aplica sobre escalas de longitud diádica, esto es potencias de 2, lo que significa que la descomposición con *wavelets* se aplica sobre ventanas de longitud 2^l. Por lo tanto, se recomienda que la longitud *nc* seleccionada sea también un múltiplo de 2^l, esto es, 2, 4, 8, 16, 32,…, etc. Si esto no se hace, la estimación en los extremos de la ventana de datos usada, podría verse levemente afectada (Addison, 2002).

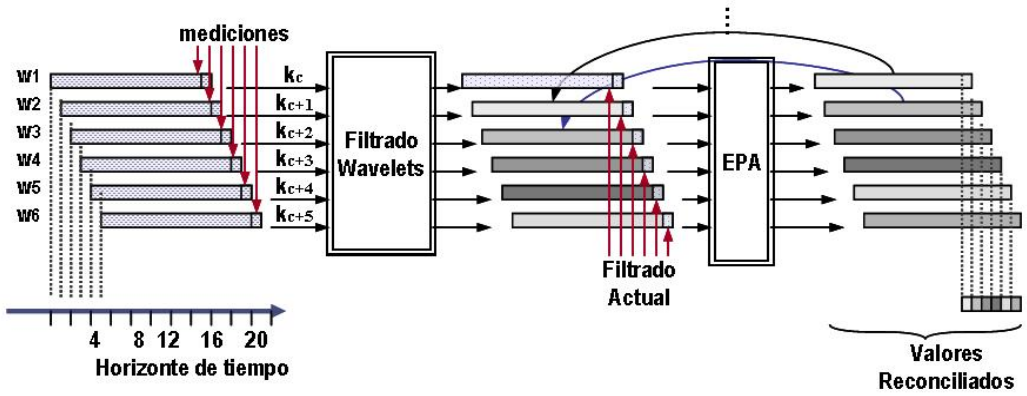

Figura 3.4. Esquema CR para Reconciliación.

3.3.2 Algoritmo de la estrategia RDD propuesta

Para la aplicación de la estrategia propuesta WEPA, se asume que se dispone en cada instante k_c de mediciones tanto de las variables de proceso **y** como de las variables de estado **x**. Asimismo, se asume que se cuenta con el modelo requerido de la planta. Los pasos a seguir son como sigue:

- **Paso 1:** Se toma una ventana de datos, de longitud *nc*, con las mediciones más recientes desde el tiempo actual k_c hasta k_c-*nc+1*, de todas las variables $\mathbf{x}_j$ y $\mathbf{y}_i$. Se generan las mediciones asociadas a las variables ficticias $\mathbf{w}_j$ por tomar la derivada de las correspondientes $\mathbf{x}_j$.
- **Paso 2:** Se lleva a cabo el filtrado de las mediciones para cada variable $\mathbf{w}_j$ y cada variable $\mathbf{y}_i$. Para ello se aplica la estrategia *levashrink*. Así, se obtiene un primer estimado (filtrado) de las tendencias de cada variable que se identifican como $\hat{\mathbf{w}}_j^{tr}$ y

$\hat{\mathbf{y}}_j^{tr}$. Si el esquema de ventana adoptado es el CR, tras el filtrado, se debe actualizar la ventana de datos según se explica en la sección 3.3.1.

- **Paso 3:** Se estima la varianza para cada una de las variables consideradas a partir de la diferencia entre la variable medida (m) y la filtrada (m^{tr}), por aplicación de la ecuación (3.14) que se rescribe como sigue:

$$\hat{\sigma}_{m_i} = \frac{1}{nc-1} \sum_{k=c-nc+1}^{c} \left(\hat{m}_i^{tr}(k) - m_i(k) \right)^2 \qquad (3.19)$$

  Donde m representa a cualquiera de las variables $\mathbf{w}_j$ o $\mathbf{y}_i$.

- **Paso 4:** Se identifican los grados de los polinomios que se ajustan a cada una de las tendencias estimadas para cada variable ($\hat{\mathbf{w}}_j^{tr}$ y $\hat{\mathbf{y}}_j^{tr}$) junto con los coeficientes polinomiales $\hat{\boldsymbol{\xi}}_j^{tr}$ y $\hat{\boldsymbol{\alpha}}_j^{tr}$ correspondientes.

- **Paso 5:** Se aplica la RD basada en EPA, según las ecuaciones (3.16) y (3.12). De este modo, se obtienen los coeficientes corregidos (consistentes con el modelo del proceso) $\hat{\boldsymbol{\xi}}_j^{*}$ y $\hat{\boldsymbol{\alpha}}_j^{*}$.

- **Paso 6:** Los coeficientes reconciliados $\hat{\boldsymbol{\xi}}_j^{*}$ y $\hat{\boldsymbol{\alpha}}_j^{*}$ se utilizan para recuperar las correspondientes tendencias reconciliadas de cada variable ($\hat{\mathbf{w}}_j^{*}$ y $\hat{\mathbf{y}}_i^{*}$). La recuperación se aplica mediante las ecuaciones (3.17) y (3.18).

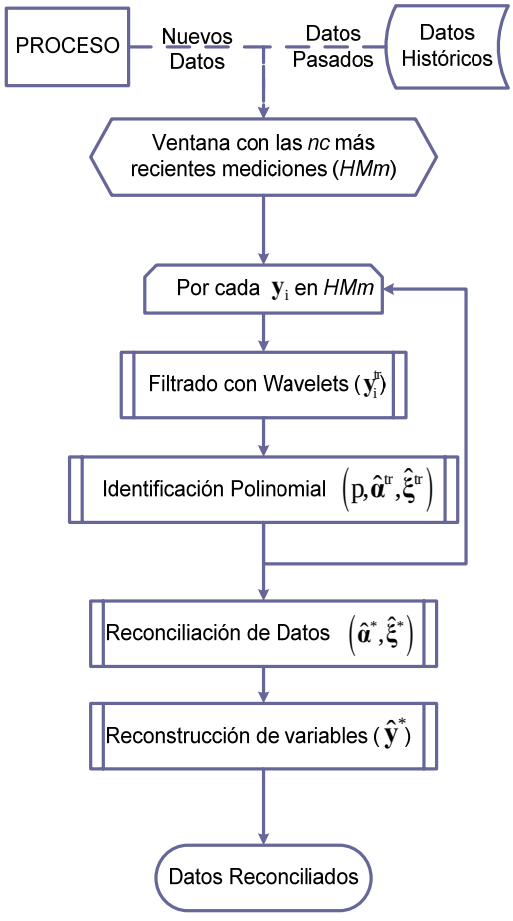

Figure 3.5. WEPA según esquema de ventana OBOR.

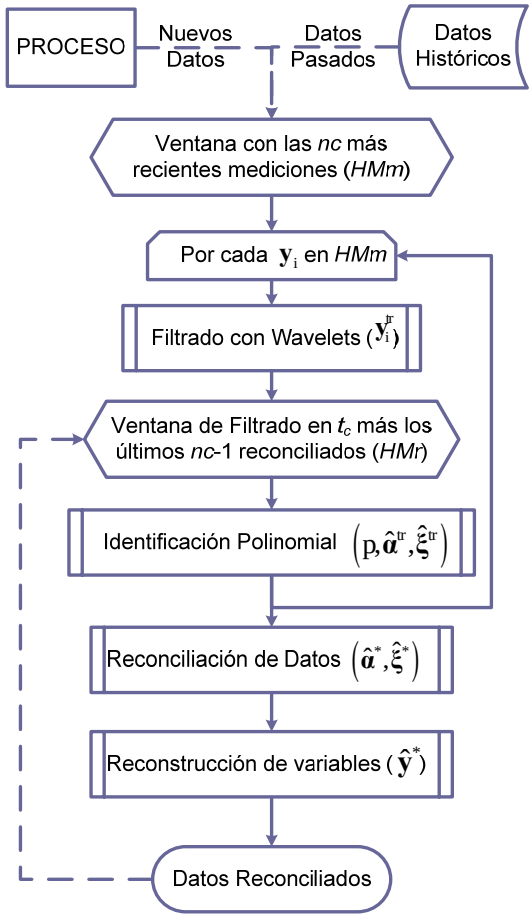

Figure 3.6. WEPA según esquema de ventana CR.

En el paso 5, los coeficientes $\hat{\boldsymbol{\xi}}_j^{tr}$ y $\hat{\boldsymbol{\alpha}}_j^{tr}$, obtenidos en el paso 4, se utilizan como la solución inicial para la optimización numérica del EPA. Esto es, se calculan los estimados iniciales de cada variable (los $\hat{\mathbf{w}}_j^*$ y $\hat{\mathbf{y}}_i^*$ iniciales) con los coeficientes $\hat{\boldsymbol{\xi}}_j^{tr}$ y $\hat{\boldsymbol{\alpha}}_j^{tr}$, y según las ecuaciones (3.10) y (3.11). Luego, en cada iteración de la optimización los estimados $\hat{\mathbf{w}}_j^*$ y $\hat{\mathbf{y}}_i^*$ se recalculan con los coeficientes corregidos en la última iteración y según las ecuaciones (3.16) y (3.12). Por otro lado, dado que se busca la consistencia de las tendencias estimadas $\hat{\mathbf{w}}_j^{tr}$ y $\hat{\mathbf{y}}_j^{tr}$, estas sustituyen a las correspondientes mediciones en la función objetivo (ecuación 3.16). De todo lo anterior resulta que la función objetivo queda representada como:

$$\min \sum_{k=c-n+1}^{c} \left\{ \left[\frac{\hat{m}^*(k) - \hat{m}^{tr}(k)}{\hat{\sigma}_m} \right]^2 \right\} \tag{3.20}$$

Donde *m* representa a cualquiera de las variables $\mathbf{w}_j$ o $\mathbf{y}_i$.

El algoritmo se puede volver a aplicar en cada nuevo instante t_c en que se generen nuevas mediciones del proceso y según se describió anteriormente. La ventana de datos se construye según se explica en la sección 3.3.1. Finalmente, En las figuras 3.5 y 3.6 se muestran los diagramas de flujo de la estrategia WEPA aplicada según los esquemas de ventana OBOR y CR

3.3.3 WEPA para Sistemas No Lineales

Debido a la complejidad de muchos procesos, los modelos aproximados a muchos de ellos se componen no solo de balances de masa y energía sino que también pueden involucrar relaciones y correlaciones de equilibrio, propiedades físicas, etc. Todo esto conduce a sistemas no lineales y problemas de RD no lineales (RD no lineal) los cuales representan un reto en la literatura actual sobre RD (ver sección 3.1).

En este capítulo, se propone un manejo sencillo y eficiente del problema PNL. De acuerdo a la formulación general del problema RDD que se presenta en la sección 3.1 las restricciones del problema que se asocian al modelo del proceso vienen dadas por las ecuaciones (3.2), (3.3) y (3.4). Asumiremos que el modelo del proceso solo vendrá representado por (3.2) y (3.3). Ambos tipos de ecuaciones pueden escribirse de manera equivalente como sigue:

$$\frac{d\hat{\mathbf{y}}_j(k)}{dk} = \mathbf{f}_j\left(\hat{\mathbf{y}}(k), \hat{\mathbf{u}}(k), \hat{\theta}(k)\right) = \sum_{i=1}^{I^z} \mathrm{f}_i\left(\hat{\mathbf{y}}(k), \hat{\mathbf{u}}(k), \hat{\theta}(k)\right) = \mathrm{f}_1(\ldots) + \ldots + \mathrm{f}_{I^z}(\ldots) \tag{3.21}$$

$$0 = \mathbf{h}_j\left(\hat{\mathbf{y}}(k), \hat{\mathbf{u}}(k), \hat{\theta}(k)\right) = \sum_{i=I^z+1}^{I} \mathrm{h}_i\left(\hat{\mathbf{y}}(k), \hat{\mathbf{u}}(k), \hat{\theta}(k)\right) = \mathrm{h}_{I^z+1}(\ldots) + \ldots + \mathrm{h}_I(\ldots) \tag{3.22}$$

La esencia del método propuesto descansa en representar cada uno de los términos individuales de cada ecuación como una nueva variable ***z***(*k*). Estas últimas, se obtienen de las variables $\hat{\mathbf{y}}(k)$ y $\hat{\mathbf{u}}(k)$, de los parámetros $\hat{\theta}$ y mediante las funciones $f_i(.)$ y $h_i(.)$, tal que el modelo en (3.21) y (3.22) puede re-expresarse como sigue:

$$\frac{d\hat{\mathbf{y}}_j(k)}{dk} = \mathrm{f}_1(\ldots) + \ldots + \mathrm{f}_{I^z}(\ldots) = \hat{z}_1(k) + \ldots + \hat{\mathrm{z}}_{I^z}(k) = \sum_{i=1}^{I^z} \hat{\mathrm{z}}_i(k) \tag{3.23}$$

$$0 = \mathrm{h}_{I^z+1}(\ldots) + \ldots + \mathrm{h}_{I}(\ldots) = \hat{\mathrm{z}}_{I^z+1}(k) + \ldots + \hat{\mathrm{z}}_{I}(k) = \sum_{i=I^z+1}^{I} \hat{\mathrm{z}}_i(k) \qquad (3.24)$$

Adicionalmente, para los términos derivativos se aplica el mismo cambio de variable que se discutió en la sección 3.2.1, lo que conduce a las variables ficticias de proceso **w**(*k*) y a que el modelo de las ecuaciones quede finalmente expresado de forma lineal como sigue:

$$\hat{\mathbf{w}}_j(k) = \mathrm{f}_1(\ldots) + \ldots + \mathrm{f}_{I^z}(\ldots) = \hat{z}_1(k) + \ldots + \hat{\mathrm{z}}_{I^z}(k) = \sum_{i=1}^{I^z} \hat{\mathrm{z}}_i(k) \qquad (3.25)$$

$$0 = \mathrm{h}_{I^z+1}(\ldots) + \ldots + \mathrm{h}_{I}(\ldots) = \hat{\mathrm{z}}_{I^z+1}(k) + \ldots + \hat{\mathrm{z}}_{I}(k) = \sum_{i=I^z+1}^{I} \hat{\mathrm{z}}_i(k) \qquad (3.26)$$

Luego, la estrategia WEPA se aplica para reconciliar ambas variables **w**(*k*) y **z**(*k*), con el modelo de las ecuaciones (3.25) y (3.26) como restricción. Los estimados de las mediciones para las **z**(*k*) y sus varianzas, se obtienen por evaluar la correspondiente función (f_i(.) o h_i(.)) con las mediciones de $\hat{\mathbf{y}}(k)$ y $\hat{\mathbf{u}}(k)$ y los valores conocidos de los parámetros $\hat{\theta}$.

3.4 Caso de estudio. Resultados y Discusión.

En esta sección se presenta la aplicación del método propuesto en secciones anteriores. Se escoge un caso típico en la literatura sobre métodos RDD (Liebman *et al*., 1992; Romagnoli y Sánchez, 2000). Este consiste en simulaciones de un reactor de tanque agitado y continuo, CSTR, con intercambio de calor externo. Los detalles del mismo se muestran en el anexo C.

El *CSTR* se utiliza para simular diferentes escenarios de operación del reactor con las correspondientes mediciones de las variables de estado *Ca, T, V,* y de las variables de entrada/salida Ca_0, T_0, q_0 y q. Las simulaciones se llevan a cabo utilizando un paso de integración de 1 unidad de tiempo y durante un horizonte de tiempo de 60 unidades. También, se añade ruido gausiano con una desviación estándar del 5% a los valores simulados de las variables medidas.

3.4.1 Caso Lineal – Reconciliados de información asociada a los balances en el reactor

Para poder plantear un modelo lineal, en esta sección se asumirá un escenario de operación donde el reactor descrito anteriormente se encuentra en operación y los operadores están interesados en los estimados de los flujos de entrada/salida del reactor así como del volumen del material en el tanque. Dado el anterior requerimiento, el modelo del proceso se puede simplificar a un sistema lineal representado mediante la ecuación C.1 (ver anexo C). Se simula la operación como una respuesta en lazo abierto. Inicialmente, el valor verdadero de q_0 es de 10 cm^3s^{-1}. En el instante 30, el caudal q_0 se ve afectado por un efecto rampa de pendiente 0.02, mientras que q se mantiene controlada (la válvula no es manipulada por el operador) a un valor de 10 cm^3s^{-1}.

En cada instante de simulación se aplica la estrategia WEPA sobre las mediciones disponibles según el esquema OBOR para el horizonte móvil. Lo anterior (la simulación y la aplicación continua de WEPA) se vuelve a hacer, pero adoptando el esquema CR para el horizonte móvil. El rendimiento de WEPA (bajo ambos esquemas OBOR y CR) se compara con la aplicación en línea de una estrategia RDD con Filtros de Kalman (FK).

En un trabajo previo de la literatura (Benqlilou *et al*., 2002), se mostró la ventaja comparativa del FK como técnica RD frente a otros métodos basados en discretización (Rollins y Devanathan, 1993) y frente a métodos basados en polinomios (Bagajewicz y Jiang, 1997). Es por esto último que se usa el FK como técnica de comparación para el presente trabajo. Adicionalmente, en el presente análisis se incluye la comparación con el EPA. En todos los casos la comparación se hace en términos de Error Cuadrático Medio (ECM) y el Error Porcentual Absoluto Medio (EPAM). El ECM es útil para contabilizar la reducción de la variabilidad en cada variable, mientras que el EPAM es útil para fijar la desviación de los estimados en términos de porcentaje. Se utilizaron ventanas de 16 valores para todas las estrategias RDD, esto es, EPA y WEPA cada una según los esquemas OBOR y CR respectivos. Para el caso de FK, la ventana no hizo falta dado que en esta técnica solo se usa un valor en cada nuevo instante (ver anexo A). Por último, se evaluó la aplicación de WEPA para diferentes *wavelets Daubechies* encontrándose que, para este caso, se obtenían los mejores estimados adoptando la *db2*.

3.4.1.1 Caso OBOR

En las figuras 3.7, 3.8 y 3.9 se muestra el comportamiento de q_0, q y V que se obtiene al operar según se describió al inicio de la sección 3.4.1. Asimismo, se muestran los reconciliados obtenidos mediante WEPA con esquema OBOR y mediante FK. En cada instante t (t=k) el reconciliado obtenido viene representado por el último valor de la ventana reconciliada en ese instante. Luego, no se muestran los reconciliados de los primeros 15 instantes (primera ventana procesada) por que hasta t=16 es que se tiene la primera ventana completa.

La precisión de la estimación para el método propuesto (etiquetado como WEPA-OBOR) es levemente mejor que la del FK para las variables q_0 y q como se muestra en la tabla 3.1. En cambio, la precisión de la estimación para V es mejor con el FK (ver tabla 3.1) aunque, como puede verse en la figura 3.9, el rendimiento del método propuesto es muy bueno con una muy leve desviación en el tiempo

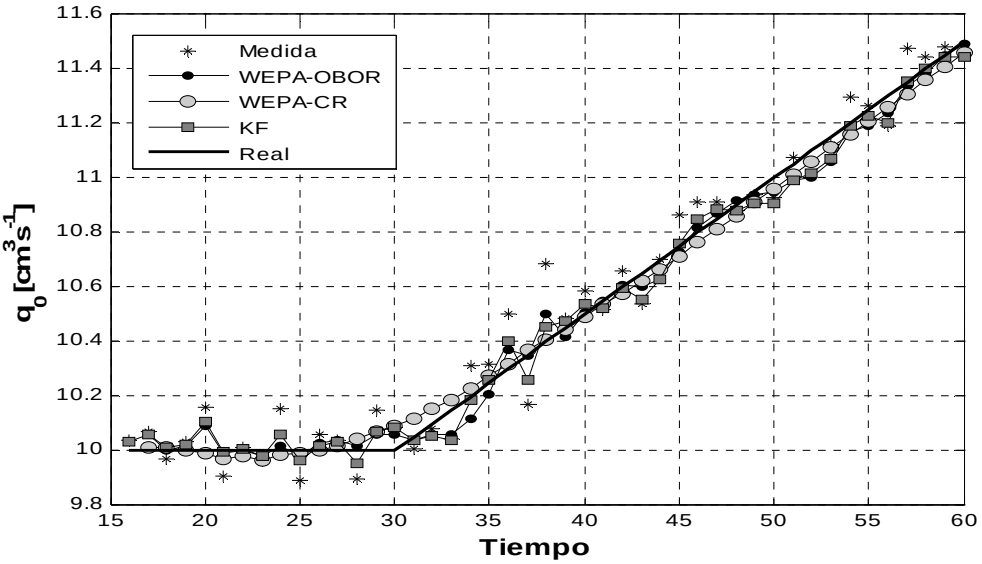

Figura 3.7. Reconciliados del flujo de alimentación (q_0).

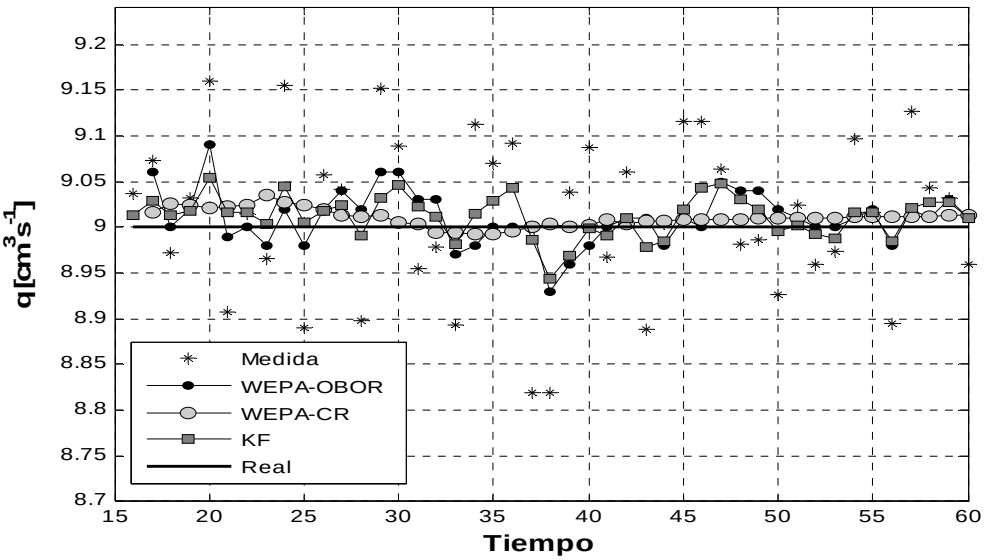

Figura 3.8. Reconciliados del flujo de salida (q).

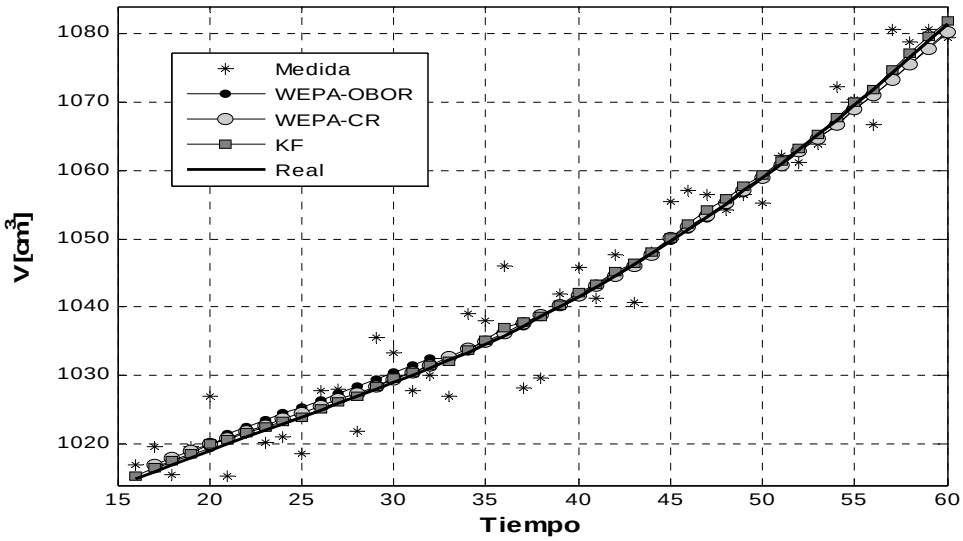

Figura 3.9. Reconciliados del Volumen en el tanque (V).

Luego, la estrategia WEPA, cuando se aplica bajo el esquema OBOR (WEPA-OBOR), no supera globalmente a la RD con FK. Esto puede deberse a 2 factores:

- Primero, en este caso los valores reconciliados en cada instante k_c corresponden a los extremos de la tendencia extraída mediante *wavelets*. Como se resalta en la literatura y pese a las soluciones existentes (Nounou y Bakshi, 1999), en ocasiones las estimaciones basadas en *wavelets* pueden ser levemente erróneas en los extremos (Addison, 2002; Misiti *et al*., 2004). Por lo tanto, el efecto de los bordes de la estimación con *wavelets* ligeramente desviados en algunos instantes de tiempo, se pudieron haber propagado a los reconciliados del WEPA.
- En segundo lugar, la variable *V* se recupera por integración. La determinación de las condiciones iniciales para la integración al tiempo k=1 se hicieron manualmente. Probablemente, la determinación de estas condiciones iniciales mediante un método más exhaustivo podría mejorar la recuperación por la integración.

Tabla 3.1. Errores asociados a las diferentes señales reconciliadas.

	ECM			EPAM		
	q_0	q	V	q_0	q	V
Mediciones	0.0096	0.0077	20.3996	0.75	0.81	0.35
$WEPA_{OBOR}$	0.0023	0.0010	0.7592	0.35	0.26	0.07
$WEPA_{CR}$	0.0007	0.0003	0.5608	0.18	0.17	0.06
EPA_{OBOR}	0.0069	0.0028	1.4423	0.6	0.43	0.09
EPA_{CR}	0.0012	0.0006	0.5276	0.25	0.23	0.05
FK	0.0030	0.0011	0.2872	0.42	0.3	0.05

Finalmente, de la tabla 3.1 se puede deducir con facilidad que la estrategia WEPA-OBOR supera a la estrategia EPA según OBOR (EPA-OBOR) en términos de la precisión de los estimados. El EPA también se ve afectado por problemas de estimación de los polinomios en los bordes y su efecto se ve reducido cuando se usan las *wavelets*. Por lo tanto, la integración del filtrado mediante *wavelets* con el EPA mejora las estrategias de reconciliación basadas en representación polinomiales.

3.4.1.2 Caso CR

En este caso se utiliza el mismo escenario de operación descrito al inicio de la sección 3.4.1. Los resultados de la reconciliación en cada instante de la simulación se muestran en las figuras 3.7, 3.8 y 3.9 así como en la tabla 3.1.

Se observa que la estrategia WEPA supera con creces a FK en términos de precisión de los reconciliados para el caso de las variables q_0 y q. Los errores de medición en la tabla 3.2 indican que la estimación de V mediante FK es mejor que la aportada por el WEPA. No obstante, el gráfico de la figura 3.9 muestra que la estimación mediante WEPA es también buena y aun mejor que la obtenida bajo el esquema OBOR. Los resultados muestran que en general y en términos de precisión de los estimados, WEPA bajo CR (WEPA-CR) es globalmente una mejor alternativa frente al FK. Los resultados para el *ECM* también muestran una reducción de la variabilidad bastante considerable así como un mejor aspecto visual cuando se compara tanto con EPA bajo OBOR como con FK. La tabla 3.1 también muestra que la aplicación de EPA bajo el esquema CR (EPA-CR) es muy buena. Esto significa que, bajo los esquemas CR, las variabilidades asociadas a la RD basados en polinomios se reducen drásticamente. No obstante, globalmente el WEPA bajo CR es mejor que el correspondiente EPA bajo CR.

Los resultados anteriores son validos para casos con variaciones de procesos suaves. Sin embargo, si una variable de proceso presenta dinámicas fuertes a un tiempo específico (por ejemplo, un cambio en escalón), el grado del polinomio de tal variable será considerablemente alto. Para tales casos (discontinuidades fuertes), se sabe que el ajuste polinómico nunca es bueno. Luego, pese a que la *wavelets* puede seguir la dinámica de las señales afectadas por discontinuidades con bastante precisión, la representación polinómica conducirá a reconciliados no satisfactorios. Para estos casos, se propone como alternativa filtrar con *wavelets* en los instantes inmediatamente posteriores a la ocurrencia de una discontinuidad fuerte con un tamaño de ventana de $nc' = nc/2$ (siempre que nc' sea mayor que 8), manteniendo el filtrado con *wavelets* hasta un tiempo en el que se tengan suficientes datos para aplicar de nuevo WEPA (8 ó mas muestras posteriores al salto). En la figura 3.10 se muestra un ejemplo de esto. Se observa claramente que las mejores estimaciones que se

obtienen en los instantes inmediatamente posteriores al salto corresponden a las del filtrado con *wavelets*.

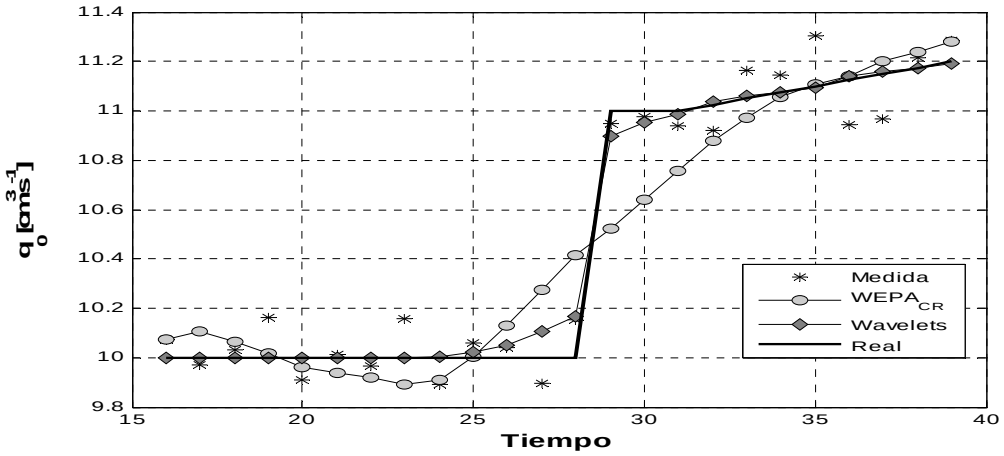

Figura 3.10. Rendimiento de WEPA ante una discontinuidad.

Todas las implementaciones de los algoritmos utilizados junto con las simulaciones llevadas a cabo se hicieron en MATLAB 7.0 y se probaron en un ordenador AMD Athlon XP+3000. El tiempo medio de resolución de la reconciliación mediante WEPA fue de 0.46 segundos, mientras que con el FK fue de 0.0001 segundos. La RDD se ha planteado en la literatura principalmente como tarea de soporte al control, la monitorización y la Optimización en tiempo real. En situaciones reales, los tiempos de muestreo son frecuentemente mayores a 1 segundo en tareas de monitorización y control, y mayores a 1 minuto para aplicaciones como la Optimización en Tiempo Real. Luego, dados los tiempos de resolución obtenidos, las estrategias propuestas se podrían aplicar sobre casos reales, de dimensiones similares a las del caso de estudio mostrado, sin problemas.

3.4.2 Caso no Lineal – Reconciliación de variables de estado

En este apartado, se presenta la aplicación de la extensión del WEPA al caso no lineal, según se expuso en la sección 3.3.3. Se utiliza el mismo caso de estudio de la sección anterior, pero asumiendo un escenario de operación donde el CSTR se encuentra en trabajando igual que antes pero los operadores están también interesados en el estimado de las concentraciones y temperaturas en las corrientes de entrada y salida del reactor. De este modo, el modelo del proceso deberá también incluir las ecuaciones (C.2), (C.3) y (C.4) (ver anexo C) haciéndose un sistema no lineal. Operando sobre el lado derecho de las ecuaciones anteriores es posible obtener las nuevas variables $\mathbf{z}(t)$ tras adoptar las siguientes substituciones:

$$\hat{\mathbf{w}}_1 = \frac{dV}{dt}, \quad \hat{\mathbf{z}}_1 = q_0, \quad \hat{\mathbf{z}}_2 = -q \tag{3.27}$$

$$\hat{\mathbf{w}}_2 = \frac{dCa}{dt}, \quad \hat{\mathbf{z}}_3 = \frac{q_0}{V}.Ca_0, \quad \hat{\mathbf{z}}_4 = \frac{q_0}{V}.Ca, \quad \hat{\mathbf{z}}_5 = K.Ca \tag{3.28}$$

$$\hat{\mathbf{w}}_3 = \frac{dT}{dt}, \quad \hat{\mathbf{z}}_6 = \frac{q_0}{V}.T_0, \quad \hat{\mathbf{z}}_7 = \frac{q_0}{V}.T, \quad \hat{\mathbf{z}}_8 = \frac{-\alpha_d.\Delta H.Ca_r.K.Ca}{\rho.C_p.T_r} \tag{3.29}$$

y

$$\hat{z}_9 = \frac{-U.A_R}{r.C_p.V}.(T - T_c) \tag{3.30}$$

Por lo tanto, es posible reformular el modelo CSTR gobernado por las ecuaciones (C.1) a (C.4) como sigue:

$$\hat{\mathbf{w}}_1 = \hat{\mathbf{z}}_1 - \hat{\mathbf{z}}_2 \tag{3.31}$$

$$\hat{\mathbf{w}}_2 = \hat{\mathbf{z}}_3 - \hat{\mathbf{z}}_4 - \hat{\mathbf{z}}_5 \tag{3.32}$$

$$\hat{\mathbf{w}}_3 = \hat{\mathbf{z}}_6 - \hat{\mathbf{z}}_7 + \hat{\mathbf{z}}_8 - \hat{\mathbf{z}}_9 \tag{3.33}$$

Luego, tomando este modelo de proceso se puede aplicar WEPA según se describió en la sección 3.3.2. En este caso, el procedimiento se evalúa a cada paso de simulación y bajo el esquema CR. La reconciliación se lleva a cabo de manera secuencial, esto es: Primero se resuelve la reconciliación asociada al balance de masa global (ecuación 3.31), luego se resuelve la reconciliación asociada al componente del balance de masa (ecuación 3.32) y, finalmente se resuelve la reconciliación asociada al balance de energía (ecuación 3.33). Tras esta resolución secuencial, se recuperan los valores reconciliados de cada variable de proceso (Ca_0, T_0,…, etc.), por trabajar sobre la variable z_i apropiada.

Los reconciliados obtenidos se muestran en las figuras 3.11 y 3.12. Se observa claramente el buen rendimiento de las estrategias propuestas.

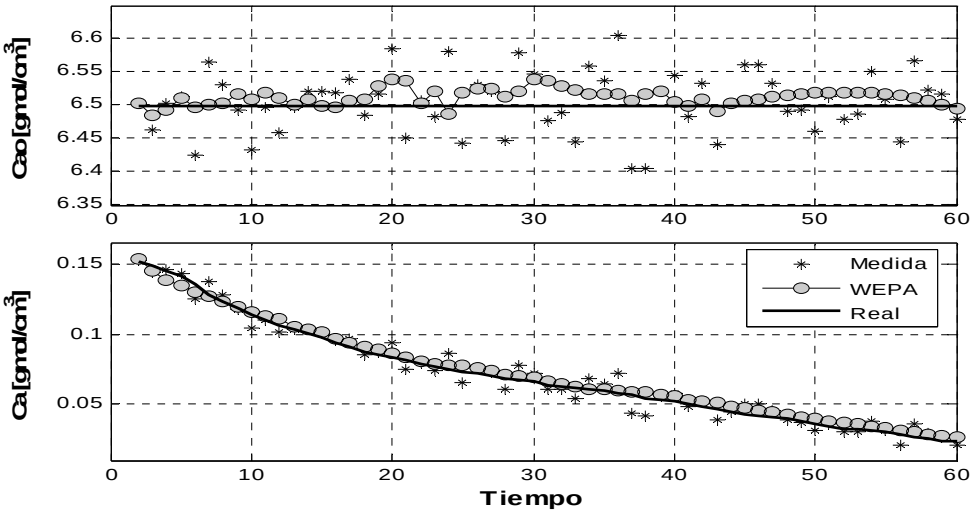

Figura 3.11. Reconciliados de las concentraciones (Ca_0-Ca).

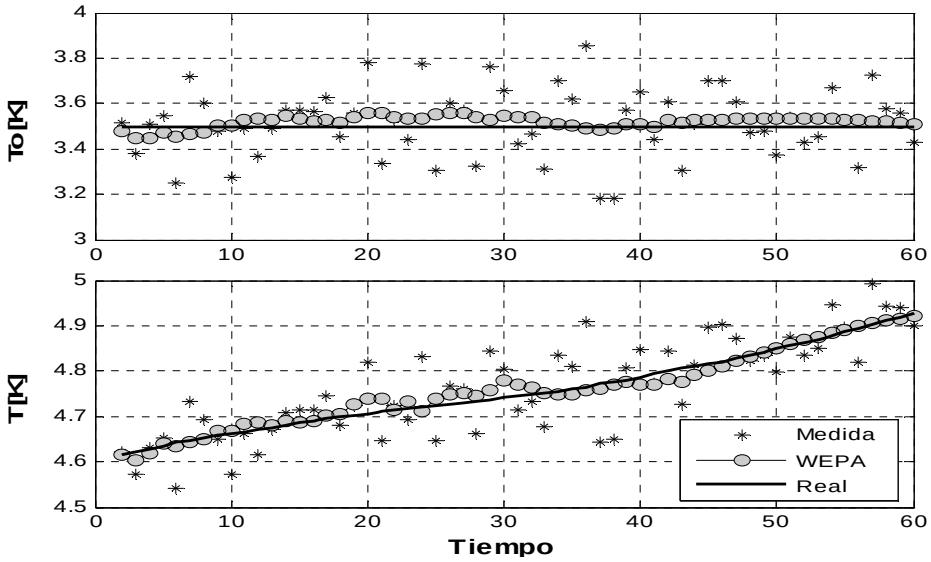

Figura 3.12. Reconciliados de las concentraciones (T_0-T).

3.5 Conclusiones

En este trabajo se presenta una técnica de reconciliación de datos para casos dinámicos lineales que combina el análisis (la extracción) de tendencias con una técnica de reconciliación basada en polinomios. La representación de las tendencias de las variables medidas se obtiene por la aplicación de filtrado mediante *wavelets*. Una vez obtenidas estas tendencias, las mismas se hacen consistentes con el modelo del proceso por optimizar los coeficientes polinomiales que mejor las representan.

Se ha visto que la aplicación previa del análisis de las tendencias mediante filtración con *wavelets* puede proporcionar beneficios significativos. En primer lugar, se reduce la complejidad de evaluar la matriz de varianzas-covarianzas de las variables del proceso, ya que la variabilidad en las tendencias se debe básicamente al proceso mismo. En segundo lugar, el preprocesamiento de las mediciones originales mediante filtrado con *wavelets* permite la eliminación de datos anormales en las mediciones, lo que conduce a una mejora en la precisión de la estimación. Todos los anteriores beneficios, se muestran claramente al hacer la comparación entre la estrategia EPA y la estrategia integrada de filtrado con *wavelets* más EPA (WEPA).

Se introdujo una extensión del método propuesto (WEPA) para el tratamiento de situaciones dinámicas no lineales. Esta extensión consiste en representar cada uno de los términos individuales en las ecuaciones del modelo del proceso considerado como una nueva variable, lo que provoca la conversión de las ecuaciones no lineales del modelo del proceso a combinaciones lineales de las nuevas variables. Luego, se puede aplicar WEPA con este nuevo modelo lineal como restricción. La aplicación del método mostró resultados alentadores: el modo en que se abordan las no linealidades es relativamente eficiente e involucra cálculos sencillos y rápidos.

Se probaron 2 esquemas de ventana móvil para aplicar la estrategia WEPA: el Esquema OBOR y el esquema CR. En ambos casos, la estrategia propuesta, al compararla con otras alternativas como DDR basado en Filtros de Kalman, se muestra competitiva en términos de

precisión de la estimación. Adicionalmente, la estrategia se mostró especialmente adecuada para su aplicación en línea sobre casos reales de baja complejidad (8-10 variables de proceso involucradas). No obstante, se deben hacer evaluaciones adicionales sobre casos de igual y mayor complejidad de modo de poder garantizar, de modo más general, la aplicabilidad en tiempo real, del método propuesto.

Ante la ocurrencia de cambios bruscos, se vio que era mejor aplicar *wavelets* en lugar de cualquiera de las estrategias DDR utilizadas dado que la *wavelets*, al aplicarse individualmente sobre cada variable, es capaz de seguir la discontinuidad en la variable inicialmente afectada por ella sin poder influir en el estimado individual de las otras variables que también serán correctamente atrapados por la *wavelets*. Luego, para situaciones donde se presente un cambio brusco se propone el uso de filtrado con *wavelets* y, una vez se tengan suficientes datos posteriores al cambio ocurrido se puede volver a aplicar la DDR propuesta. Alternativamente se podría pensar en reconciliar directamente los coeficientes *wavelets* y no los polinomiales basados en las tendencias para ver si se mejora la repuesta de la DDR propuesta ante situaciones drásticas como la anteriormente descrita. Esto será abordado en futuros trabajos.

NOMENCLATURA

$\boldsymbol{Ci}$	Matrices del sistema.
$\mathbf{f}()$	Conjunto de restricciones mediante ecuaciones diferenciales.
$\mathbf{g}()$	Conjunto de restricciones de desigualdad.
$\mathbf{h}()$	Conjunto de restricciones mediante ecuaciones algebraicas.
I	Número de variables de proceso.
J	Número de variables de estado.
m	Variable medida genérica.
$\hat{m}$	Estimación de una variable genérica.
n	Número de observaciones en $\mathbf{y}(k)$.
nc	Longitud de la ventana (horizonte de tiempo) móvil.
p	Grado del polinomio.
p_{i}	Grado del polinomio ajustado a la i-ésima variable de proceso.
p_{j}	Grado del polinomio ajustado a la j-ésima variable de estado.
$p^{\max}$	Grado del polinomio más alto entre las variables analizadas.
$\hat{\mathbf{Q}}$	Matriz covarianza de las variables de proceso estimadas.
k	tiempo de muestreo ($k = t$)
$\hat{\mathbf{u}}$	Estimados de las variables de entrada.
$\mathbf{w}$	Variable de proceso ficticia
$\hat{\mathbf{w}}$	Estimado de una variable de proceso ficticia
$\hat{\mathbf{x}}(k)$	Estimados de las variables de estado.
$\mathbf{x}(k)$	Estimados de las variables de estado.
$\mathbf{y}(k)$	Variable de proceso medida (mediciones afectadas por ruido).
$\hat{\mathbf{y}}(k)$	Estimado de la variable de proceso.

$\mathbf{y}^*(k)$ Variable de proceso real (no contaminada con ruidos).
$\hat{\mathbf{y}}^*(k)$ Estimado de la variable de proceso real.
$\mathbf{z}(k)$ Variable ficticia que representa los términos individuales de las ecuaciones del modelo de proceso utilizado.

LETRAS GRIEGAS

$\hat{\alpha}_{i,r}$ r-ésimo coeficiente del polinomio estimado para la variable de proceso $\mathbf{y}_i$.

$\hat{\theta}$ Estimados de los parámetros del modelo.

$\hat{\sigma}_y$ Desviación estándar de las variables de proceso **y**.

$\hat{\sigma}_w$ Desviación estándar de las variables ficticias **z**.

$\hat{\xi}_{j,r}$ r-ésimo coeficiente del polinomio estimado para la variable ficticia $\mathbf{w}_j$.

$\mathbf{\Lambda}$ Matriz de coeficientes polinómicos de $\hat{\mathbf{y}}$ en orden decreciente de tiempo.

$\mathbf{\Xi}$ Matriz de coeficientes polinómicos de $\hat{\mathbf{w}}$ en orden decreciente de tiempo.

SUPERÍNDICES

tr Tendencia de variable filtrada.

p grado del polinomio.

SUBÍNDICES

i Variable de proceso.

j Término derivativo (variable de estado).

c Tiempo o instante actual.

r *r*-esimo coeficiente polinomial

ACRÓNIMOS

ECM Error Cuadrático Medio.
EPA Estrategia Polinomial Extendida.
EPAM Error Porcentual Absoluto Medio.
FK Filtro de Kalman.
IQP Industria de Procesos Químicos.
PNL Programación No Lineal.
RD Reconciliación de Datos.
RDD Reconciliación de Datos Dinámica.
QSS *Quasi Steady State* o condiciones de trabajo cercanas al estado estacionario.
WEPA *Wavelets con EPA.*

CAPÍTULO 4. MONITORIZACIÓN DE SITUACIONES CON PERTURBACIONES LARGAS

RESUMEN
En este capítulo se presenta un análisis comparativo entre estrategias que combinan la filtración de datos con el Análisis de Componentes Principales (ACP). Las estrategias analizadas se enfocan a la monitorización, especialmente para casos donde se produzcan anormalidades pequeñas o de aparición lenta (perturbaciones largas). La comparación incluye métodos existentes en la literatura y algunas combinaciones de filtración basada en *wavelets* con ACP, utilizando los resultados que se proponen en el capítulo 2. La comparación se mide en tiempos de respuesta para detectar una anormalidad. Adicionalmente, incluye una discusión sobre la definición de los límites asociados a los gráficos de control que se generan a través de la monitorización con ACP para obtener mejores respuestas de detección. Se utilizan 4 casos de estudio académicos e industriales para ser monitorizados: Un reactor continuo, un reactor de polimerización, una planta con reciclo y la planta Tennesse Eastman. Los resultados muestran que la combinación de la filtración *Levashrink* con el ACP proporciona resultados de monitorización superiores a la mayoría de las estrategias actuales, para escenarios diversos que incluyan la ocurrencia de anormalidades pequeñas o de perturbaciones largas y con la ventaja de una disminución de la generación de falsas alarmas.

4.1 Introducción

Como se discute en el capítulo 1 (sección 1.3.2), las Técnicas Estadísticas Multivariables (TEM) se han mostrado como herramientas potencialmente útiles para el análisis y supervisión de procesos. Dentro de las TEM, el ACP ha sido la técnica más estudiada en investigación y la que ha exhibido resultados más prometedores en aplicaciones industriales (Qin, 2003; MacGregor, 2004).

4.1.1 Análisis de Componentes Principales

La idea básica del ACP es extraer un nuevo conjunto de variables combinadas, llamadas componentes principales, que describen las variaciones y tendencias clave en los datos de operación en un espacio de dimensiones más pequeño que el original. Supóngase que se dispone de un conjunto de datos de operaciones sin fallos de proceso **Y** de dimensiones *m*x*n*, donde *m* es el número de muestreos disponibles y *n* es el número de variables medidas. Además, se asume que los datos de las *n* variables en Y se han estandarizado previamente[a] y que la distribución de los datos en **Y** se aproxima a una distribución normal. Luego, se puede obtener la siguiente descomposición:

$$\mathbf{Y} = \mathbf{t}_1 \cdot \mathbf{p}_1^T + \mathbf{t}_2 \cdot \mathbf{p}_2^T + ... + \mathbf{t}_n \cdot \mathbf{p}_n^T = \sum_{i=1}^{n} \mathbf{t}_i \cdot \mathbf{p}_i^T \qquad (4.1)$$

o

$$\mathbf{Y} = \mathbf{T} \cdot \mathbf{P}^T \qquad (4.2)$$

[a] La estandarización no es más que llevar todas las escalas de medida a una escala común de media 0 y varianza 1. Esto se hace restando a cada variable su media y luego dividiendo el resultado por su desviación estándar.

T es la llamada matriz de *scores*, de dimensiones *m*x*n*. Sus columnas, representadas por $\mathbf{t}_i$ de *m*x1, se conocen como las variables coordenadas o variables *scores* y representan las coordenadas de los datos al proyectarse en el nuevo espacio de los Componentes Principales (CP). **P** es la matriz de *loadings*, de dimensiones *n*x*n*, cuyas columnas las forman los diferentes $\mathbf{p}_i$. Los $\mathbf{p}_i$, conocidos como cargas o *loadings* y de dimensión *n*x1, son los vectores propios que se derivan de la matriz de correlación de las variables en **Y**; vectores unitarios que indican las direcciones de los CP y forman una base ortonormal en $\mathbb{R}^n$. Luego, se cumple que $\mathbf{P}^T \cdot \mathbf{P} = \mathbf{P} \cdot \mathbf{P}^T = 1$ y $\mathbf{P}^T = \mathbf{P}^{-1}$. La matriz **P** y la descomposición aportada por el ACP (ecuaciones 4.1 ó 4.2), se suelen determinar mediante un método conocido como Descomposición en Valores Singulares o DVS (Jackson, 1991). Durante la resolución del ACP, también se obtiene una matriz diagonal **Λ** que contiene en su diagonal los valores propios λ_i, asociados a cada $\mathbf{p}_i$, ordenados en orden decreciente $\lambda_1 > \lambda_2 > \ldots > \lambda_n$.

El nuevo conjunto de variables obtenido con el ACP tiene la particularidad de que concentra la máxima varianza o variabilidad del proceso (asociada por tanto a las variaciones de causa conocida) en un número de componentes principales que es significativamente menor al número de variables medidas, lo cual hace que la información del proceso, útil para tareas de análisis y supervisión, sea mucho más fácil de manejar sobre todo si el proceso en estudio involucra muchas variables. Así, el modelo en las ecuaciones (4.1) ó (4.2) puede rescribirse como:

$$\mathbf{Y} = \mathbf{T}_a \cdot \mathbf{P}_a^T + \mathbf{E} = \hat{\mathbf{Y}} + \mathbf{E} \qquad (4.3a)$$

$$\hat{\mathbf{Y}} = \mathbf{T}_a \cdot \mathbf{P}_a^T \qquad (4.3b)$$

En las ecuaciones anteriores, el sub-índice *a* es el número de CP que retienen la máxima variabilidad mientras que **E** es la matriz de residuos del modelo obtenido. La dimensión de las matrices $\mathbf{T}_a$ y $\mathbf{P}_a$ queda definida como *m*x*a* y *n*x*a* respectivamente. $\hat{\mathbf{Y}}$ representa la reconstrucción de los datos y se puede considerar como el modelo aproximado del proceso operando en condiciones normales. Su definición solo depende de la selección adecuada de los *a* componentes principales. En la literatura actual se cuenta con varios métodos bastante fiables para la selección de los CP, siendo los más usados el Porcentaje de Varianza Explicada y la selección basada en validación cruzada o *cross-validation* (Jackson, 1991; Wise y Gallagher, 1996).

4.1.2 Monitorización con ACP

Una vez establecido el modelo del proceso (ecuación 4.3b), este se puede utilizar para monitorizar. Para hacer esto primero se calculan los siguientes estadísticos:

- Estadística de Hotelling o T^2:
  Si el ACP se obtiene a partir de datos históricos de operaciones sin fallo el modelo resultante, según la ecuación 4.3, representará un Modelo del Proceso Bajo Control (MP-BC). Luego, se puede inferir el estado actual del proceso a través del cálculo de la T^2, que no es más que la distancia mahalanobis[b] entre los nuevos datos, proyectados en el espacio de los CP, y el MP-BC. Para calcular esta desviación se toma en cada

[b] La distancia de mahalanobis es una generalización de la distancia euclidiana entre 2 vectores en la que se tiene en cuenta la dispersión de las variables y su dependencia. Un valor alto de la distancia de Mahalanobis indica que el punto se aleja del centro de la nube.

instante k en que se muestrea el proceso, el correspondiente vector de mediciones $\mathbf{y}(k)$, de dimensión 1xn, y se proyecta en el espacio de los CP como sigue:

$$\mathbf{t}(k) = \mathbf{y}(k) \cdot \mathbf{P}_a \tag{4.4}$$

En la ecuación anterior $\mathbf{t}(k)$ son los *scores* representando la proyección de $\mathbf{y}(k)$ en el instante k. Luego, T^2 se determina como sigue:

$$T^2 = \mathbf{t}^T(k) \cdot \mathbf{\Lambda}_a^{-1} \cdot \mathbf{t}(k) \tag{4.5}$$

donde $\mathbf{\Lambda}_a$ es la matriz diagonal de valores propios λ_i asociados a los $\mathbf{p}_i$ de los a CP seleccionados.

- Error de predicción al cuadrado o *SPE*:
  El *SPE* (*Squared Prediction Error*), también conocido como *Q*, se define como la suma del error al cuadrado entre la señal original y la señal reconstruida después de la reducción que produce el ACP. Se calcula como sigue:

$$\mathbf{e}(k) = (\mathbf{I} - \mathbf{P}_a \cdot \mathbf{P}_a^T) \cdot \mathbf{y}(k) \tag{4.6}$$

$$SPE = \mathbf{e}(k) \cdot \mathbf{e}(k)^T \tag{4.7}$$

  Donde $\mathbf{I}$ es una matriz identidad de nxn y $\mathbf{e}$ es el error entre los datos y el ajuste por el modelo.

Luego, la monitorización con ACP se efectúa como sigue:

- Con los nuevos datos $\mathbf{y}(k)$ que se recogen en cada nuevo instante k, se obtienen los correspondientes *scores* según la ecuación 4.4.
- Se calcula el error entre los datos recibidos $\mathbf{y}(k)$ y el ajuste por el modelo según la ecuación 4.6.
- Con la información anterior se calculan el T^2 y el *SPE*, según las ecuaciones 4.5 y 4.7 respectivamente.
- Se comparan el T^2 y el *SPE* obtenidos con sus respectivos límites de control (T^2_{lim} y SPE_{lim}) tal que:

$$T^2 \leq T^2_{\text{lim}} \tag{4.8}$$

$$SPE \leq SPE_{\text{lim}} \tag{4.9}$$

  Si alguna o ninguna de las 2 condiciones anteriores se cumplen, entonces el sistema se encuentra fuera de sus condiciones normales de operación. Las comparaciones anteriores (4.8 y 4.9) típicamente se aplican mediante gráficos de monitorización o control donde se observan las señales de T^2 y el *SPE* con sus respectivos límites.

4.1.2.1 Cálculo de los límites de control

El cálculo del límite de control para el *SPE* se ha propuesto en la literatura (Nomikos y MacGregor, 1995) como sigue:

$$SPE_{\lim} = \theta_1 \left[\frac{c_\upsilon \sqrt{2\theta_1 h_0}}{\theta_1} + 1 + \frac{\theta_2 \cdot h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \tag{4.10}$$

$$\theta_i = \sum_{j=a+1}^{n} \lambda_j^i \quad , con \quad i = 1,2,3 \tag{4.10b}$$

$$h_0 = 1 - \frac{2\theta_1 \theta_2}{3\theta_3} \tag{4.10c}$$

Donde c_υ es el valor crítico de la desviación normal estándar al límite de confianza υ deseado y los λ son valores propios contenidos en la diagonal de la matriz $\mathbf{\Lambda}$ y corresponden a los (n-a) valores propios no considerados en el modelo de los CP. Para el caso de los valores de T^2, se asume que se aproximan a una distribución F (Jackson, 1991) y su límite se calcula como sigue (Nomikos y MacGregor, 1995):

$$T^2_{\lim} = \frac{a(m-1)}{m-a} F_{\upsilon(a,m-a)} \tag{4.11}$$

Donde m es el número de muestras utilizados para obtener el modelo inicial, a es el número de CP que retienen la máxima variabilidad y F_υ es el valor crítico de distribución F a un nivel de confianza α y con a y m-a grados de libertad.

Adicionalmente, se pueden usar gráficos de los *scores*, individuales o combinados, como gráficos de monitorización y para el caso de diagnosis de fallos se puede añadir el uso de otros estadísticos basados en los componentes principales como las contribuciones al *SPE* y al T^2.

De manera similar, se puede llegar a establecer una metodología de monitorización usando alguna técnica multivariable similar al ACP como es el caso de la técnica Mínimos Cuadrados Parciales o MCP (mejor conocida como *PLS* o *Partial Least Squares*), que se aplica para casos en que es imprescindible relacionar las variables de proceso con las de calidad (principalmente procesos discontinuos) y predecir el valor de estas últimas en tiempo real (Davis *et al*., 2000).

4.1.3 Monitorización de anormalidades de lenta aparición

A pesar de que las primeras propuestas y aplicaciones ya pusieron de relieve el potencial de monitorización del ACP (ver sección 1.3.2), se ha reconocido que el ACP clásico adolece de ciertas incapacidades para manejar correctamente algunas características del proceso o de los fallos que se puedan presentar.

Una de estas incapacidades, discutida por varios autores en la literatura (Wold, 1994; Wachs y Lewin, 1999; Chen *et al*., 2001), se refiere al hecho de que el ACP no es capaz de detectar rápida ni exactamente la aparición de ciertos fallos de proceso con perturbaciones suaves y largas como un decaimiento suave en las condiciones de operación (por ejemplo, un ensuciamiento de un equipo). En estos casos, existe una dependencia significativa entre el comportamiento del punto de trabajo actual del proceso y los puntos de trabajo recientes (Wold, 1994; Chen *et al*., 2001) y, no obstante, el ACP asume independencia de los puntos de trabajo. Asimismo, ante la presencia de perturbaciones como un salto leve en las condiciones de operación (por ejemplo, debidas a la aparición de una fuga muy leve en una bomba de

alimentación a un reactor) el ACP no es capaz de detectar estos cambios si la magnitud del cambio no es muy alta.

4.1.3.1 Combinación de filtrado con ACP

Como una alternativa para mejorar la respuesta del ACP frente a este tipo de anormalidades, se ha propuesto el modelar los datos, frente a situaciones como las anteriores, aplicando un filtrado inicial a las variables de proceso y luego utilizando estos filtrados con el ACP.

- Propuestas iniciales
  En una de las primeras propuestas se utiliza ACP sobre los datos originales y, luego, se procesan los *scores* del modelo ACP reducido con un filtro de suavizado exponencial o *EWMA*, resultando el *EWMA*-ACP (Wold, 1994). En otro trabajo se propone el uso de una estadística llamada *s-summed*, lo que da lugar al *summed-scores con ACP o Summed-ACP* (Wachs y Lewin, 1999). El *s-summed* consiste en una variante al filtro de media móvil, que intenta un estimado similar al del *EWMA* pero sin la complejidad asociada a la determinación del parámetro de suavización del *EWMA* (Wachs y Lewin, 1999). En cada uno de los trabajos anteriores, los métodos propuestos muestran una mayor resolución y un menor tiempo de respuesta en la detección de las perturbaciones discutidas cuando se comparan a la monitorización con el ACP tradicional. No obstante y como lo resaltan Chen et al., (2001) tanto en el *EWMA*-ACP como en el *summed*-ACP se nota la ausencia de definición de límites de control apropiados para usar con estadísticos derivados de estos métodos.

- La combinación MEWMA-ACP
  Este método fue propuesto como una variante al método *EWMA*-ACP (Chen *et al*., 2001). Primero, se utiliza el *MEWMA*, que no es más que el *EWMA* (ver sección 2.1.1.1.1) en forma matricial para obtener un filtrado simultáneo de las variables y usando un mismo valor de la constante de suavización α para todas las variables, como sigue:

$$\mathbf{z}(k) = \alpha \cdot \mathbf{y}(k) + (1-\alpha) \cdot \mathbf{z}(k) \tag{4.12}$$

  Sobre la matriz de datos filtrados **z** se aplica el ACP. Luego, aplicando diversas transformaciones (Chen *et al*., 2001) se llega a que el *SPE* **t** y los *scores* se computan según las ecuaciones 4.6 y 4.7 respectivamente, mientras que el T^2 se calcula como sigue:

$$T^2 = \left[\mathbf{z}(k) \cdot \mathbf{P}_a \cdot \left(\frac{\alpha}{2-\alpha} \cdot \mathbf{\Lambda}_a \right)^{-1} \cdot \mathbf{P}_a^T \cdot \mathbf{z}^T(k) \right] = \left[\mathbf{z}(k) \cdot \mathbf{P}_a \cdot \mathbf{\Lambda}_z^{-1} \cdot \mathbf{P}_a^T \cdot \mathbf{z}^T(k) \right] \tag{4.13}$$

  Donde $\mathbf{\Lambda}$ esta asociada a la matriz covarianza de los datos originales y $\mathbf{\Lambda}_Z$ esta asociada a la matriz covarianza de los datos filtrados. Según esta propuesta (Chen *et al*., 2001) los límites de control para el T^2 se calculan de la misma manera que en el ACP tradicional, esto es, basado en una distribución *F*, mientras que en el caso de *SPE* proponen una corrección a la ecuación 4.10 como sigue:

$$(SPE_{\lim})_{MEWMA-ACP} = \frac{\alpha}{2-\alpha} \cdot (SPE_{\lim})_{ACP} \tag{4.14}$$

- La combinación de s-summed con ACP
  Este método, identificado como *MSSUM*-ACP, se propuso en paralelo al *MEWMA*-ACP y con un propósito similar: mejorar la aplicabilidad del s-summed con ACP

(Chen *et al*., 2001). En este caso, primero se crea una matriz basada en la suma acumulada de *s* observaciones pasadas como sigue:

$$\mathbf{s}(k) = \sum_{i=k-s+1}^{k} \mathbf{y}(i) \tag{4.15}$$

Sobre la matriz de datos **s** se aplica el ACP. De manera similar al método anterior, aplicando diversas transformaciones (Chen *et al*., 2001) se llega a que los *scores* **t** y el *SPE* se computan según las ecuaciones 4.6 y 4.7 respectivamente, mientras que el T^2 se calcula como sigue

$$T^2 = \mathbf{s}(k) \cdot \mathbf{P}_a \cdot (s \cdot \mathbf{\Lambda}_a)^{-1} \cdot \mathbf{P}_a^T \cdot \mathbf{s}^T(k) = \mathbf{s}(k) \cdot \mathbf{P}_a \cdot \mathbf{\Lambda}_S^{-1} \cdot \mathbf{P}_a^T \cdot \mathbf{s}^T(k) \tag{4.16}$$

Donde $\mathbf{\Lambda}_S$ está asociada a la matriz covarianza de los datos convertidos mediante la ecuación 4.15. También en este caso, la propuesta (Chen *et al*., 2001) establece que los límites de control para el T^2 se calculan de la misma manera que en el ACP tradicional, esto es, basado en una distribución *F*, mientras que en el caso de *SPE* proponen una corrección a la ecuación 4.10 como sigue:

$$(SPE_{\lim})_{\text{MSSUM-ACP}} = s \cdot (SPE_{\lim})_{ACP} \tag{4.17}$$

En el trabajo de *Chen et al.* (2001), los métodos *MEWMA*- ACP y *MSSUM*- ACP se muestran superiores en capacidad de detección tanto al *EWMA*- ACP como al *Summed*-ACP. Particularmente el *MEWMA*- ACP se observa como muy superior al resto.

4.1.3.2 Uso de wavelets

Kosanovich y Piovoso presentaron una estrategia de monitorización en la que filtran las variables originales mediante un Filtro Híbrido basado en la Mediana (*FHM*) para luego procesarlos con *wavelets* (Kosanovich y Piovoso, 1997). Sobre los coeficientes aportados por la *wavelets*, los autores aplican el ACP para monitorizar a diferentes escalas, aunque no establecen claramente el manejo de la información a diferentes escalas, mostrando solamente los gráficos de pares de *scores* sin definir ni utilizar límites de control orientativos para ayudar a establecer la ocurrencia de un fallo. Bakshi (1998)propuso otra metodología basada en la combinación de *wavelets* con ACP, llamada *MsPCA* o *multiscale PCA*, que pretende explotar las capacidades multiescala que aporta el análisis *wavelets* para mejorar el diagnóstico y la detección con ACP. En este caso, se añade una definición de los gráficos *SPE* y T^2 que incluyen límites para detección a distintas escalas (Bakshi, 1998). En su trabajo Bakshi (1998) incluye un caso de estudio donde consigue la detección de una perturbación de aparición lenta pero con bastante retraso respecto del tiempo real de aparición de dicha perturbación. Luego, la discusión final de los resultados la orientan hacia la detección de otro tipo de perturbaciones sin concluir sobre la capacidad real del *MsPCA* para el manejo de perturbaciones de aparición lenta.

4.2 Métodos combinados de Filtrado wavelets con ACP para monitorización

Vistas las propuestas existentes para monitorizar eventos anormales de aparición lenta y suave, en esta sección se propone una estrategia similar a las ya existentes pero en las que se explotará el filtrado con *wavelets* según los desarrollos del capítulo 2.

La combinación que se propone es como sigue: primero se corrigen los datos con un filtro *wavelets* apropiado y luego se aplica el ACP sobre los datos filtrados. La figura 4.1 ilustra el esquema de combinación a usar.

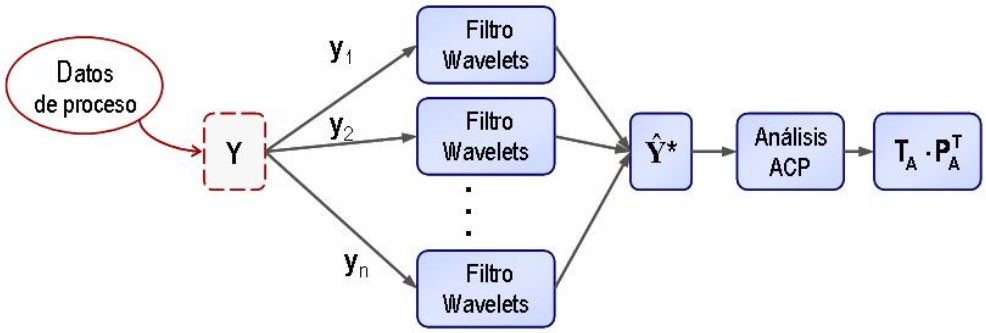

Figura 4.1. Esquema de combinación Filtrado-*wavelets* más ACP.

Donde $\hat{\mathbf{Y}}^*$ es una matriz con los filtrados de las mediciones de las variables de proceso y contenidas en **Y**. Para la filtración, se propone la evaluación de varias estrategias basadas en *wavelets*:

- Estrategia 1: Filtrado *waveshrink* con *thresholding soft* y utilizando como función la *wavelets Haar* o *db1*.
- Estrategia 2: Filtrado *levashrink* con *thresholding soft* y utilizando como función la *wavelets Haar* o *db1*.
- Estrategia 3: Filtrado basado en la combinación de varias *wavelets dbN*. Las funciones a combinar son la *wavelets db1*, *db4* y *db8*.

La Estrategia 1 se toma por que es usada en un trabajo previo de monitorización con ACP (Kosanovich y Piovoso, 1997). No obstante, en este caso se trabaja con la señal reconstruida tras el filtrado y no con los coeficientes de las *wavelets*. Así, se evita el tener que trabajar con un número de señales mayor al original, ya que cada escala a la que se aplique la descomposición con *wavelets* se obtienen distintos vectores de coeficientes. La estrategia 2 se selecciona de acuerdo a los resultados obtenidos en el capítulo 2. La estrategia 3 se toma para evaluar la propuesta de rectificación combinada que también se propuso en el capítulo 2. Una vez se obtiene la matriz de filtrados con cualquiera de las estrategias propuestas (1 a 3), se aplica el ACP sobre dichos datos.

4.3 Análisis Comparativo de detección de fallos para Monitorización

En esta sección se presenta un estudio comparativo entre diversas estrategias descritas en la sección anterior. Cada estrategia se utiliza para monitorizar una serie de diferentes escenarios de proceso que incluyen operaciones en condiciones normales y operaciones bajo condiciones anormales. Se considera no solo la detección de anormalidades asociadas a perturbaciones de aparición lenta, sino también a fallos más bruscos del tipo escalón ya que, aunque se considere que un proceso se pueda ver afectado frecuentemente por perturbaciones lentas, éste también estará inevitablemente sujetos a otro tipo de perturbaciones. Así, los métodos a evaluar deben ser capaces de manejar distintos tipos de anormalidades.

En primer lugar, se establecen una serie de criterios y herramientas que sirven para llevar a cabo el análisis. Luego, se describen los casos de estudio utilizados. Finalmente, se presenta el análisis de resultados.

4.3.1 Criterios y herramientas para la detección y para la comparación de las detecciones

4.3.1.1 Límites de control basados en la distribución empírica de los datos

Los límites para los gráficos de control *SPE* y T^2 (ver ecuaciones 4.10 y 4.11) tradicionalmente se construyen bajo la suposición de que estos estadísticos obedecen a distribuciones específicas: la distribución *F* para el caso del T^2 y la normal para el *SPE* (Nomikos y MacGregor, 1995). En una fase inicial de este trabajo (Musulin *et al*., 2002) se observó que los límites obtenidos de esta forma conducen a frecuentes errores en la detección. Esto se debe a que en la práctica y para diferentes procesos, los valores obtenidos de los estadísticos no necesariamente cumplirán la distribución a la que teóricamente deberían aproximarse.

Luego, se adopta como alternativa la estrategia propuesta en Musulin, Tona, Ruiz, Espuña y Puigjaner de utilizar límites basados en la distribución empírica de los datos o límites BDE (Musulin *et al*., 2002). Para obtenerlos basta con aplicar el siguiente procedimiento:

- Calcular para cada estadístico su función empírica de distribución acumulada.
- Sobre la curva de distribución obtenida fijar un límite tal que la probabilidad de un porcentaje *xy* de los valores de los estadísticos permanezca por debajo de dicho límite. Los valores típicos para *xy* serán 95 ó 99.

Tras lo anterior se tendrá:

$$(CL_{\lim})_{\,Estrategia-i} = CL_{\lim-BDE} \tag{4.18}$$

Donde *CL* indica cualquiera de los estadísticos usados (*SPE* o T^2) y $CL_{lim\text{-}BDE}$ son los límites obtenidos mediante la distribución empírica de los *CL*.

Más adelante, en la sección 4.3.4, se ilustra el beneficio de adoptar estos límites para la monitorización. Así, se ofrece una evidencia experimental y comparativa sobre esta recomendación que en trabajos precedentes (Musulin *et al*., 2002; Lu *et al*., 2003) se propuso y se adoptó sin dejar una evidencia.

4.3.1.2 Corrección por cada escala

En el trabajo de Bakshi (1998) donde se combina *wavelets* con ACP, se calculan los límites a múltiples escalas, según las expresiones 4.10 y 4.11. Sin embargo, cada límite en cada escala lo multiplican por un factor de corrección que ajusta el límite fijado a las variaciones de causa común en cada escala. Este factor de corrección también se utiliza en este trabajo tal que la ecuación 4.18 se modifica como sigue:

Para el caso de la combinación *waveshrink*-ACP:

$$(CL_{\lim})_{\,waveshrink-ACP} = 100 - \left(\frac{1}{(L-1)} \cdot (100 - CL_{\lim-BDE}) \right) \tag{4.19}$$

Para el caso de la combinación *levashrink*-ACP:

$$(CL_{\lim})_{\,levashrink-ACP} = 100 - \left(\frac{1}{(Lm-1)} \cdot (100 - CL_{\lim - BDE}) \right) \qquad (4.20)$$

La diferencia básica entre 4.19 y 4.20 está en que la *L* (la escala) en la ecuación 4.19 es fijada por el analista, mientras que en 4.20 la *Lm* es la *L* determinada dentro del procedimiento *levashrink* (ver sección 2.3).

4.3.1.3 La regla de detección

La manera tradicional de detectar fallos con la monitorización basada en ACP consiste en tomar en cada instante los nuevos valores calculados de *SPE* y T^2, proyectarlos en los gráficos de control de dichos estadísticos y comprobar si superan los límites de control en dichos gráficos. Luego, si durante *k* periodos consecutivos (típicamente *k*=3) se produce una desviación se genera una alarma de fallo (figura 4.2). Aunque esta no es la regla universal se utiliza ampliamente, y por tanto es también la que se utilizará en este capítulo para todos los métodos.

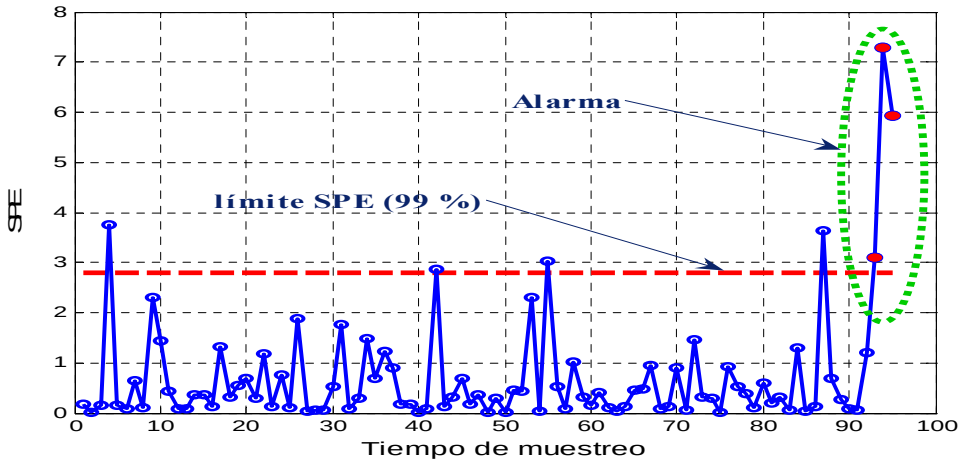

Figura 4.2. Esquema de combinación filtrado-*wavelets* más ACP.

4.3.1.4 Evaluación comparativa entre métodos

Para poder comparar la detección entre distintos métodos se utiliza la siguiente estrategia de comparación:

- Se toman los tiempos ti_{EST} donde el sistema de monitorización utilizado indica una alarma de detección. ti_{EST} puede representar a ti_{SPE} (tiempo de detección usando el *SPE*) o ti_{T2} (tiempo de detección usando el T^2).

- Se toma la diferencia entre el tiempo en que se detecta (ti_{EST}) y el tiempo real en que se sabe que se produce la perturbación (t_{REAL}):

$$t_{EST} = ti_{EST} - t_{REAL} \qquad (4.21)$$

  En este caso t_{SPE} y t_{T2} representan el tiempo que se ha tardado en detectar un fallo bien sea mediante los *SPE* o mediante los T^2 obtenidos con un método de monitorización dado.

- De cada par t_{SPE} - t_{T2} se escoge el valor más pequeño y se asigna a la variable *Td*. Luego, la variable *Td* representa el tiempo mínimo de detección con un método dado (*Td*).

$$Td = \min(t_{SPE}, t_{T2}) \tag{4.22}$$

- Tras estimar los *Td* de cada método de monitorización, se pueden tabular para compararlos entre sí y poder establecer el método que ofrece la detección más rápida.

Adicionalmente, se propone en este trabajo un índice de detección *Id* con el que se intenta puntuar la capacidad de detección de un método. Con dicha puntuación se pueden categorizar los métodos para establecer cual es el que ofrece una mejor detección. El índice propuesto se obtiene mediante la siguiente expresión:

$$Id = \begin{cases} 1 & , \quad Td \le 3 \\ 3/Td & , \quad Td > 3 \end{cases} \tag{4.23}$$

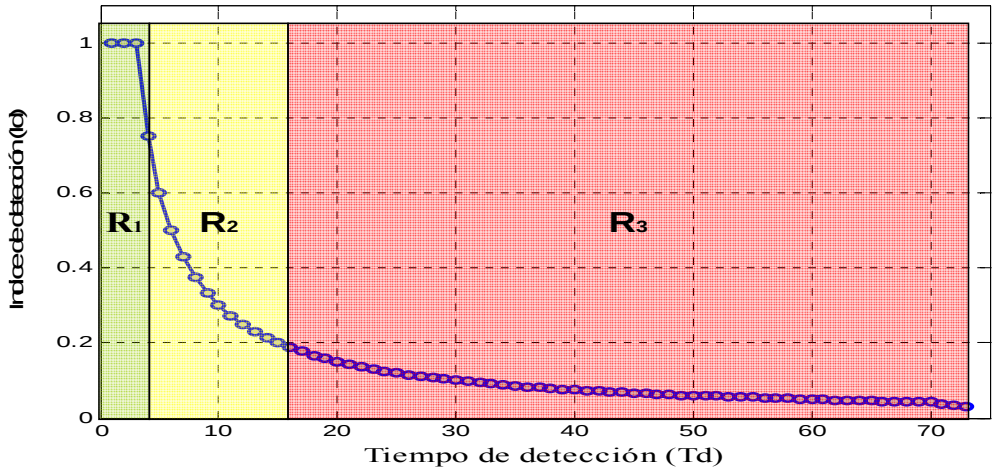

Figura 4.3. Gráfico del índice de detección *Id* en función de los *Td*.

El parámetro *Td* sigue siendo el mismo que se muestra en la ecuación 4.22. En el gráfico de la figura 4.3 se muestra una curva de valores de *Id* en función de los tiempos de detección (que podrían venir medidos en cualquier unidad de tiempo) y se ilustra su interpretación. Se establecen 3 regiones de valores *Td* que corresponden a diferentes niveles de calidad *Id*.

- Región $\boldsymbol{R_1}$. Incluye los valores de *Td* entre 1 – 4. Para ellos, los correspondientes *Id* toman la más alta puntuación. Así, *Id* máximo indican detecciones rápidas.
- Región $\boldsymbol{R_2}$. Incluye los valores de *Td* entre 5 – 16. En este caso los valores de *Id* indican detecciones aceptables, a medida que se corresponden *Td* con un cercano a 5 y que posiblemente se podrían mejorar bien sea por una revisión del método aplicado o por adoptar algún otro método para la detección.
- Región $\boldsymbol{R_3}$. En este caso los valores de *Id* se asocian a detecciones considerablemente tardías (*Td* > 16) lo que sugiere que el sistema de monitorización en uso es insuficiente para monitorizar el tipo de fallo que se ha presentado en el proceso.

4.3.2 Casos de estudio

4.3.2.1 Operación de un Reactor Continuo

Este caso consiste en un modelo de reactor *CSTR*, el cual se describe en la sección C.2 del anexo C. Se asume que durante la operación del reactor (simulaciones) se dispone de mediciones continuas de las siguientes variables de proceso y de producto: Flujo de alimentación al reactor (q_0), Temperatura de alimentación al reactor (T_0), Concentración del reactivo A en el reactor (Ca), Temperatura del reactor (T), Flujo del refrigerante (q_C), Temperatura del refrigerante (T_C), Flujo de salida del reactor (q), y Volumen del reactor (V). Todas estas variables se recuperan fácilmente de la simulación e incluyen la adición durante la simulación de errores aleatorios de medición en las variables. Adicionalmente, el modelo incluye un control de V mediante el flujo de salida q y otro control para la temperatura del reactor T mediante el q_C. Con este modelo se lleva a cabo una simulación del proceso en estado normal y otra que incluye durante su desarrollo la ocurrencia de diversos fallos asociados a las variables de entrada Ca_0 (Concentración de la Alimentación), T_0 (Temperatura de la alimentación) y T_C. Los cambios que se introducen durante la operación se describen en la tabla 4.1.

Tabla 4.1. Descripción de escenarios para el Reactor Continuo.

Escenario	Operación	Descripción	toe(min.)	tts (min.)
E1.1	Normal	-	-	126
E1.2	Anormal	Anormalidad 1	17	222
E1.3	Anormal	Anormalidad 2	56	222
E1.4	Anormal	Anormalidad 3	110-120	222
E1.5	Anormal	Anormalidad 4	165	222

En la tabla anterior:

- La Anormalidad 1 consiste en un desajuste en la válvula del caudal de entrada que provocó un salto del caudal de 40 a 42 lt/min.
- La Anormalidad 2 consiste en un desajuste en la válvula del caudal de entrada que provocó un salto del caudal de 40 a 40.8 lt/min.
- La Anormalidad 3 consiste en un desajuste intermitente y progresivo de T_{C0}.
- La Anormalidad 4 consiste en variaciones escalonadas (no continuas) en Ca_0 (esta es una variable no medida).

Asimismo, *toe* representa el tiempo real en que ocurre o se inicia una perturbación y *tts* es el tiempo total de simulación del escenario indicado.

4.3.2.2 Reactor Continuo Industrial para la producción de Acetato de Polivinilo

Este caso consiste en la simulación de un reactor continuo industrial para la fabricación de Acetato de Polivinilo. El modelo teórico fue introducido en una serie de trabajos (Teymour y Ray, 1992a; Teymour y Ray, 1992b), mientras que el modelo de simulación fue desarrollado por Jeffrey DeCicco (1998), bajo el entorno Simulink del MATLAB. El modelo asociado es relativamente complejo debido al tipo de reacción química que se produce: Una polimerización del Acetato de Vinilo vía radicales libres (DeCicco, 1998).

Tabla 4.2. Descripción de escenarios para el reactor de Polimerización.

Escenario	Operación	Descripción	**toe** (min.)	**tts** (min.)
E2.1	Normal	-	-	161
E2.2	Anormal	Salto brusco de T_{OR} por problema del precalentador de la corriente de alimentación.	51	176
E2.3	Anormal	Apertura leve y continua de válvula de alimentación al reactor (q_0).	23	241
E2.4*a* **E2.4*b***	Anormal	2 Saltos bruscos de T_{OR} por problema del precalentador de la corriente de alimentación.	Salto1-15 Salto2-96	163
E2.5	Normal	-	-	

Se asume que durante la operación del reactor se dispone de mediciones continuas de las siguientes variables de proceso y de producto: Flujo de alimentación al reactor (q_0), Temperatura de alimentación al reactor (T_0), Concentración del iniciador en la alimentación al reactor (CI_{0R}), Concentración del iniciador en el reactor (CI_R), Temperatura del reactor (T), Temperatura del refrigerante (T_{0C}), Polidispersidad (PD). Todas estas variables se recuperan fácilmente de la simulación e incluyen la adición durante la simulación de errores aleatorios de medición en las variables y en el proceso, alcanzándose un gran parecido con la operación real. Con el simulador se proponen 5 escenarios de operación los cuales se resumen en la tabla 4.2. Los escenarios se numeran como siguiendo la numeración de escenarios de los casos anteriores.

4.3.2.3 Planta Química con Reciclo

Este caso de estudio consiste en un ejemplo académico de una planta química continua con reciclo (Belanger y Luyben, 1997). La planta esta formada por 2 unidades principales: Un reactor y una columna de destilación. Al reactor se envía la alimentación fresca, consistiendo de reactante *A* y algo de producto *B*. El reactor es continuo y la reacción que se lleva a cabo es irreversible y de primer orden. La corriente de salida del reactor es enviada a la columna de destilación. Allí, se separa la mayor parte del *A* que no reacciono del producto *B*. La corriente de producto *B*, junto con una pequeña fracción de *A* (X_{AB}), se obtiene por la corriente de fondo de la columna de destilación mientras que la corriente del tope, con la mayor parte del *A* que no reacciono, es recirculada al reactor.

Tabla 4.3. Variables medidas en la Planta con Reciclo.

Variable	Descripción
X_f	Composición de entrada a la columna.
D	Flujo de destilado.
$O_pV_{a\ cond}$	Apertura de la válvula que controla el nivel del condensador.
$HP_{cond\ err}$	Error con respecto al punto de consigna del nivel del condensador.
B	Caudal de salida de la planta.
$OPV_{\ reb}$	Apertura de la válvula que controla el nivel de la base de la columna.
$HP_{reb\ err}$	Nivel de la base de la columna.
B_{up}	Caudal del vapor del reboiler.
$OpV_{a\ boil}$	Apertura de la válvula de vapor del reboiler.
X_{Aerr}	Error con respecto al punto de consigna de la composición de salida.
V_r	Error con respecto al punto de consigna del nivel del reactor.
F	Caudal de salida del reactor.
F_0	Caudal de entrada al reactor = $D+F_{00}$

El esquema de la planta se muestra en la figura 4.4. Para los análisis del presente trabajo se dispone de un simulador de la planta desarrollado previamente en Simulink-MATLAB (Musulin *et al*., 2002; Musulin *et al*., 2004). Se asume que durante las operaciones simuladas, se dispone continuamente de las mediciones descritas en la tabla 4.3.

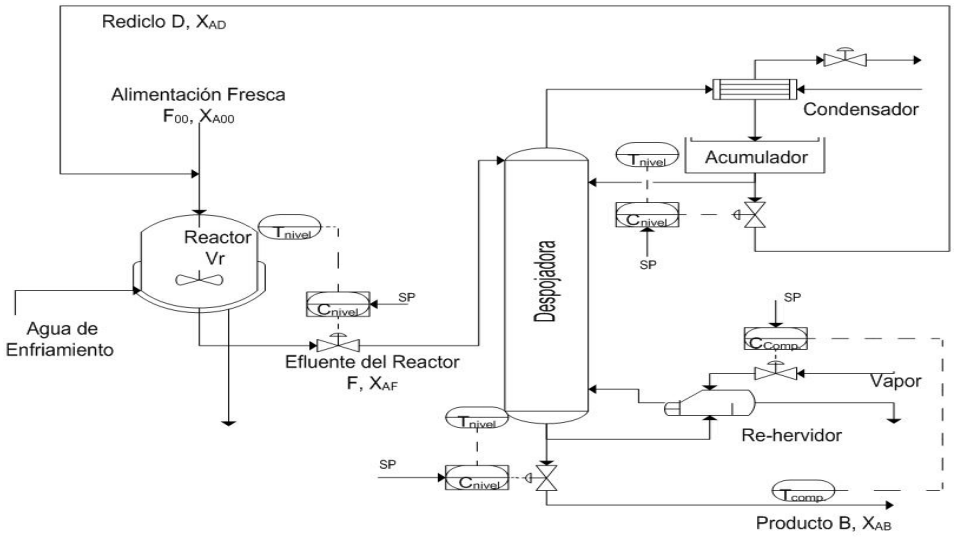

Figura 4.4. Esquema de la Planta con Reciclo

Como en los casos anteriores, se generan varios escenarios de operación, la mayoría de ellos afectados por anormalidades diversas. El detalle de estos escenarios se muestra en la tabla 4.4.

Tabla 4.4. Descripción de escenarios para la Planta con Reciclo.

Escenario	Operación	Descripción	toe(min.)	tts (min.)
E3.1	Normal	-	-	384
E3.2	Anormal	*Drift* en la variable V_R	98	167
E3.3	Anormal	Perdida de B_{UP}	95	175
E3.4	Anormal	Fallo en el sensor F	48	235

4.3.2.4 Proceso Tennessee Eastman

El proceso *Tennesse Eastman* (TE) fue propuesto a la comunidad científica por Downs y Vogel a principios de la década de los 90 (Downs y Vogel, 1993). Consiste en la simulación de una planta real que involucra la producción de 2 productos, G y H, a partir de 4 reactantes A, C, D y E. Adicionalmente ocurren 2 reacciones laterales y la presencia de un inerte B. Todas las reacciones son irreversibles y exotérmicas.

El proceso se compone de 5 unidades de operación principales: (1) Un reactor bifásico y exotérmico, (2) El condensador de producto, (3) Un separador flash, (4) Un desorbedor y (5) Un compresor de reciclo. En la figura 4.5, se muestra un diagrama del proceso. Los gases reactantes se alimentan al reactor donde reaccionan para formar productos líquidos. Para ayudar a la reacción de los gases se utiliza un catalizador no volátil disuelto en el líquido. El calor generado en el reactor se extrae con ayuda de un refrigerante. Los productos, junto con los reactantes que no se consumieron, dejan el reactor en forma de vapor y pasan a través de un enfriador para condensar los productos, y luego a un separador de líquido-vapor. Los componentes no condensados se devuelven al reactor, a través de un compresor centrífugo.

Los componentes condensados se procesan en una columna de separación (desorbedor), para extraer los reactantes remanentes. Los productos G y H que salen del desorbedor, se envían a otra sección de la planta donde se separan. Está última parte del proceso no se considera en el problema propuesto de la TE. Los inertes y los subproductos se extraen del sistema como vapores, a través de una purga en el separador de líquido-vapor. En el modelo del proceso intervienen 41 variables medidas y 12 variables manipuladas. Durante las simulaciones se asume que se dispone de mediciones continuas de cada una de estas 52 variables.

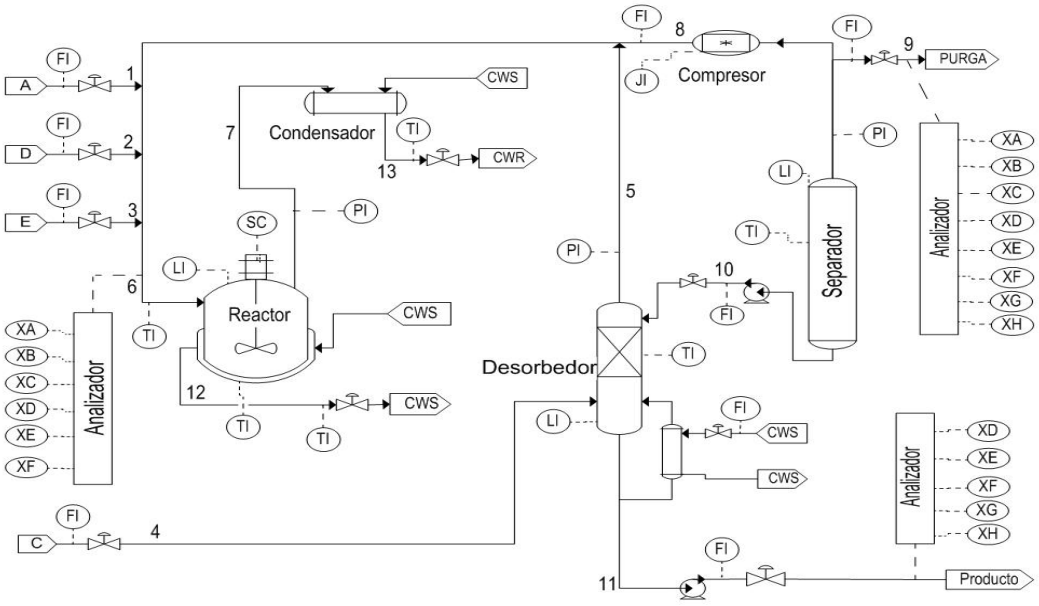

Figura 4.5. Diagrama de flujo del proceso Tennesse Eastman.

En la propuesta inicial de este caso de estudio (Downs y Vogel, 1993), aparece un conjunto de 17 diferentes perturbaciones que podrían afectar al proceso. Para el análisis actual se seleccionan las perturbaciones que se listan en la tabla 4.5.

Tabla 4.5.Descripción de escenarios para la planta con reciclo

Escenario	**Operación**	**Descripción**	**toe** (min.)	**tts** (min.)
E4.1	Anormal	Cambio brusco en la temperatura del agua de enfriamiento que entra al reactor.	60	1115
E4.2	Anormal	Variación aleatoria de la temperatura de la alimentación del reactivo D.	50	1115
E4.3	Anormal	Desajuste de la válvula de agua de enfriamiento al reactor.	50	1115

4.3.3 Realización de experimentos de simulación

Para evaluar el rendimiento de la monitorización con varios de los métodos presentados a lo largo de las secciones 4.1 y 4.2, se llevó a cabo la simulación de la monitorización de cada uno de los escenarios descritos en la sección 4.3.2. Los métodos evaluados se etiquetan según aparecen en la tabla 4. 5.

Tabla 4.5.Descripción de escenarios para la planta con reciclo

Método	Descripción
M1	ACP clásico.
M2	Combinación de *MEWMA* con ACP.
M3	Combinación *MSSUM*-ACP.
M4	Combinación *waveshrink-db1* con ACP.
M5	Combinación *levashrink -db1* con ACP.
M6	Filtro Combinado (*db1-db4-db8*) integrado con ACP.

En una primera etapa se generó un Conjunto de Operación Normal (CON) de cada uno de los casos: Uno asociado al *CSTR*, otro al reactor de polimerización, otro a la planta con reciclo y otro al proceso *Tennesse Eastman*. Cada uno de estos conjuntos CON se procesó con cada método a fin de obtener los *SPE* y T^2 (límites de estos) para utilizar posteriormente en las monitorizaciones. Los límites se calculan de 2 maneras: basados en las formulas que se proponen para cada método (Límites Clásicos descritos en la sección 4.1.2.1) y basados en la distribución empírica de los datos según se explica en la sección 4.3.1.1 (límites BDE). Una vez hecho lo anterior, se procede a la simulación de cada escenario descrito en la sección 4.3.2. La detección con cada método se aplicó según se explica en la sección 4.3.1.3. Todas las implementaciones de los métodos, los casos de estudio y las simulaciones mismas se llevaron a cabo en MATLAB 7.0.

4.3.4 Análisis de resultados. Comparación de límites

En esta sección se muestra la comparación entre las monitorizaciones aplicando límites para el *SPE* y T^2 calculados de 2 formas: a la manera clásica (ver sección 4.1.2.1) y según la distribución empírica de los estadísticos obtenidos (ver sección 4.3.1.1). Dado que en todos los casos de estudio los resultados obtenidos son similares solo se presenta el análisis para los casos de estudio 1 (*CSTR*) y 2 (Reactor de polimerización).

En las tablas 4.7 a 4.10 se resumen los resultados de la monitorización para las simulaciones de los escenarios con operación anormal asociados a los casos 1 y 2. Se muestran los valores del tiempo mínimo de detección *Td* de cada método sobre cada escenario, que reflejan la diferencia de tiempo entre el momento en que se produce la detección y el momento en que el sistema de detección genera la alarma de fallo detectado. La casilla **ODt** indica el orden que ocupa cada método en cuanto a la rapidez de detección. En algunas casillas de *Td* aparece **NF**. Esta indica que no se llegó a detectar el fallo.

Tabla 4.7. Caso *CSTR*-Luyben con Límites clásicos.

	E1.2		**E1.3**		**E1.4**		**E1.5**	
	Td	**ODt**	***Td***	**ODt**	***Td***	**ODt**	***Td***	**ODt**
M1	3	-	**NF**	**NF**	13	4	24	6
M2	3	-	8	3	6	1	17	4
M3	3	-	8	3	8	2	18	5
M4	4	-	5	1	12	5	10	2
M5	4	-	6	2	11	3	4	1
M6	3	-	5	1	14	6	15	3

Tabla 4.8. Caso CSTR-Luyben con Límites BDE.

	E1.2		**E1.3**		**E1.4**		**E1.5**	
	Td	**ODt**	***Td***	**ODt**	***Td***	**ODt**	***Td***	**ODt**
M1	3	-	5	-	11	3	24	6
M2	3	-	5	-	6	1	12	3
M3	3	-	5	-	8	2	12	3
M4	3	-	5	-	8	2	5	2
M5	3	-	6	-	11	3	4	1
M6	3	-	5	-	14	4	15	4

En un primer análisis se comparan las detecciones obtenidas para el caso 1 cuando se utilizan los límites clásicos (tabla 4.7) y cuando se utilizan los límites BDE (tabla 4.8). En el caso del primer escenario (**E1.2**) se observa que los tiempos de detección con cada método y en cada tabla son muy similares y pequeños indicando una detección rápida con cualquiera de los métodos. Debido a esto, la jerarquización **ODt** se ve innecesaria en este caso. Solo se observa una leve mejora en la detección cuando se adoptan los límites BDE de modo que, por ejemplo, ***M4*** y ***M5*** reducen los tiempos de detección en 1 minuto**.** El escenario **E1.3** es similar al anterior. No obstante, se nota una mejora más apreciable en los tiempos de detección de algunos métodos al pasar de utilizar límites clásicos a límites BDE. Así, el ***M1*** con límites clásicos para el *SPE* y el T^2 es incapaz de detectar el fallo (ver la figura 4.6), mientras que utilizando límites BDE (ver la figura 4.7) alcanza la detección en un tiempo equivalentemente tan rápido como el del resto de los otros métodos. De manera similar, se pueden establecer conclusiones semejantes para los casos de ***M2*** y ***M3***.

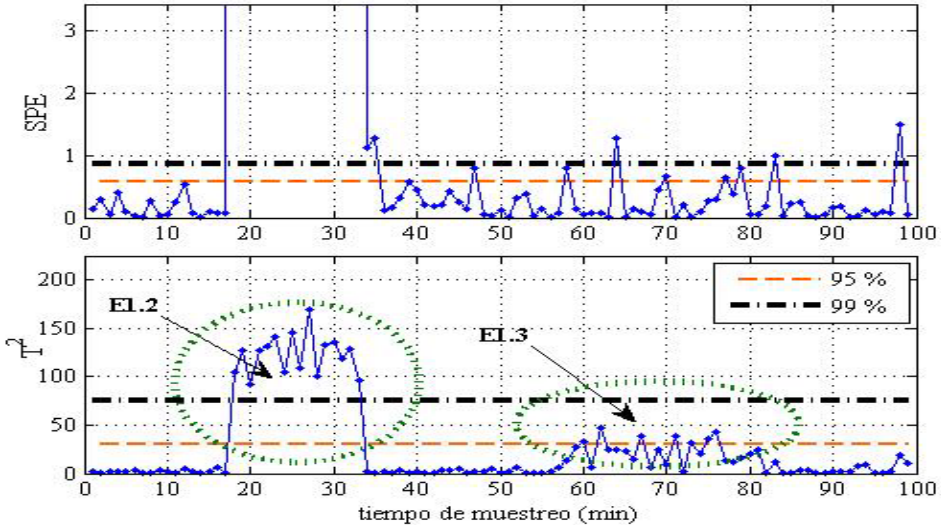

Figura 4.6. Detección con ACP clásico y límites clásicos para E1.2-E1.3.

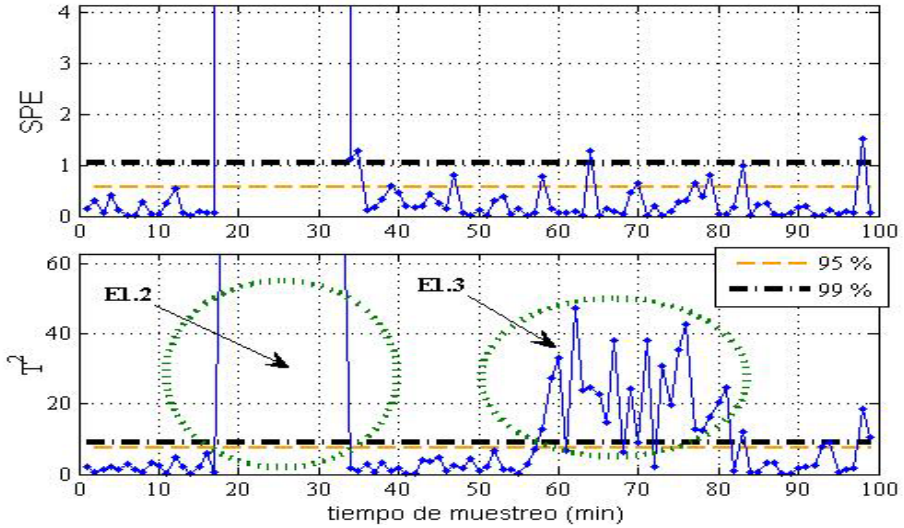

Figura 4.7. Detección con ACP clásico y límites BDE para E1.2-E.13.

Los siguientes escenarios (**E1.4**) y (**E1.5**), muestran la detección de 2 casos con perturbaciones menos bruscas que los casos anteriores. Como en el caso del escenario **E1.3**, la mejora en los tiempos de detección tras la adopción de los límites BDE y para los casos **E1.4** y **E1.5**, también se pueden apreciar aquí pero de una forma más dramática. Por ejemplo, si se considera el caso de ***M2*** aplicado a **E1.5**, se puede ver que los tiempos de detección se reducen desde 27 hasta 12 con el T^2.

Tabla 4.9. Caso reactor de polimerización con Límites clásicos.

	E2.2		**E2.3**		**E2.4*a***		**E2.4*b***	
	Td	**ODt**	***Td***	**ODt**	***Td***	**ODt**	***Td***	**ODt**
M1	12	5	69	5	6	2	5	-
M2	6	2	47	3	6	2	3	-
M3	7	3	42	2	8	3	3	-
M4	9	4	38	1	3	1	4	-
M5	4	1	77	6	3	1	4	-
M6	4	1	54	4	3	1	4	-

Pasando al análisis de las tablas 4.9 y 4.10, se observa que las reducciones en los tiempos de detección tras adoptar los límites BDE son también importantes en muchos casos. Por ejemplo, en el caso del escenario **E2.2**, al monitorizar con ***M1***, los tiempos de detección se reducen de 22 a 3 para el T^2 y de 12 a 3 para el *SPE*, y en el caso de ***M4*** los tiempos se reducen de 13 a 4 para el T^2 y de 9 a 4 para el *SPE*. Se pueden apreciar otros resultados aún más evidente para el caso del escenario **E2.3** y en los casos en que se monitoriza con los métodos ***M1*** a ***M5***.

Tabla 4.10. Caso reactor de polimerización con Límites BDE.

	E2.2		**E2.3**		**E2.4*a***		**E2.4*b***	
	Td	**ODt**	***Td***	**ODt**	***Td***	**ODt**	***Td***	**ODt**
M1	3	2	63	8	4	-	3	-
M2	3	2	35	3	3	-	3	-
M3	3	1	33	1	5	-	3	-
M4	4	3	38	4	3	-	4	-
M5	3	2	34	2	3	-	4	-
M6	4	3	54	7	3	-	4	-

Los resultados anteriores dejan ver claramente las ventajas de basar los límites de control de las estadísticas *SPE* y T^2 en distribuciones empíricas más que en los límites tradicionalmente propuestos en la literatura. En efecto, el tiempo de detección de muchos fallos disminuye en muchos casos. Incluso, en algunos casos donde el tiempo de reducción de la detección es muy considerable (ver caso E1.3 en tablas 4.7 y 4.8 y casos E2.2 y E2.3 en tablas 4.9 y 4.10), se puede decir que con los límites BDE no solo se aumenta la capacidad de detección, sino que también se reduce la posibilidad de trabajar bajo "falsos periodos de tiempo bajo estados de operación sin fallos".

4.3.5 Análisis de resultados. Comparación de la detección entre varios métodos

Los resultados de las detecciones con cada método y para cada escenario se muestran en las tablas 4.7, 4.9, 4.11 y 4.12. Para los resultados de estas tablas se utilizan siempre los límites BDE.

Aún cuando los escenarios monitorizados corresponden a casos de estudio distintos, muchos de ellos representan situaciones anormales similares. En vista de esto, se procedió a reagruparlos según el tipo de perturbación similar como sigue:

- Grupo-Operación 1: Escenarios con perturbaciones tipo salto (**E1.2**, **E1.3**, **E2.2**, **E2.4***a*, **E2.4***b*, **E3.3**, **E4.1**).
- Grupo-Operación 2: Escenarios con perturbaciones tipo incrementos/decrementos suaves y continuos (**E2.3**, **E3.2**) o intermitentes del valor de la variable (**E1.4**, **E1.5**).
- Grupo-Operación 3: Escenarios con perturbaciones tipo variación aleatoria del valor de la variable (**E4.2**, **E4.3**).
- Grupo-Operación 4: Escenarios de operación normal (**E1.1**, **E2.1**, **E2.5**, **E3.1**).

Tabla 4.11. Caso Planta con reciclo con Límites BDE.

	E3.2		**E3.3**		**E3.4**	
	Td	**ODt**	***Td***	**ODt**	***Td***	**ODt**
M1	18	4	3	-	4	-
M2	4	1	3	-	3	-
M3	8	2	5	-	3	-
M4	11	3	4	-	3	-
M5	11	3	6	-	5	-
M6	26	**I**	6	-	4	-

Tabla 4.12. Caso proceso *Tennesse Eastman* con Límites BDE.

	E4.1		**E4.2**		**E4.3**	
	Td	**ODt**	***Td***	**ODt**	***Td***	**ODt**
M1	25	-	20	-	25	2
M2	25	-	20	-	60	4
M3	25	-	20	-	120	5
M4	30	-	20	-	35	3
M5	15	-	15	-	15	1
M6	20	-	15	-	35	3

A continuación, se contabilizó el índice de detección *Id*, que se explicó en la sección 4.3.1.4, para cada escenario en cada grupo y los resultados se añadieron a las tablas 4.13 a 4.15. En

estas tablas también se añade un índice *Idm* que representa el promedio de los *Id* asociados a un ***Mi*** en la tabla.

Análisis del Grupo-Operación 1:
Si se analiza la tabla 4.13, que resume los *Id* para casos de perturbaciones rápidas del tipo escalón, vemos que el ***M2*** (*EWMA*-PCA) detecta más rápidamente este tipo de anormalidades. Aunque, como puede verse de los valores de los *Id* para cada escenario por separado, las detecciones de los distintos ***Mi*** son muy semejantes y en la mayoría de los casos (excepto para el caso **E1.3**) todos caen en la región $\boldsymbol{R_1}$ del gráfico de la figura 4.3. Esto indica que la adopción de un ***Mi*** ú otro para manejar este tipo de perturbaciones no es crítica. Incluso, puede verse que la monitorización con ACP (***M1***) clásico es mejor que muchos ***Mi*** y es equivalente a utilizar la combinación ACP-*levashrink* (***M5***).

Tabla 4.13. *Id* para los escenarios del Grupo-Operación 1

	E1.2	**E1.3**	**E2.2**	**E2.4a**	**E2.4b**	**E3.3**	**E4.1**	*Idm*
M1	1	0.60	0.75	0.75	1	1	1	0.87
M2	1	0.60	1	1	1	1	1	0.94
M3	1	0.60	1	0.60	1	0.75	1	0.85
M4	1	0.60	0.75	1	0.75	1	0.75	0.84
M5	1	0.60	1	1	0.75	0.75	1	0.87
M6	1	0.60	0.75	1	0.75	0.60	1	0.81

Se debe notar que para el caso **E1.3** todos los ***Mi*** muestran una diferencia significativa de detección respecto a los otros escenarios afectados por perturbaciones similares. Esto se debe a que el salto que se produce es más leve en magnitud que por ejemplo el que se produce con la perturbación en **E1.2**. Así, en la gráfica 4.7 se observa que el salto brusco de las condiciones para **E1.2** se nota con claridad tanto en el *SPE* como en el T^2, mientras que el salto brusco de las condiciones para **E1.3** no es detectado en *SPE*. No obstante, al utilizarse ambos estadísticos logra ser detectado por todos los métodos.

Análisis del Grupo-Operación 2:
La tabla 4.14 muestra casos de perturbaciones que aparecen poco a poco, lo cual produce valores bastante bajos de *Id*, sobre todo en el caso **E2.3**. Como en el análisis del grupo anterior, la ventaja de ***M2*** sobre el resto de métodos es notable siendo la mejor opción tanto globalmente (*Idm*) como en los casos específicos de **E1.4** y **E2.3**. Igualmente, ***M5*** vuelve a ser el segundo globalmente y el primero en el caso de **E1.4**, aunque su rendimiento puede llegar a ser superado en algún caso por ***M3*** o ***M4***. Por último, se nota que ***M1*** y ***M5*** no son recomendables para manejar estos casos por lo que deberían ser descartados como opciones para procesos afectados frecuentemente por perturbaciones de este tipo.

Tabla 4.14. *Id* para los escenarios del Grupo-Operación 2

	E1.4	**E1.5**	**E2.3**	**E3.2**	*Idm*
M1	0.27	0.13	0.05	0.17	0.15
M2	0.50	0.25	0.09	0.75	0.40
M3	0.38	0.25	0.09	0.38	0.27
M4	0.38	0.60	0.08	0.27	0.33
M5	0.33	0.75	0.09	0.27	0.36
M6	0.21	0.20	0.06	0.12	0.15

Análisis del Grupo-Operación 3:
La tabla 4.15 muestra casos de perturbaciones asociadas a oscilaciones aleatorias como las que se ilustran en la figura 4.8.

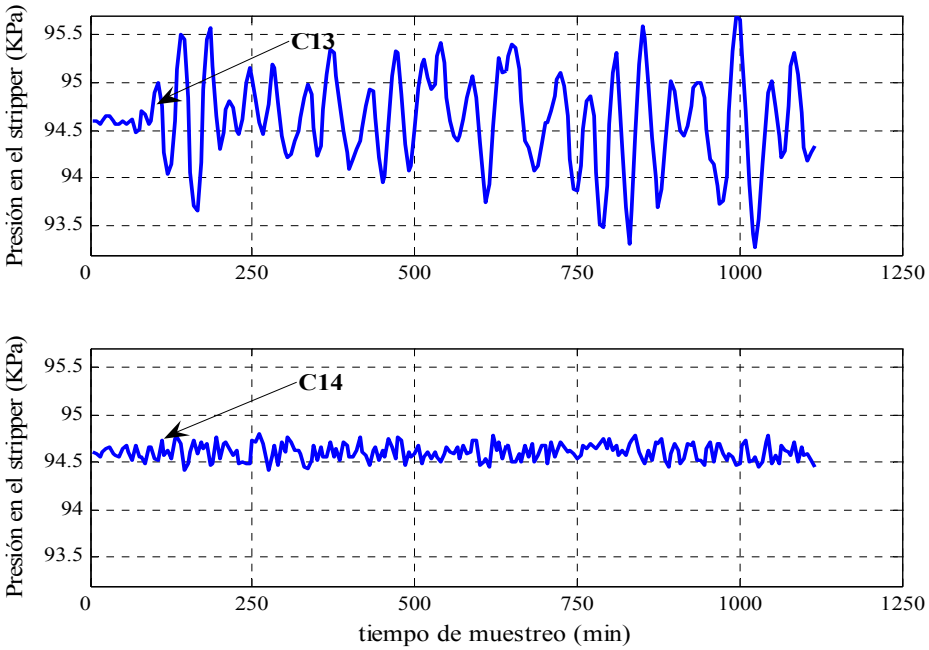

Figura 4.8. Perturbaciones asociadas a los casos **E4.2-E4.3**.

En el caso de **E4.2** las oscilaciones son fuertes por lo que se detectan sin problemas con cualquiera de los métodos. No obstante, en el caso **E4.3** las oscilaciones son mucho más pequeñas lo que produce problemas en la detección utilizando los métodos ***M2***, ***M3***, ***M4*** y/o ***M5***, mientras que ***M1*** y ***M6*** si son capaces de detectar. Así, para este caso tanto ***M1*** como ***M5*** son las mejores opciones de trabajo.

Tabla 4.15. *Id* para los escenarios del Grupo-Operación 3

	E4.2	**E4.3**	*Idm*
M1	0.75	1	0.88
M2	0.75	0.30	0.53
M3	0.75	0.15	0.45
M4	0.75	0.60	0.68
M5	1	1	1
M6	1	0.60	0.80

Análisis del Grupo-Operación 4:
Por último se analiza la tabla 4.16, que presenta la detección en los casos donde la operación es normal. Idealmente no debería detectarse ninguna anormalidad ya que se esta operando en condiciones normales. No obstante, se observa que se generan diversas falsas alarmas de fallo.

Tabla 4.16. Escenarios con operación normal (Grupo-Operación 4)

	t_{SPE}				t_{T2}			
	E1.1	***E2.1***	***E2.5***	***E3.1***	***E1.1***	***E2.1***	***E2.5***	***E3.1***
M1			9,93			191	103	
M2			9,62,85,99	156		177	103	
M3		64	11,59,87,99	58,138,157		177	103	23,177,314
M4	33			137,160,260				
M5	35							
M6				157	26			

El ***M5*** es el que produce menos falsas alarmas de fallo (solo una falsa alarma generada en 4 casos analizados) mientras que los otros generan muchas más, particularmente los métodos ***M2*** y ***M3***. Teniendo en cuenta este último resultado, la mejor opción para evitar falsas alarmas durante la operación sería adoptar el método ***M5***.

Comportamiento global:
Vistos los resultados anteriores, el método *EWMA*-ACP (***M2***) ofrece los mejores resultados para la monitorización de procesos usualmente afectados por perturbaciones de aparición lenta aunque con las siguientes desventajas: a) no detecta de forma correcta aquellos casos donde se presentan anormalidades de tipo oscilaciones suaves; b) genera muchas alarmas falsas que podrían llegar a confundir a los operadores o, incluso, si se producen muy poco antes de una perturbación, podrían llegar a oscurecer la detección.

La alternativa clara sería el uso del la combinación *levashrink*-ACP (***M5***) que detectaría algunas perturbaciones largas con un poco más de lentitud en comparación al *EWMA*-ACP pero que reduciría drásticamente los inconvenientes de mala detección para casos como los de la tabla 4.15 y la generación de alarmas falsas como las mostradas en la tabla 4.16.

Finalmente, se debe observar que la propuesta *levashrink*-ACP (***M5***) es mejor opción que el MSSUM-ACP (***M3***) el cual se propone en la literatura para manejo de perturbaciones largas (Chen et al, 2001). También se debe hacer notar que la estrategia de rectificación combinada (según el planteamiento propuesto en el capítulo 2) y etiquetada en los experimentos como ***M6***, no llega a impactar positivamente al integrarse en la monitorización.

4.4 Conclusiones

En este capítulo se ha presentado una comparación de métodos de monitorización basados en ACP y orientados principalmente a procesos que pueden verse frecuentemente afectados por anormalidades de aparición lenta como por ejemplo una apertura leve y continua de una válvula.

De entre los métodos comparados, el *EWMA*-ACP se mostró como el más adecuado para el manejo de este tipo de perturbaciones coincidiendo con resultados previos publicados. No obstante, este método puede presentar problemas (no indicados en la literatura existente) de mala detección de anormalidades tipo oscilación aleatoria y por la generación de alarmas falsas.

La propuesta *Levashrink*-ACP presenta un rendimiento levemente menos efectivo en tiempo de detección que el *EWMA*-ACP para la detección de perturbaciones suaves, aunque presenta la ventaja de mínima ocurrencia de falsas alarmas en los casos estudiados y un buen manejo de la detección de perturbaciones tipo oscilación aleatoria por lo que aparece como una alternativa valida para sustituir al *EWMA*-ACP.

El método *MSSUM*-ACP, que también se propone en la literatura, presenta un rendimiento más pobre que los métodos ya discutidos, incluyendo la propuesta *Levashrink*-ACP.

Por último, se hizo la comparación de la detección utilizando límites de control tradicionales y según la distribución empírica de los datos. A través de la comparación mostrada se demuestra que el uso de límites basados en la distribución empírica conduce a mejores detecciones que el uso de límites calculados según las expresiones comúnmente propuestas en la literatura.

NOMENCLATURA

a	Número de CP que retienen la máxima varianza.
c_υ	Desviación normal estándar correspondiente al percentil superior (1-υ).
$\mathbf{E}$	Residuales o subespacio de residuales del modelo.
Id	Índice de detección.
k	tiempo de muestreo.
m	Número de muestreos disponibles
n	Número de variables medidas
$\mathbf{P}$	Matriz de *loadings*.
$\mathbf{p}_i$	i-ésimo *loading* de los CP.
SPE	Estadística del el error de predicción al cuadrado o *SPE* (*Squared Predictive Error*).
$\mathbf{s}$	Vector con la suma acumulada de s valores pasados de $\mathbf{y}$ en cada instante k.
$\mathbf{T}$	Matriz de *scores*.
$\mathbf{t}_i$	i-ésimo vector de *scores* de los CP.
T^2	Estadístico de *Hotelling*.
Td	Tiempo que ha tardado la detección con un método dado (*Td*)
ti_{SPE}	Tiempo en que el sistema de monitorización utilizado indica una alarma de fallo detectado a través de los valores del *SPE*.
ti_{T2}	Tiempo en que el sistema de monitorización utilizado indica una alarma de fallo detectado a través de los valores del *T2*.
$\mathbf{t}(k)$	*Scores* en el instante k.
toe	Tiempo real en que ocurre o se inicia una perturbación.
t_{REAL}	Tiempo real en que se sabe que se produce una perturbación.
t_{SPE}	Diferencia de tiempo entre la detección con *SPE* y el tiempo real en que se produjo la perturbación.
t_{T2}	Diferencia de tiempo entre la detección con el T^2 y el tiempo real en que se produjo la perturbación.
tts	Tiempo total gastado en la simulación del escenario indicado.
$\mathbf{Y}$	Matriz de datos de proceso de dimensiones mxn,
$\mathbf{y}(k)$	Vector de muestras de las Variables de proceso al instante k.
$\mathbf{z}(k)$	Filtrado de $\mathbf{y}(k)$ obtenido mediante el *EWMA*.

LETRAS GRIEGAS

α	Constante de suavización para el método *EWMA*.
λ	Valores propios contenidos en la diagonal de la matriz $\mathbf{\Lambda}$.
$\mathbf{\Lambda}$	Matriz diagonal, obtenida durante la solución de ACP y a partir de la matriz de correlación muestral de **Y**.
$\mathbf{\Lambda}_S$	Matriz semejante a $\mathbf{\Lambda}$ pero obtenida a partir de los datos en **s**.
$\mathbf{\Lambda}_Z$	Matriz semejante a $\mathbf{\Lambda}$ pero obtenida a partir de los datos filtrados con *EWMA*.

ACRÓNIMOS

ACP	Análisis de Componentes Principales (*PCA* según las siglas en ingles)
CON	Conjunto de Operación Normal.
CP	Componentes principales.
EWMA	*Exponentially Weighted Moving Average* o filtro de suavización
MCP	Mínimos Cuadrados Parciales o *Partial Least Squares (PLS)*.
	Exponencial.

CAPÍTULO 5. COMPARACIÓN DE ESTRATEGIAS BASADAS EN CLUSTERING PARA ANÁLISIS DE PROCESOS MULTIOPERACIONALES

RESUMEN

En la última década, se han explorado diversas estrategias de análisis de datos para explotar la información contenida en los datos de proceso. Una de estas propuestas es la que combina Técnicas Estadísticas Multivariables (TEM) con técnicas de *Clustering* basadas en Lógica Difusa (*CLD*) o TEM-*CLD* para el análisis y la supervisión de procesos sujetos a frecuentes cambios en las condiciones de operación (procesos multioperacionales). En este tipo de propuestas, no se ha establecido la capacidad real de las mismas para superar ciertos problemas típicos de los análisis con *clustering* como la identificación de grupos con formas diversas, el manejo de *outliers*,…, etc., y que pueden conducir a conclusiones erróneas en el análisis y la monitorización. Asimismo, dado el número existente de propuestas se hace necesaria una comparación entre las mismas que ayude a establecer cual es la mejor opción sobre un rango diverso de procesos.

En este capítulo se intenta dar respuesta a los retos anteriores con el fin de asegurar el máximo rendimiento de las estrategias TEM-*CLD* para la monitorización de procesos multioperacionales. Se comienza por hacer una revisión exhaustiva de las estrategias existentes para, luego, comparar la eficacia de cada una cuando se utilizan para analizar datos históricos de diferentes casos de procesos multi-producto. La comparación, que incluye algunas propuestas noveles, busca establecer la mejor técnica en términos de identificación de *clusters* de formas diversas y manejo apropiado de *outliers*. Adicionalmente, se considera el problema de la estimación del número de *clusters* para casos donde esta información no es conocida. Se utilizan distintos casos de estudio que permiten establecer la comparación de las diversas estrategias consideradas sobre un conjunto amplio de situaciones, mostrando cómo las mejoras propuestas en las estrategias TEM-*CLD* garantizan una aplicación mas fiable de las mismas para el análisis y monitorización de procesos.

5.1 Introducción

Recientemente, las estrategias *Knowledge Discovery in Databases* (*KDD*) y *Minería de Datos* (MD) han atraído la atención de investigadores y operadores en la Industria Química y de Procesos (IQP) y en otras áreas como herramientas para explotar información en los datos de trabajo. Entre ellas, el análisis *clustering* es uno de los que más ha llamado la atención dada la capacidad para asistir en la organización de los datos de un sistema e inferir información a partir de ello.

5.1.1 El Análisis Clustering

El análisis *clustering* se puede definir como el proceso de agrupar objetos o datos dentro de clases, grupos o *clusters* con el fin de que los objetos dentro de un *cluster* sean similares entre sí y que al mismo tiempo sean muy distintos a los objetos de los otros *clusters*. Las técnicas de *clustering* son aquellas que permiten lograr el objetivo anterior. Estas técnicas, con raíces en diversos campos (estadística, química analítica, biología molecular, aprendizaje con maquinas, KDD y MD, etc.) han dado lugar a una amplia gama de propuestas (Jain *et al*., 1999; Han y Kamber, 2001).

Se han propuesto diversas clasificaciones a los métodos existentes, siendo una de las más completas y exhaustivas la de Han y Kamber (2001) y que se expone a continuación:

- Métodos de Partición: Obtienen *k* particiones de los datos originales a través de la optimización de una función objetivo que involucra la distancia entre los datos individuales y los posibles centros de *cluster* como medida de similaridad o semejanza para los *clusters*. Ejemplos de éstas son las diversas variantes de *Clustering* basado en Lógica Difusa (*CLD*) o el *k-means* (Jain *et al*., 1999). Son, junto a los métodos jerárquicos, los más populares y, como se verá en la sección 5.1.2, se han utilizado bastante para supervisión de procesos químicos.
- Métodos Jerárquicos: La idea base es crear una estructura jerárquica de los datos, bien sea aglomerándolos o separándolos consecutivamente entre sí. La información resultante se muestra en una especie de árbol, llamado dendograma, que puede ser muy útil y fácil de manejar si el conjunto de datos analizados no es muy grande (menos de 100-140 datos) pero si el conjunto de datos es grande (como es el caso de muchos procesos químicos) se hace difícil descubrir la estructura agrupada de los datos. Jain *et al*., (1999) hacen una buena revisión de estos métodos.
- Métodos basados en densidades: Se utiliza una noción de densidad de los datos contabilizando para cada dato un valor potencial que indica cuan agrupado (si tiene muchos puntos cercanos alrededor) o aislado (si tiene muy pocos puntos cercanos a su alrededor) se encuentra, identificando con esta información regiones de alta densidad (*clusters*) y de baja densidad (separación entre *clusters*). A pesar de lo intuitivo de su idea base, son muy dependientes de diversos parámetros como el radio de vecindad para medir la cercanía de puntos lo que los hace poco autónomos y con altos requerimientos de interacción con el usuario. Ejemplo de estos métodos son el STING o el DENCLUE (Han y Kamber, 2001).

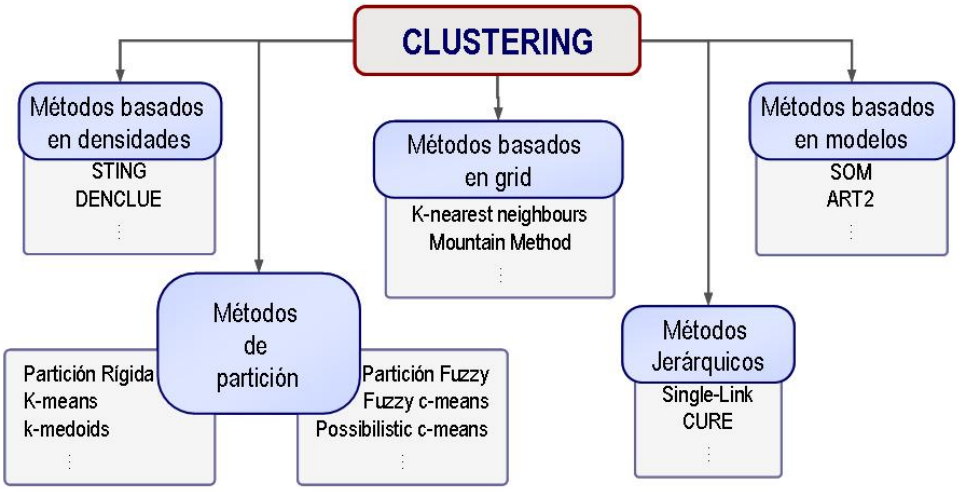

Figura 5.1. Clasificación de los métodos de *clustering*.

- Métodos basados en *grids*: En estos, se divide cada dirección del espacio de los datos en cuadrículas (*grids*) lo que produce celdas hiper-cúbicas. Sobre cada celda se establece la densidad o algún tipo de estadístico que la caracterice. Luego, se utiliza esta información para establecer la semejanza entre celdas y la existencia de *clusters*. Así, estos métodos se hacen independientes del número de datos aun cuando algunos autores resaltan la alta dependencia de estos métodos a la definición de la dimensión

de la rejilla (Han y Kamber, 2001). Un ejemplo de métodos basados en *grids* es el método *Mountain* (Yager y Filev, 1994).

- Métodos basados en modelos: Proponen un modelo para diferentes subconjuntos en los datos y luego miden la semejanza de los mismos para poder establecer los grupos existentes en los datos. Ejemplos de estos métodos son las redes SOM y ART2.

Dada la gran disponibilidad de métodos, se han llevado a cabo comparaciones teóricas entre los métodos para ayudar en la decisión de cual elegir a la hora de hacer una aplicación pero sin llegar a un resultado claro que sirva para la selección (Jain *et al*., 1999; Han y Kamber, 2001).

5.1.2 Utilización del Análisis Clustering en la IQP

En los últimos 10-12, se han explorado las capacidades del *clustering* para asistir en problemas de supervisión de procesos. Así podemos encontrar en la literatura una serie de estrategias de supervisión que integran el *clustering* dentro de sus propuestas y que se pueden reagrupar como sigue:

- Estrategias TEM-*CLD*: En estos casos se utilizan las TEM para obtener una representación reducida del proceso o de las tendencias de las variables del proceso y, luego, estas representaciones reducidas se analizan mediante técnicas *CLD*. Dentro de esta área:
  - Teppola y colaboradores utilizan 2 variantes *CLD* en la estrategia anterior (*Fuzzy c-means* (*FCM*) y *Possibilistic c-means* (*PCM*), para identificar cambios en las condiciones de operación de unas plantas de tratamiento de efluentes afectadas por cambios estacionales y para la monitorización del proceso (Teppola y Minkkinen, 1999; Teppola *et al*., 1999).
  - Sebzalli y Wang obtienen representaciones reducidas de las tendencias de las variables del proceso y las con ayuda del *FCM* para definir las condiciones de operación para la producción de distintos grados de un mismo producto (lo que denominan diseño de productos) en una planta simulada (Sebzalli y Wang, 2001).
  - En una etapa inicial del trabajo de esta tesis (Tona *et al*., 2001) se utiliza la combinación ACP-*FCM* para ayudar a caracterizar el ciclo de ensuciamiento de un reactor y mejorar la operación del proceso.
  - Más recientemente, Lee y colaboradores, adoptan como técnica *CLD* al método *Credibilistic Fuzzy c-means* o *CFCM* (Choi *et al*., 2003; Yoo *et al*., 2003). Los resultados del *CFCM* se usan para crear el sistema de monitorización de varios procesos multioperacionales e incluyen un índice de discriminación para analizar transiciones y una versión adaptativa del *CFCM* para la monitorización en línea.

- Estrategias con S_{ACP}: En éstas, se comienza por dividir los datos en grupos de tamaños que pueden responder a un ordenamiento natural (lotes de producción de procesos discontinuos), o según un criterio apropiado para el análisis actual (duración esperada de los fallos conocidos). Luego, se procesa cada grupo con ACP y, finalmente, se utiliza un índice S_{ACP} que mide la similaridad entre los modelos ACP de cada grupo (Krzanowski, 1979).
  - Seborg y colaboradores (Johannesmeyer *et al*., 2002; Singhal y Seborg, 2002b) proponen una modificación al método de *clustering k-means* que consiste en añadir un criterio de similaridad S_{ACP}, basado en las direcciones y el ángulo de

separación entre los Componentes Principales (CP) obtenidos de distintos conjuntos de datos tratados ACP. El método lo aplican al análisis de operaciones pasadas de un proceso de fermentación por lotes (Singhal y Seborg, 2002b) y, en otro trabajo, lo utilizan para asistir en la definición de patrones de anormalidades pasadas (Singhal y Seborg, 2002a).

- En otro trabajo reciente (Srinivasan *et al*., 2004) se utiliza una variante del S_{ACP} para asistir en la identificación de estados estacionarios y transiciones. La variante del S_{ACP} consiste en que este índice se calcula sobre los CP obtenidos con versiones dinámicas del ACP, tal que se añada la dinámica del proceso al cálculo del S_{ACP}.

- Estrategias con Redes Neuronales Artificiales (RNA). Son similares a los TEM-*CLD*, dado que aplican una reducción inicial de los datos, usualmente con ACP, y sobre estos aplican el análisis clustering con redes SOM o con redes *ART2*.
  - Hwang *et al*., aplican redes basadas en modelos de mapas autoorganizados o *SOM* (*Self Organizing Maps*), para identificar regiones de operación anormal en un horno de incineración industrial (Hwang y Han, 1999). Previo al análisis *SOM* los datos se preprocesan con un filtro basado en la media y, luego, se reduce la dimensionalidad con ACP.
  - Li y Wang obtienen representaciones reducidas de las tendencias de las variables del proceso con ayuda del ACP (Li y Wang, 1999). Luego, las analizan mediante una red neuronal basada en Teorías de Resonancia Adaptativa (*ART2* o *Adaptive Resonance Theory 2*) y mediante *FCM*.
  - También, Witheley *et al.*, trabajan con redes *ART2* para explorar y monitorizar las regiones normales y anormales en un reactor continuo operado a un solo estado de operación normal (Whiteley *et al*., 1996).
  - Chen *et al*., exploran una variante de redes *ART2* para monitorización (Chen *et al*., 2002). Con la variante explorada se intentan reducir ciertos problemas de ajuste de diversos parámetros de las redes *ART2*.

5.1.2.1 Comentarios sobre las estrategias actuales

Haciendo una revisión detallada de las estrategias anteriores se puede ver que:

- Las mismas se han mostrado muy útiles para asistir en la identificación de regiones de operación históricas, utilizar esta información para asistir en la redefinición de las regiones óptimas de operación (diseño de productos) y en tareas de diseño y aplicación de sistemas de monitorización para procesos multioperacionales. Asimismo, pueden ser potencialmente útiles para ayudar a supervisar variaciones o cambios a largo plazo, en las condiciones de operación del proceso como por ejemplo efectos estacionales sobre las demandas de producción.
- Normalmente obedecen a un esquema tipo *KDD* (ver sección 1.2) con número de etapas variables, siendo lo más típico una primera etapa extracción de características (para reducir la dimensionalidad de los datos con ACP) y una segunda etapa de identificación de patrones con c*lustering*. Ocasionalmente, se ha usado algún tipo de preprocesamiento inicial como por ejemplo el filtrado de los datos (Hwang y Han, 1999) o la división del conjunto original datos en ventanas sucesivas de datos (Tona *et al*., 2001; Johannesmeyer *et al*., 2002).
- Algunos autores han discutido sobre la problemática de la parametrización de las *RNA* usadas para *clustering* (Rinta-Runsala, 2001; Chen y Liao, 2002; Singhal y Seborg, 2002a) y los esfuerzos de cálculo significativos involucrados en el entrenamiento de

las mismas viendo en ello un gran obstáculo para la adopción de estrategias que utilizan *clustering* basado en redes neuronales.

- Las estrategias TEM-*CLD* son las más exploradas y también las más aplicadas en entornos industriales (Teppola y Minkkinen, 1999; Rosen, 2001; Yoo *et al*., 2003) aunque con diversas variantes en el uso de la técnica *CLD*, y donde el cambio en la *CLD* obedece a ciertas limitaciones de algunas de ellas (discutidas solo teóricamente) en cuanto a identificación en presencia de *outliers*, aporte de información diversa (inter e intra *cluster*), etc.
- Las estrategias con S_{ACP} involucran la comparación de varios conjuntos de datos de un tamaño predefinido más que la exploración única de un solo conjunto como en las otras estrategias, lo que implica tiempos de procesamiento prohibitivos para situaciones de plantas reales. Pese a esto, este tipo de estrategias ofrece una alternativa interesante para comparar directamente señales de procesos (o grupos de estas) entre sí.
- Aun cuando las técnicas de *clustering* se consideran no supervisadas en el sentido de que son capaces de agrupar diferentes objetos con mínima o ninguna intervención humana, en todas normalmente se asume el conocimiento a-priori del número de grupos en los datos. Pero este valor podría ser una incógnita en ciertos casos de análisis de proceso, de manera que sería aconsejable establecer mecanismos básicos para obtener esta información a través de la técnica de *clustering* a explorar. La determinación del número de *clusters* se ha estudiado de manera insuficiente en la literatura siendo lo más recomendado el análisis basado en índices de validación (ver sección 5.3).
- En alguno de los trabajos existentes se ha hecho la comparación entre los algoritmos *FCM* y *PCM* (Teppola y Minkkinen, 1999). No obstante, sería deseable establecer una comparación más amplia que sirviese para establecer las ventajas y desventajas entre las distintas técnicas utilizadas.

En las secciones que siguen, se muestra un estudio comparativo entre diversas estrategias de análisis de datos basadas en *clustering*. Es un primer intento de establecer la comparación entre las diversas estrategias discutidas anteriormente. Solo se incluyen las técnicas usadas dentro de las estrategias TEM-*CLD*, por ser las más usadas en la literatura y por que en aplicaciones reportadas sobre procesos reales se han mostrado como las más atractivas y útiles.

5.2 Métodos de Clustering basados en Lógica Difusa (CLD)

Los métodos *CLD* caen dentro del grupo de métodos de partición (sección 5.1.1). Es por esto que para poder entenderlos definimos primeramente la idea básica de los métodos de partición. La idea de estos métodos es obtener una partición de los datos en *c* grupos y para ello se han propuesto 2 tipos de particiones posibles que condicionan a los diferentes métodos:

- Partición rígida: El objetivo del *clustering* es repartir el conjunto de datos **Y** en *c* grupos tal que cada objeto en el conjunto de datos pertenece a uno y solo un grupo. Así, sea $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m\}$ un conjunto finito de datos de proceso, donde cada muestra *m* es un vector en $\mathbb{R}^n$ y sea *c* un entero con valores $2 \leq c < m$. La matriz de pertenencias $\mathbf{U}=[\mu_{ik}]$, representa la partición rígida de los datos en **Y** si y solo si sus elementos satisfacen las siguientes propiedades:

$$\mu_{ik} \in 0,1, \quad 1 \le i \le m, \quad 1 \le k \le c \tag{5.1}$$

$$\sum_{k=1}^{c} \mu_{ik} = 1, \quad 1 \le i \le m, \tag{5.2}$$

$$0 < \sum_{i=1}^{m} \mu_{ik} < m, \quad 1 \le k \le c \tag{5.3}$$

Donde μ_{ik} es la pertenencia individual de un objeto o muestra i a un grupo k. Claramente, las pertenencias μ_{ik} asignan cada objeto a un solo grupo. La principal crítica que se le ha hecho a esta forma de ver la partición de los datos consiste en que si hay presencia de *outliers* o si hay puntos o muestras en una región muy cercana al limite entre 2 grupos la simple asignación (=1) o no (=0) a un grupo podría ser información insuficiente acerca de porque la asignación se ha hecho a un grupo y no a otro.

- Partición Difusa (*fuzzy*): Es una generalización de la partición rígida, mediante un enfoque de lógica difusa, que permite valores en el intervalo [0,1] para las μ_{ik}. Esto permite que los valores de pertenencia no solo indiquen a que grupo pertenece cada objeto sino también como se clasifica un objeto en cada grupo. Así, sea $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m]$ un conjunto finito de datos de proceso y sea c un entero con valores $2 \le c < m$, la matriz de pertenencias $\mathbf{U}=[\mu_{ik}]$, representa la partición difusa de los datos en $\mathbf{Y}$ si y solo si sus elementos satisfacen las siguientes propiedades:

$$\mu_{ik} \in [0,1], \quad 1 \le i \le m, \quad 1 \le k \le c \tag{5.4}$$

$$\sum_{k=1}^{c} \mu_{ik} = 1, \quad 1 \le i \le m, \tag{5.5}$$

$$0 < \sum_{i=1}^{m} \mu_{ik} < m, \quad 1 \le k \le c \tag{5.6}$$

La k-ésima columna de $\mathbf{U}$ contiene los valores de la función de pertenencia del k-ésimo subconjunto difuso de $\mathbf{Y}$. La información de pertenencia actual al proporcionar información acerca de la pertenencia de un objeto a cada *cluster* puede permitir un mejor manejo de datos anormales, esto es, *outliers*. En efecto, si a un punto se le asignan valores bajos de pertenencia a cada grupo, esto es un indicativo de que no esta suficientemente relacionado a ninguno de los grupos y por tanto puede ser un valor anormal (*outlier*).

Basado en las anteriores definiciones de partición, se han desarrollado diferentes métodos de *clustering*.

5.2.1 Método k-means

Es el método de partición más simple y una de las técnicas de *clustering* más populares. La idea es tomar una matriz de datos $\mathbf{Y}$ de dimensiones mxn (n es el número de variables medidas, m es el número de observaciones) y reagrupar los datos de $\mathbf{Y}$ en c grupos distintos, tal que se logre minimizar el siguiente criterio de suma de cuadrados:

$$\sum_{k=1}^{c} \sum_{i=1}^{m} \left\| \mathbf{y}_i - \mathbf{v}_k \right\|^2 \tag{5.7}$$

donde la diferencia entre dobles barras denota la norma de una distancia la cual se toma comúnmente como la distancia euclidiana. $\boldsymbol{v}_k$ es la media de los puntos dentro del *cluster k*, y se le conoce como prototipo del *cluster* o centro de *cluster*. Para obtener cada uno de los $\boldsymbol{v}_k$ se utiliza la siguiente ecuación:

$$\mathbf{v}_k = \frac{\sum_{i=1}^{N_k} \mathbf{y}_i}{N_k} \tag{5.8}$$

Donde N_k es el número de observaciones de **Y** en el *cluster k*. El algoritmo para esta técnica es como sigue:

1. Se dividen los datos de **Y** en *k* grupos o particiones iniciales. Típicamente esto se hace de manera aleatoria o mediante alguna regla heurística.
2. Se calcula $\mathbf{v}_k$ por cada grupo usando para ello la ecuación 5.8.
3. Se redefine cada partición por asociar los $\mathbf{y}_i$ de cada partición con el $\boldsymbol{v}_k$ más cercano.
4. Se repiten los pasos 2 y 3 hasta que se alcance la convergencia del algoritmo. Lo usual es tomar la no variación de los $\boldsymbol{v}_k$ como criterio de convergencia.

Algunos autores señalan que el algoritmo es significativamente sensible a la selección de los centros de *clusters* iniciales (Duda *et al*., 2001; Han y Kamber, 2001). Asimismo, se ha visto que el algoritmo solo trabaja bien si en los datos a analizar los *clusters* son esféricos por naturaleza. Finalmente, al trabajar con particiones rígidas se hace muy problemático el manejo de *outliers*.

5.2.2 Método Fuzzy C-Means (FCM)

Se basa en el concepto de particiones difusas (técnicas *CLD*). Sea la matriz de datos **Y** de dimensiones *mxn*. Se intenta determinar la pertenencia de cada objeto de **Y** a cada uno de los *c clusters* mediante la minimización de la siguiente función objetivo:

$$J(\mathbf{Y}, \mathbf{U}, \mathbf{V}) = \sum_{k=1}^{c} \sum_{i=1}^{m} (\mu_{ik})^{\delta} d_{ik}^2 \tag{5.9}$$

$$d_{ik}^2 = \left| \mathbf{x}_i - v_k \right|^2 \tag{5.10}$$

La anterior función objetivo tiene como restricciones a las ecuaciones (5.4) a (5.6). Adicionalmente, δ es un índice de difusividad tal que $\delta \in [1,\infty)$ y es muy importante ya que si se toman valores del mismo cercanos a uno, la partición de datos resultante será rígida ($\delta \rightarrow 1$), mientras que a valores de δ cada vez más lejanos de 1 la partición que resulte será muy difusa ($\delta \rightarrow \infty$). Teppola *et al* compararon los resultados del *FCM* cuando utiliza valores de δ entre 1 y 3 no encontrando diferencias significativas cuando este valor varía entre 1.5 y 3, por lo que recomiendan el valor típico de la literatura $\delta = 2.5$ (Teppola *et al*., 1999).

La minimización de la ecuación 5.11 involucra un problema de optimización no lineal que se podría resolver de distintas maneras. No obstante, el método más popular para resolver este problema se conoce como la iteración de Picard (Bezdek, 1981) tal que en cada iteración los prototipos y las pertenencias de los *clusters* (centros de *clusters*) se recalculan como sigue:

$$\mathbf{v}_k = \frac{\sum_{i=1}^{m} \mu_{ik}^{\delta} \cdot \mathbf{y}_i}{\sum_{i=1}^{n} \mu_{ik}^{\delta}}, \quad 1 \le k \le c \tag{5.11}$$

$$\mu_{ik} = \frac{1}{\sum_{l=1}^{C} \left(d_{ik}^2 / d_{il}^2 \right)^{\frac{1}{(\delta-1)}}}, \quad 1 \le k \le c, \quad 1 \le i \le m, \tag{5.12}$$

Las ecuaciones anteriores, derivadas usando multiplicadores de Lagrange, corresponden a las condiciones para el extremo local asociado al problema representado por el conjunto de ecuaciones 5.4 a 5.6 y 5.9 (Bezdek, 1981). La función de distancia (d_{ik}) utilizada en este caso es la distancia euclidiana. Debido a esto, el *FCM* solo detecta *clusters* con la misma forma (básicamente esféricos). El detalle del algoritmo se muestra a continuación:

- Se especifican los valores para las variables c, δ. y ni=0 (número de iteraciones).
- Se fijan valores aleatorios para los μ_{ik} de cada $\mathbf{y}_i$ tal que se cumplan las ecuaciones 5.4 a 5.6. Luego $\mathbf{U}^{\mathbf{ni}} = [\mu_{ik}]$.
- Se calculan los centroides $\mathbf{v}_i$ mediante la ecuación 5.11.
- **repetir**
    a. Se actualizan las pertenencia μ_{ik} mediante la ecuación 5.12 y $\mathbf{U}^{\mathbf{ni}}$.
    b. Se calcula $du = (\mathbf{U}^{\mathbf{ni}} - \mathbf{U}^{\mathbf{ni-1}})$
    c. Se actualizan los centroides $\mathbf{v}_i$ mediante la ecuación 5.11.

- **mientras** ($du < tolerancia$), donde *tolerancia* se fija entre 0.01 y 0.001.

5.2.3 La modificación de Gustafson-Kessel (GK)

Este método propuesto por Gustafson y Kessel (Gustafson y Kessel, 1979) y mejorado en un trabajo posterior (Babuska *et al*., 2002), corresponde a una extensión al *FCM* en la que se emplea una norma de distancia adaptativa que permita detectar *clusters* de formas geométricas diversas.

Por cada *cluster* se tiene una matriz $\mathbf{A}_k$ que induce la norma de la distancia, lo que conduce a:

$$d_{ikA}^2 = \left(\mathbf{y}_i - \mathbf{v}_k \right)^T \mathbf{A}_k \left(\mathbf{y}_i - \mathbf{v}_k \right), \quad 1 \le k \le c, \quad 1 \le i \le m, \tag{5.13}$$

Las matrices $\mathbf{A}_k$ se introducen como variables de optimización en la función objetivo del *FCM*, para permitir la adaptación de la norma de la distancia a la estructura topológica local de los datos. Así, la ecuación (5.11) se rescribe como sigue:

$$J(\mathbf{Y}, \mathbf{U}, \mathbf{V}, \mathbf{A}) = \sum_{k=1}^{c} \sum_{i=1}^{m} (\mu_{ik})^{\delta} d_{ikA_k}^2 \tag{5.14}$$

Donde $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_c)$. Puede ocurrir que el número de muestras m sea muy pequeño o que los datos dentro de un *cluster* estén linealmente correlacionados. Bajo tales circunstancias Babuska *et al*., han observado que las matrices $\mathbf{A}_k$ pueden hacerse singulares lo que conducirá a problemas en la resolución numérica del algoritmo de *clustering*. Para evitar este tipo de situaciones propuesto que el determinante de $\mathbf{A}_k$ se restrinja como sigue (Babuska *et al*., 2002):

$$\left\| \mathbf{A}_k \right\| = \rho_k, \quad \rho > 0 \tag{5.15}$$

Donde ρ_k es fijo para cada *cluster*. Lo anterior permite que la matriz $\mathbf{A}_k$ varíe con su determinante fijo lo que equivale a optimizar la forma del *cluster* mientras que su volumen

permanece constante. Aplicando multiplicadores de Lagrange, se deriva la siguiente expresión para $\mathbf{A}_k$:

$$\mathbf{A}_k = \left[\rho_k \cdot \det(\mathbf{F}_k)\right]^{1/m} \cdot \mathbf{F}_k^{-1} \tag{5.16}$$

Donde $\mathbf{F}_i$ es la matriz covarianza difusa del k-ésimo *cluster* y viene definida como:

$$\mathbf{F}_k = \frac{\sum_{k=1}^{m} (\mu_{ik})^{\delta} (\mathbf{y}_i - \mathbf{v}_k)(\mathbf{y}_i - \mathbf{v}_k)^T}{\sum_{i=1}^{n} (\mu_{ik})^{\delta}} \tag{5.17}$$

Las anteriores expresiones para el cálculo adaptativo de la norma se insertan fácilmente en el algoritmo de *FCM*. Aunque este método no se ha probado en la literatura previa para el análisis y supervisión de procesos químicos, se verá más adelante que puede ser más útil que algunas de las propuestas previas.

5.2.4 Método Possibilistic C-Means (PCM)

Antes de explicar el PCM, se explora un caso de análisis propuesto en la literatura (Krishnapuran y Keller, 1993). Se tiene un conjunto de datos en el que c=2. En dicho conjunto existen 2 puntos *X11* y *X12* que son equidistantes a cada centro de cada *cluster* (ver la Figura 5.2). Al aplicar *FCM* sobre este conjunto de datos se observa que las μ asignadas *X11* y *X12* son todas iguales a 0.5 aun cuando el punto *X12* se encuentra mucho más desviado de los *clusters* que *X11* o es menos *típico* a ambos *clusters* que *X11*. Lo anterior se debe a que las pertenencias asignadas por el *FCM* están inversamente relacionadas con la distancia relativa de cada punto a los centros v_k sin considerar el valor absoluto de las distancias de cada punto a los v_k.

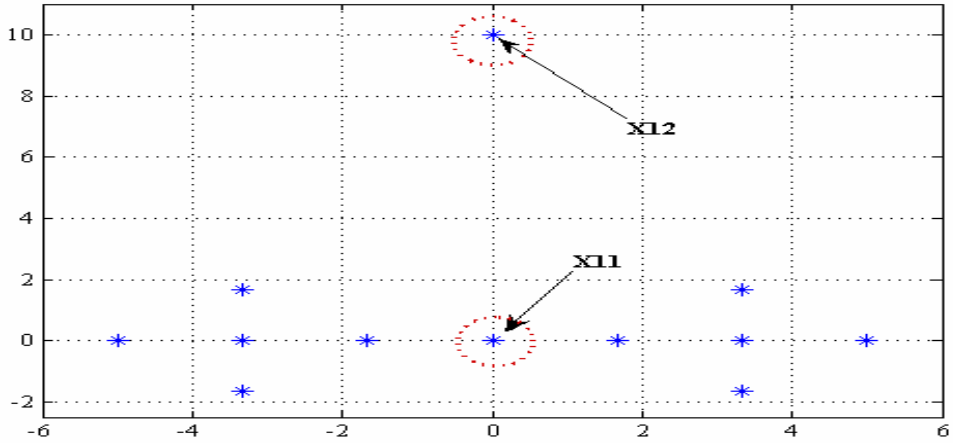

Figura 5.2. Problema de puntos equidistantes con el *FCM*.

PCM es un intento de resolver el problema anterior. Para ello se relaja la restricción probabilística impuesta por la ecuación (5.5) al *FCM* como sigue:

$$\tau_{ik} \geq 0, \quad \forall i,k \tag{5.18}$$

Lo anterior permite reinterpretar los valores de pertenencias (μ_{ik} de la ec. 5.5) como la tipicalidad (τ_{ik}) de $\mathbf{y}_i$ relativa al *cluster k*. La función objetivo se ve modificada según se muestra a continuación:

$$J(\mathbf{Y},\mathbf{T},\mathbf{V}) = \sum_{k=1}^{c}\sum_{i=1}^{m} (\tau_{ik})^{\delta} d_{ik}^2 + \sum_{k=1}^{c} \eta_k \sum_{i=1}^{m} (1-\tau_{ik})^{\eta} \tag{5.19}$$

El primer término es el mismo que aparece en la función objetivo del *FCM* (5.9) pero con τ (la tipicalidad) en lugar de μ (la pertenencia), mientras que el segundo término intenta forzar los valores de la tipicalidad a ser tan altos como sea posible. El parámetro η se usa para establecer una zona de influencia de un punto y_i sobre los prototipos a ser estimados. Al considerarlo dentro de 5.19 se intenta darle robustez al método resultante en el sentido estadístico de ser poco sensible a los efectos de los *outliers*. **T** es equivalente a **U** en el *FCM*, pero en este caso contiene las tipicalidades τ y no las pertenencias μ. La optimización se resuelve por un procedimiento similar al del *FCM*. Los valores de τ se calculan en cada iteración como sigue:

$$\eta_k = \frac{\sum_{i=1}^{m} \tau_{ik}^{\delta} \cdot d_{ik}^2}{\sum_{i=1}^{m} \tau_{ik}^{\delta}}, \quad \forall k \tag{5.20}$$

$$\tau_{ik} = \frac{1}{1 + \left(d_{ik}^2 / \eta_k \right)^{\frac{1}{(\delta - 1)}}} \tag{5.21}$$

Evaluando los datos del ejemplo (ver figura 5.2) con *FCM* y con *PCM* se obtienen los resultados que se muestran en la tabla 5.1. Se observa que para el caso del punto *X11*, ambas técnicas asignan un mismo valor de pertenencia o de tipicalidad respecto de cada *cluster*. Dicho resultado es lógico dado que *X11* se encuentra en un punto equidistante a la nube de puntos de ambos *clusters*, es decir es igualmente aceptable o típico a ambos grupos. En el caso del punto *X12* ocurre algo similar. Los valores de μ_{ik} son iguales para cada grupo y los valores de τ_{ik} también son iguales para ambos grupos. No obstante, *X12* esta mucho más lejos de las nubes de puntos de cada *cluster* y por tanto es poco típico a cualquiera de los 2 grupos. Esto queda reflejado en los correspondientes valores de τ_{ik} que son más pequeños al resto de las τ_{ik} de los demás datos, siendo con ello un posible candidato a *outlier*. En el caso del *FCM* las correspondientes μ_{ik} caracterizan a *X12* de manera similar a *X11* por lo que no se puede llegar a deducir con esta información si este punto es un *outlier*.

Tabla 5.1. Ejemplo de Krishnapuran – Keller para comparar FCM y PCM.

Datos	*FCM* pertenencias		*PCM* tipicalidades	
	Cluster 1	*Cluster* 2	*Cluster* 1	*Cluster* 2
	μ_{i1}	μ_{i2}	τ_{i1}	τ_{i2}
X_{1k}	0.94	0.06	0.49	0.13
X_{2k}	0.97	0.03	0.66	0.19
X_{3k}	0.99	0.01	0.85	0.21
X_{4k}	0.90	0.10	0.65	0.19
X_{5k}	0.92	0.08	0.97	0.35
X_{6k}	0.08	0.92	0.35	0.97
X_{7k}	0.03	0.97	0.19	0.66
X_{8k}	0.01	0.99	0.21	0.85
X_{9k}	0.10	0.90	0.19	0.65
X_{10k}	0.06	0.94	0.13	0.49
X_{11k}	0.50	0.50	0.63	0.63
X_{12k}	0.50	0.50	0.07	0.07

Teppola y Minkkinen utilizan una variante del *PCM* con distancias mahalanobis en lugar de distancias euclidianas (Teppola y Minkkinen, 1999). Muestran cómo los valores de tipicalidad pueden llegar a ser más ventajosos que las pertenencias del *FCM* para la monitorización del caso que estudian. No obstante, otros autores en áreas distintas a ingeniería química (Barni *et al*., 1996; Pal *et al*., 1997) critican ciertos problemas del *PCM* como la sensibilidad del método al paso de inicialización y que en ocasiones puede generar *clusters* coincidentes (que tienen los mismos prototipos de *clusters*) con graves efectos en los resultados. Estos problemas se han observado también en el trabajo de este capítulo (ver sección 5.5.2).

5.2.5 Método Credibilistic Fuzzy C-Means (CFCM)

Este método, propuesto por Chintalapudi y Kam (1998), se derivó para intentar resolver problemas en el *FCM* como el que se discutió al inicio de la sección 5.2.4, para que aportase tanto información Intra-*cluster* (algún tipo de tipicalidad) como información Inter-*cluster* (pertenencias) y que con ello fuese capaz de obtener buenas detecciones de *clusters* aún en presencia de *outliers* (Chintalapudi y Kam, 1998). La función objetivo es la misma que la del *FCM* (ver ecuación 5.9). Sin embargo, la restricción probabilística del *FCM* (ecuación 5.5) se cambia como sigue:

$$\sum_{k=1}^{c} \mu_{ik} = \psi_i \tag{5.22}$$

ψ_i es una variable de credibilidad que representa la tipicalidad de y_i al conjunto completo de datos en $\mathbf{Y}$. Partiendo de la idea de que la distancia entre 2 objetos es una medida de su similaridad, la credibilidad intenta establecer el grado de aislamiento de un objeto en un espacio característico y utilizar esto como medida de tipicalidad. Así, sea el conjunto $\{\mathbf{y}_i^{i2} \in \mathbf{Y} \mid i2 = 1,\ldots,\sigma\}$ representando a los σ objetos más próximos a $\mathbf{y}_i$, en términos de alguna norma de distancia $||\ ||_C$, la credibilidad del vector $\mathbf{y}_i$ será:

$$\psi_i = 1 - \frac{\left(\kappa_i - \min(\kappa_1, \kappa_2, \ldots, \kappa_n)\right)}{\left(\max(\kappa_1, \kappa_2, \ldots, \kappa_n) - \min(\kappa_1, \kappa_2, \ldots, \kappa_n)\right)} \tag{5.23}$$

$$\kappa_i = \frac{\sum_{i2=1}^{\sigma} \left(\left\|\mathbf{y}_i^{i2} - \mathbf{y}_i\right\|\right)}{\sigma} \tag{5.24}$$

$$\sigma = \gamma \frac{m}{c} \tag{5.25}$$

κ_i es la distancia media entre $\mathbf{y}_i$ y los σ objetos más cercanos y γ es una constante con valores en [0,1]. Chintalapudi y Kam demuestran que los resultados de *clustering* son relativamente insensibles al valor de γ y proponen trabajar con un valor fijo de 0.5 (Chintalapudi y Kam, 1998). También comprueban que cualquier objeto tipo *outliers* tiene valor bajo de la credibilidad, mientras que objetos no-*outliers* que son similares a muchos otros objetos tienden a tener valores muy altos de credibilidad. Asimismo, resaltan que los valores de ψ_i para cualquier conjunto de datos siempre caen dentro del intervalo [0,1] mientras que los *outliers* toman valores tendientes a 0.

La resolución del problema *CFCM* es similar a la del *FCM*. Tomando la credibilidad y operando adecuadamente sobre la función objetivo en (5.9) junto a la restricción en (5.22) se llega a que los prototipos de *clusters* y las pertenencias se calculan como sigue:

$$v_k = \frac{\sum_{i=1}^{m} \mu_{ik}^{\delta} \cdot x_k}{\sum_{i=1}^{m} \mu_{ik}^{\delta}} \tag{5.26}$$

$$\mu_{ik} = \frac{\psi_k}{\sum_{l=1}^{c} \left(d_{ik}^2 / d_{il}^2 \right)^{\frac{1}{(\delta-1)}}} \tag{5.27}$$

5.2.6 Método Fuzzy Possibilistic C-Means (FPCM)

De manera similar al *CFCM*, este método fue propuesto por Pal *et al.,* como un intento por resolver problemas en el *FCM* como el que se discutió al inicio de la sección 5.2.4, para que aportase información intra-*cluster* (algún tipo de tipicalidad) e inter-*cluster* (pertenencias) simultáneamente, y que con ello fuese capaz de obtener buenas detecciones de *clusters* aun en presencia de *outliers* (Pal *et al*., 1997). El problema de optimización para este caso queda como sigue:

$$J(\mathbf{Y}, \mathbf{U}, \mathbf{T}, \mathbf{V}) = \sum_{k=1}^{c} \sum_{i=1}^{m} (\mu_{ik}^{\delta} + \tau_{ik}^{\eta})^{\delta} d_{ik}^2 \tag{5.28}$$

$$\sum_{k=1}^{c} \mu_{ik} = 1 \tag{5.28b}$$

$$\sum_{k=1}^{c} \tau_{ik} = 1 \tag{5.28c}$$

η es una constante similar a la del *PCM* pero en este caso los autores proponen que por tomar un valor entero entre [2 , 4] se llega a resultados fiables y más o menos iguales. El algoritmo de resolución es similar al del *FCM* y los μ_{ik}, τ_{ik} y v_k se pueden actualizar en cada iteración como sigue:

$$\mu_{ik} = \frac{1}{\sum_{l=1}^{c} \left(d_{ik}^2 / d_{il}^2 \right)^{\frac{1}{(\delta-1)}}} \tag{5.29}$$

$$\tau_{ik} = \frac{1}{\sum_{s=1}^{m} \left(d_{ik}^2 / d_{is}^2 \right)^{\frac{1}{(\eta-1)}}} \tag{5.30}$$

$$\mathbf{v}_k = \frac{\sum_{i=1}^{m} \left(\mu_{ik}^{\delta} + \tau_{ik}^{\eta} \right) \cdot \mathbf{y}_i}{\sum_{i=1}^{m} \left(\mu_{ik}^{\delta} + \tau_{ik}^{\eta} \right)} \tag{5.31}$$

En el caso de conjuntos con muchos datos, los valores de τ tienden a ser muy bajos por lo que se recomienda que sean escalados para una más fácil interpretación de los mismos.

5.3 Estimación del número de clusters

Todos los métodos anteriores requieren como entrada el número de *clusters c* en que se dividen los datos. En muchas situaciones este valor puede conocerse a-priori. Por ejemplo, supóngase que se quiere explorar una base de datos históricos correspondiente a operaciones normales de un proceso donde se fabrican *h* productos con distintos grados de calidad. Se sabe que cada grado de producto corresponde a una producción con condiciones de operación distintas a las del resto. Por lo tanto, se espera que al aplicar el *clustering*, con el dato de entrada *c=h*, la información de las operaciones puntuales a lo largo de los históricos se clasifique en *c* grupos o categorías. Puede haber otras situaciones en que no se conozca esta información del proceso. En estos últimos casos, es deseable tener una herramienta que asista en la determinación del número de grupos en que se dividen los datos. A continuación se describen algunas estrategias para ello y para el caso de técnicas *CLD*.

5.3.1 Estimación basada en índices

Las estrategias de estimación de *c* basadas en índices siguen siempre un esquema básico (Halkidi *et al*., 2001):

1 Se fija un *c* mínimo $c_{mín}$= 2 y un *c* máximo $c_{máx}$=*CM*, donde *CM* es un valor entero fijado por el usuario.
2 Se aplica el *FCM* sobre la matriz de datos en estudio (**Y**) y con $c_{actual} = c_{mín}$.
3 Se selecciona un índice de eficacia o validación I_{VAL} y se evalúa con información obtenida en el paso anterior. El valor de I_{VAL} se guarda.
4 Se incrementa el valor de c_{actual} en una unidad. Se repiten los pasos 2 y 3.
5 Se repite 2, 3 y 4 hasta que se cumpla que $c_{actual} = c_{máx}$.
6 Dependiendo del I_{VAL} utilizado, se selecciona *c* igual al óptimo de la curva I_{VAL} vs. c_{actual}.

En la literatura se han propuesto diversos índices de validación I_{VAL} (Pal *et al*., 1997; Halkidi *et al*., 2001). Ninguno de ellos produce resultados ampliamente superiores al resto. A continuación se describen los más usados en la literatura.

5.3.1.1 Coeficiente de Partición

Este coeficiente, etiquetado como *PC* (siglas del nombre en ingles), mide la cantidad de solapamiento entre *clusters* (Pal *et al*., 1997):

$$PC = \frac{1}{m}\sum_{k=1}^{c}\sum_{i=1}^{m}(\mu_{ik})^2 \qquad (5.32)$$

La interpretación de los resultados de este índice es como sigue:

- Si los objetos en **Y** se han agrupado correctamente, entonces el valor de *PC* tiende a 1 y **U** tiende a representar una partición rígida de **Y** tal que:
  $$PC \to 1 \qquad (5.33a)$$
- Si *PC* tiende a valores cercanos a (1/c), esto esta queriendo decir que **U** muestra ausencia de una tendencia de *clusters* en los datos.
  $$\left\{PC \to (1/c), \quad \mu_{ik} \to (1/c)\right\} \qquad (5.33b)$$

En la literatura se recomienda que el *c* seleccionado deberá corresponderse con un máximo en la curva I_{VAL} vs. *c*.

5.3.1.2 Coeficiente de Entropía de la Partición (CE)

Es similar al índice anterior. Se define como sigue (Halkidi *et al*., 2001):

$$CE = -\frac{1}{m}\sum_{i=1}^{c}\sum_{k=1}^{m}\mu_{ik}\cdot\log(\mu_{ik}) \tag{5.34}$$

Los valores del *CE* caerán dentro del intervalo [*0*, *log*(*c*)]. La interpretación de los resultados de este índice es como sigue:

- Si los objetos en **Y** se han agrupado correctamente, entonces **U** tiende a representar una partición rígida de **Y** tal que:
  $CE \to 0$ (5.35a)
- Si los objetos en **Y** se han agrupado en cada grupo con valores de μ_{ik} muy cercanos o iguales a (1/*c*), entonces **U** indica ausencia de una tendencia de *clusters* en los datos.
  $CE \to \log(c)$ (5.35b)

En la literatura, se recomienda que el *c* seleccionado deberá corresponderse con un mínimo en la curva I_{VAL} vs. *c*.

5.3.1.3 Índice de Xie-Beni (XB)

La meta con este coeficiente es cuantificar la relación entre la variación total dentro de los *clusters* y la separación entre ellos (Pal *et al*., 1997):

$$XB = \frac{\sum_{k=1}^{c}\sum_{i=1}^{m}(\mu_{ik})^{\delta}\left\|\mathbf{y}_i-\mathbf{v}_k\right\|^2}{m\cdot\min\left\|\mathbf{v}_i-\mathbf{v}_j\right\|^2} \tag{5.36}$$

Con este índice *XB* se espera identificar el número óptimo de *clusters c* como el mínimo en la curva I_{VAL} vs. *c*. Al usar este índice se ha visto que si el valor del $c_{máx}$ utilizado dentro del algoritmo descrito al inicio de la sección 5.3.1 es muy alto (por ejemplo, si $c_{máx}$ tiende a *m*), los correspondientes $I_{VAL} = XB$ se hacen monótonamente decrecientes (Halkidi *et al*., 2001). Al producirse esto, se afecta la correcta identificación del *c* óptimo. Para evitar el problema anterior Halkidi *et al*., (2001) proponen que al trabajar con este índice inicialmente se explore la curva I_{VAL} vs. *c*, con un c_{max} muy alto, y se identifique el punto de inicio del comportamiento monótono de la curva (Halkidi *et al*., 2001). Luego, se fija el c_{max} igual a dicho punto de inicio y, finalmente, se identifica el *c* óptimo dentro del intervalo [2, c_{max}].

5.3.2 El método Subtractive Clustering (MSCl)

Este método fue propuesto por Chiu como una mejora a un método similar llamado método *Mountain* (Chiu, 1994; Yager y Filev, 1994) . Ambos métodos se plantearon para determinar *c* en un conjunto de datos así como un estimado inicial de los centros de cada *cluster* que sirviese a otros métodos como el *FCM* o las redes *SOM*.

El método se basa en considerar inicialmente cada punto (dato en el conjunto **Y**) como un centro de *cluster* en potencia. Luego, se define una función que mida para cada punto su potencial P_i como centro de *cluster*. Con ayuda de estos P_i y mediante un procedimiento iterativo se logra una estimación del número *c* de *clusters* en **Y**. A continuación se describe el algoritmo:

1 Se normalizan los datos en **Y** con lo que se restringe la dimensión original de los datos a un hipercubo de igual dimensión pero teniendo como rango máximo de valores el intervalo [0,1] en cada dimensión.
2 Se calcula el potencial de cada punto $\mathbf{y}_i$.

$$P_i = \sum_{l=1}^{m} e^{-\alpha \|\mathbf{y}_i - \mathbf{y}_l\|^2} \quad (5.37)$$

$$\alpha = \left(2/r_a\right)^2 \quad (5.38)$$

Donde P_i representa el potencial del *i*-ésimo punto y r_a es el radio definiendo una vecindad (ver sección 5.3.2.1). P_i es función de las distancias del punto *i* al resto de los (*m-1*) puntos en **Y**.

3 Una vez calculados los P_i para cada punto, se selecciona el punto con el mayor P_i como el primer centro de *cluster* y sus coordenadas como las coordenadas del centro de *cluster*:

$$P_1^* = \max\left(P_i\right) \quad (5.39)$$

4 Se corrige el potencial de cada punto $\mathbf{y}_i$. Para ello, se resta a cada uno una cantidad que es función de su distancia al primer potencial de punto detectado:

$$P_{i-corr} = P_i - P_1^* e^{-\beta \|y_i - y_1^*\|^2} \quad (5.40)$$

$$\beta = \left(2/r_b\right)^2 \quad (5.41)$$

Donde $r_b = 1.25 * r_a$. Al aplicar (5.40), los potenciales de los puntos cercanos al primer centro serán los que se vean más fuertemente reducidos.

5 Se detecta el siguiente centro de *cluster* como el punto con el máximo potencial corregido:

$$P_i^* = \max\left(P_{i-corr}\right) \quad (5.42)$$

6 Se repiten 3 y 4 de forma continua y según el resultado del siguiente *criterio de terminación*:

Si $P_i^* > \overline{\varepsilon} \cdot P_1^*$

Se acepta y_i^* como un centro de *cluster* y se vuelve a 3.

Sino si $P_i^* < \underline{\varepsilon} \cdot P_1^*$

Se rechaza y_i^* y se termina el proceso de *clustering*.

Sino

Sea $d_{\min}$ la distancia más corta entre $\mathbf{y}_i^*$ y todos los centros de *clusters* previamente encontrados:

Si $\frac{d_{\min}}{r_a} + \frac{P_i^*}{P_1^*} \geq 1$

Se acepta y_i^* como un centro de *cluster* y se vuelve a 3.

Sino

Se rechaza $\mathbf{y}_i^*$ y se fija su potencial en 0. Se selecciona el siguiente punto con $P_{\text{i-corr}}$ más alto y se vuelve a evaluar todo el *criterio de terminación*.

Fin

En el criterio anterior $\overline{\varepsilon}$ representa un umbral sobre el cual se acepta (sin lugar a dudas) un punto como centro de *cluster*. Por el contrario, $\underline{\varepsilon}$ representa un umbral bajo el cual se rechaza (sin lugar a dudas) un punto como centro. La región entre estos 2 valores se evalúa según la parte final del criterio.

5.3.2.1 Variaciones sobre el método Subtractive Clustering

El método anterior es dependiente de diversos parámetros $\overline{\varepsilon}$, $\underline{\varepsilon}$ y r_a. En la literatura se proponen valores para $\overline{\varepsilon}(0.5)$ y $\underline{\varepsilon}(0.15)$ para los cuales se ha visto que variándolos conducen a resultados similares y aceptables. No sucede lo mismo para el caso de r_a. La opción por defecto es experimentar para distintos valores de r_a y confrontar los estimados de c con el análisis visual de los datos. Esta opción es definitivamente costosa en tiempo. Diversos autores han propuesto métodos para obtener un estimado del mismo sin llegar a conclusiones satisfactorias (Paiva *et al*., 1999; Demirli *et al*., 2003).

En este capítulo se propone una estrategia para estimar r_a. La idea del método propuesto es sencilla. Se propone estudiar la distribución de las distancias en cada dimensión y ver algún valor típico que represente adecuadamente a r_a, tal que tome en cuenta las distancias de cercanía entre puntos más típicas (distancias entre puntos dentro de un *cluster*) y deje a un lado las distancias más largas entre puntos (distancias entre puntos en distintos *clusters*). La estrategia se describe a continuación:

- Tras la normalización en el paso 1 del algoritmo *MSCl*, se calcula la distancia d_{il} de cada punto $\mathbf{y}_i$ al resto de los puntos en $\mathbf{Y}$ ($\mathbf{y}_l$), obteniéndose con ello la matriz $\mathbf{D}$.
- Por cada columna de $\mathbf{D}$, se calcula la distribución empírica de las d_{il} y se selecciona el valor de r_a tal que su valor típico α esté en el intervalo de probabilidad [80 % - 90 %]. Con esto se obtiene un vector $\mathbf{r_a} = <r_{a1}, r_{a2},\ldots, r_{as}>$ con un valor de vecindad adecuado para cada dimensión de los datos.

La anterior estrategia se etiqueta como *MSCl-1*. Dado que el objetivo global del trabajo en este capítulo es la evaluación y mejora de las estrategias TEM-*CLD*, se propone una variante a la *MSCl-1*, en la que los datos de $\mathbf{Y}$ se preprocesan con ACP y, luego, se trabaja con los *scores* que aporta el modelo ACP de $\mathbf{Y}$. Esta segunda variante se etiqueta como *MSCl-2*.

Adicionalmente, para el caso de *MSCl-2*, si se toma cada variable en $\mathbf{Y}$, y se sustituye por la aproximación obtenida mediante una *wavelets Daubechies db1* a un nivel de descomposición l=2 (ver secciones 2.1.1.2.2 y 2.1.2), se puede mejorar la estimación del nº de *clusters*. Esto se ha visto válido para el caso de datos con ruido y más ventajoso que la alternativa de trabajar sobre el filtrado de los datos lo que significa que para el caso del análisis *clusters*, una leve reducción del ruido presente en los datos mejora la detección de los mismos.

5.3.3 Validación mediante índice

La validación de los *clusters* se refiere al problema de evaluar la asignación de objetos a cada partición tras haber aplicado una técnica de *clustering*. La validación del *clustering* también se propone en la literatura mediante el uso de índices, como el PC, el CE o el XB (Bezdek, 1981; Halkidi *et al*., 2001), que den alguna indicación de la calidad de las particiones o grupos creados y la asignación de objetos. La ventaja de usar estos índices radica en que solo necesitan como entradas la información (parcial o completa) que se obtiene tras aplicar la técnica *clustering* utilizada. No obstante, muchos autores critican la falta de conexión de los mismos con respecto a alguna propiedad de los datos (Halkidi *et al*., 2001; Balasko *et al*.,

2004). Más recientemente, Singhal y Seborg propusieron 2 índices de validación que toman en cuenta información propia del proceso en estudio (Singhal y Seborg, 2002b).

5.3.3.1 Pureza del Cluster

Mediante este índice se intenta caracterizar la pureza de cada *cluster* en términos de cuantos lotes, producciones o muestras (objetos) con una condición de operación asociada *j* están presentes en cada *cluster*. Matemáticamente se expresa como:

$$P_{r_k} = \frac{\left(\max_j N_{kj}\right)}{Np_k} * 100\%, \quad k = 1, 2, ..., c. \tag{5.43}$$

Donde N_{kj} es el número de objetos con una condición de operación *j* que están presentes en el *k*-ésimo *cluster* y donde el máximo es para indicar que la operación dominante en el *k*-ésimo *cluster* es aquella con el valor más largo de N_{kj}. Np_k es el número de objetos en el *cluster k*.

5.3.3.2 Eficiencia del Cluster

Este índice se utiliza para intentar caracterizar la extensión a la cual una condición de operación *j* se distribuye en diferentes *clusters*. Así, intenta penalizar valores amplios de *c* cuando una condición de operación se distribuye en distintos *clusters*. Se expresa como sigue:

$$\xi_k = \frac{\left(\max_k N_{kj}\right)}{N_{DB\,j}} * 100\% \tag{5.44}$$

Donde N_{DBj} es el número total de objetos pertenecientes a la condición de operación *j* en los datos disponibles.

Los índices anteriores requieren un conocimiento detallado de los datos usados y del proceso en estudio por lo que se ven particularmente útiles para estudios de comparación de técnicas en los que se tiene bastante conocimiento de los casos estudiados (ver secciones 5.5.2 y 5.5.3).

5.4 Estrategias de análisis y monitorización de procesos basadas en clustering

En la sección 5.1.2 ya se discutió sobre el uso de estrategias basadas en *clustering* para aplicaciones de ingeniería química, particularmente de las estrategias TEM-*CLD* para análisis e procesos multioperacionales. Se debe recalcar que la razón de aplicar una Técnica Estadística Multivariable (TEM) previo al *clustering* descansa en la evidencia experimental de que trabajando sobre los *scores* obtenidos del ACP (o de alguna técnica TEM similar) se obtienen *clusters* mejor definidos que cuando se trabaja sobre los datos originales (Teppola y Minkkinen, 1999; Teppola *et al*., 1999).

El esquema general de estas estrategias es siempre el mismo y se describe a continuación:

- Etapa de análisis:
  - Se toma la matriz de datos de proceso **Y**, con operaciones normales pasadas y se procesa con una técnica TEM, usualmente el ACP. En **Y**, las columnas representan las variables del proceso y las filas representan las observaciones (muestreos) de dichas variables en el tiempo. Se seleccionan los CP del modelo obtenido y, luego, se guardan los correspondientes *scores* (t_i) del modelo (ver detalles sobre modelos ACP en sección 4.1).

- Se toman los t_i y se procesan mediante una técnica *CLD*. Tras esto se obtiene el agrupamiento de las distintas observaciones (mediante sus μ_{ik}) según la única información de partida del procesos que consiste en el número de condiciones de operación (o el número de productos fabricados) *c*. Asimismo, se obtienen los prototipos de *clusters* $\mathbf{v}_k$, que definen el patrón más característico de cada grupo. Los resultados se intentan verificar bien sea por exploración visual de los resultados o mediante algún índice de validez (ver sección 5.3) de manera que aporten información significativa sobre el comportamiento del proceso en operaciones pasadas (condiciones de operación de cada producto). El proceso se describe en la Figura 5.3.

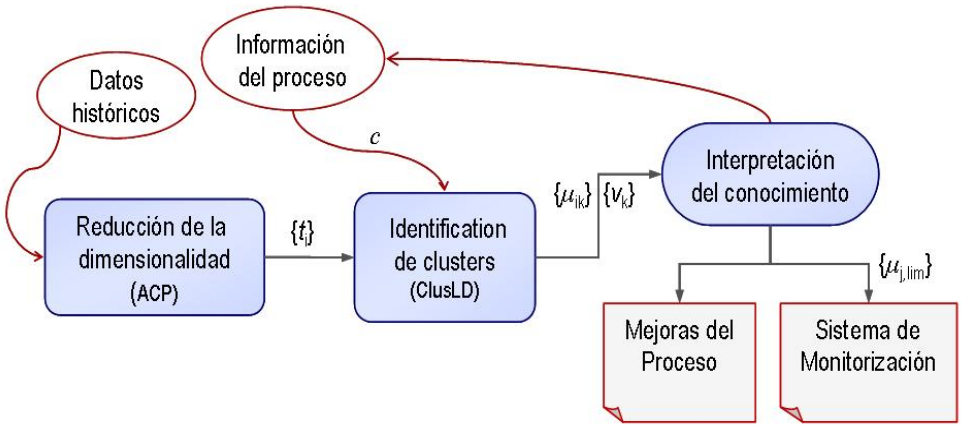

Figura 5.3. Etapa de Análisis del Proceso.

- Etapa de monitorización continua:
  - En cada nuevo instante de tiempo se transforman las observaciones obtenidas a sus correspondientes *scores*, se calcula la pertenencia a cada grupo *k* ($\mu_{tk} = \mu_{ik}$) y se evalúa la siguiente regla de control:

$$\mu_{jk} = \max(\mu_{ik}) \geq \mu_{j,\lim} \qquad (5.45)$$

Donde los μ_{lim} se han calculado con los valores de μ_{ik} obtenidos al final de la etapa de análisis. Si el máximo de las μ_{ik} es mayor qué el $\mu_{j,lim}$ de alguno de los grupos, el proceso está en control y dentro de la condición de operación *j*. Por el contrario, si ninguno de los límites se supera, el sistema presenta alguna anormalidad. El proceso se describe en la Figura 5.4. Alternativamente, se pueden usar los gráficos de control basadas en los estadísticos *SPE* y T^2, que se obtienen del ACP (Ver sección 4.1), junto con el análisis de las μ_{ik}. Yoo *et al.*, utilizan dicha combinación (Yoo *et al.*, 2003) y concluyen que por el uso de ambos tipos de controles se brinda una información más completa para la monitorización de los casos que estudian.

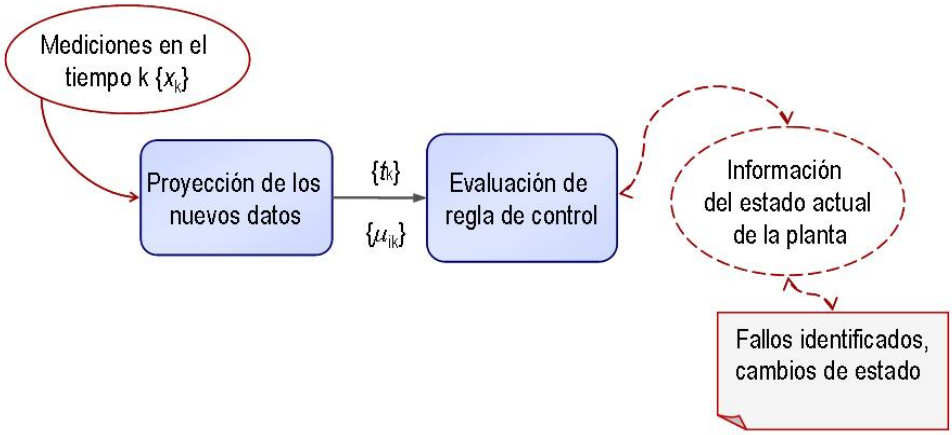

Figura 5.4. Etapa de Monitorización del Proceso.

Lo que se describe como etapa de análisis se puede encontrar en todos los trabajos donde se aplica la combinación TEM-*CLD*. En alguno de estos trabajos este análisis es usado para definir las regiones de operación históricas para distintos grados de un mismo producto y lo definen como diseño del producto (Sebzalli y Wang, 2001). En otros trabajos este mismo análisis se utiliza como base para hacer inferencias acerca del proceso y proponer mejoras sobre el mismo (Næs y Mevik, 1999). Finalmente, en el resto de los trabajos esta etapa sirve como base para el diseño de sistemas de monitorización (Teppola y Minkkinen, 1999; Li y Wang, 2001; Rosen, 2001; Choi *et al*., 2003; Yoo *et al*., 2003).

La etapa de monitorización consiste en tomar en cada nuevo instante las observaciones que se obtienen del proceso y evaluarlas con reglas obtenidas en la etapa de análisis de manera que se puede identificar el estado actual de la planta.

Las diversas variantes que se han ido proponiendo en la literatura básicamente se diferencian en la técnica *CLD* a utilizar. La más usada en los trabajos precedentes es el *FCM* (Næs y Mevik, 1999; Teppola *et al*., 1999; Rosen, 2001), aunque también se ha probado el uso de *PCM* (Teppola y Minkkinen, 1999) y el *CFCM* (Choi *et al*., 2003; Yoo *et al*., 2003). En la comparación que se muestra en la siguiente sección se propone la evaluación de las variantes anteriores y otras más que consisten en el uso de la técnica *FPCM*, descrita en la sección 5.2.6, y en el uso de las todas las técnicas *CLD* anteriores (*FCM*, *PCM*, *CFCM* y *FPCM*) modificadas con la variante *GK* para el cálculo de la distancia propuesto por *Gustafson* y *Kessel* (ver sección 5.2.3). Para el caso del *PCM*, solo se utiliza la variante *GK* ya que la estrategia basada en distancias euclidianas conduce a muchísimos errores por lo que se descarta directamente del estudio de comparación. Asimismo, la comparación se lleva a cabo solo para la etapa de análisis. Más adelante (capítulo 6), se muestra el rendimiento de las estrategias TEM-*CLD* para la etapa de monitorización en línea.

5.5 Análisis comparativo de estrategias de clustering

En esta sección se presenta un estudio comparativo entre diversas estrategias TEM-*CLD*. Primeramente, se describen una serie de casos de estudio sobre los que se desarrolla la comparación. En una segunda parte se comparan las estrategias TEM-*CLD* en el análisis de los casos de estudio propuestos. En una tercera parte se repite el análisis anterior pero con presencia de *outliers* en los datos analizados. Finalmente, se presenta la comparación de estrategias de estimación del número de *clusters*.

5.5.1 Casos de estudio

5.5.1.1 Casos 1 y 2

Estos casos consisten de 2 conjuntos de datos con 2 variables, correspondiendo cada uno a un proceso diferente. En lo sucesivo, cada uno de los casos se etiqueta como **E1** y **E2**. Los casos son adaptaciones de datos típicos usados en varias demostraciones de programas que implementan *CLD* (MATLAB 7.0, Paquete R 1.8.1) y en varios trabajos de la literatura (Amiri, 2003; Balasko *et al*., 2004). El número de muestras es variable en cada caso. En las figuras Figura 5.5 y Figura 5.6 se muestran las curvas correspondientes a cada conjunto de datos.

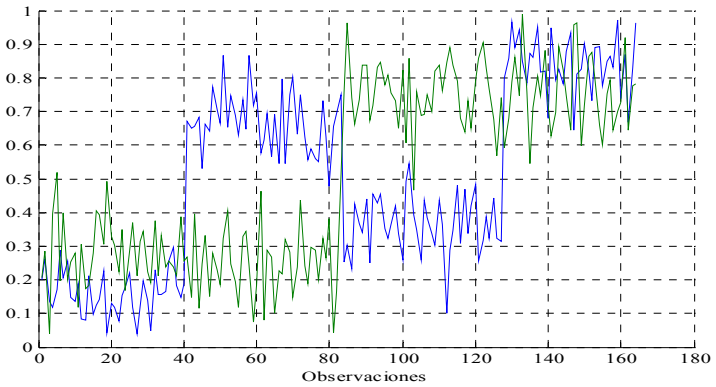

Figura 5.5. Datos del caso de estudio **E1**.

Es de esperar que en el caso **E1** se agrupen los datos en 4 regiones de operación distintas (como se puede deducir de la Figura 5.5). En cuanto al caso **E2**, se espera que la información se estructure en 3 grupos (ver Figura 5.6).

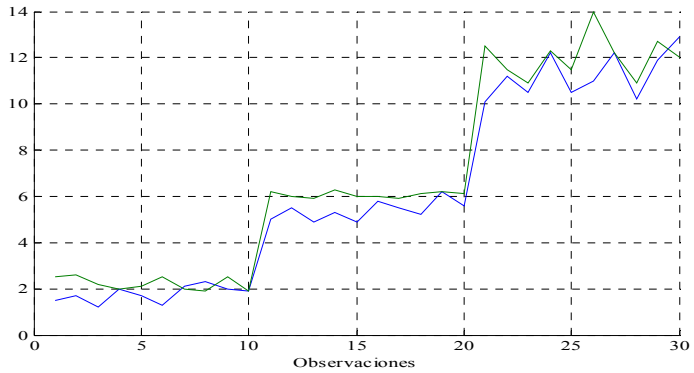

Figura 5.6. Datos del Caso de estudio **E2**.

5.5.1.2 Caso 3 - Operación de un reactor CSTR

Se muestra nuevamente el modelo del *CSTR* (ver sección 4.3.2.1 y anexo C). En este caso, se utiliza para generar el siguiente escenario de trabajo: Se fabrica un producto (*A*) a diferentes grados de calidad y para satisfacer distintas demandas de clientes. Los diferentes grados de producto se consiguen con solo cambiar la temperatura del reactor como se indica en la tabla 5.2. Se realiza una primera simulación donde se opera el reactor como sigue:

- El producto *A1* se fabrica continuamente y siempre que no haya demanda de *A2* ni *A3*, dado que mantiene una demanda constante, continua y alta.
- El producto *A2* es similar al *A1* pero su demanda se puede cubrir con una programación fija de aprox. 10 horas de producción cada 4 días.

- El producto *A3*, es un producto de mayor valor que los otros y obedece a una demanda variable tal que el día que se recibe su pedido se fabrica, siendo su tiempo de operación un periodo estable de alrededor de 12 ½ horas de trabajo.

Tabla 5.2. Diferentes Grados del producto A.

Producto	Temperatura (*T*)	Concentración Final (*Ca*)
A1	600 K	0.25
A2	610 K	0.2
A3	590 K	0.32

Dado lo anterior, la planta está sometida a 3 cambios en las condiciones de operación. Se desea estructurar los datos históricos según esta información básica. En lo sucesivo este caso se etiqueta como **E3**.

5.5.1.3 Caso 4 - Planta Química con reciclo (E4)

Este caso también se utilizó en el capítulo 4 (ver sección 4.3.2.3). Se etiqueta como **E4**. Se propone el siguiente escenario: Un cambio momentáneo de la apertura de la válvula de alimentación al reactor inesperadamente condujo a un grado de producto fuera de especificación. El cambio se debió a un error inducido por el operador. Posteriormente, un cliente probó y compró el producto. Se desea caracterizar el comportamiento del proceso durante el cambio anterior en las condiciones de operación y así poder adaptar el sistema de monitorización de manera que pueda indicar si en algún momento, tras producirse un error se llega a esta región aceptable de producto o simplemente para poder monitorizar la producción de este nuevo grado de producto si se pide en el futuro.

5.5.2 Comparación de estrategias TEM-CLD

En esta sección se evalúan las diferentes estrategias que se describen en la sección 5.4. Se utilizan todos los casos descritos en la sección 5.5.1. En cada uno de ellos ya se conoce su valor de *c* (ver la tabla 5.3) por lo que los métodos de *clustering* que se utilizan reciben como entrada este valor.

Tabla 5.3. Valores reales de *c* para cada caso de estudio.

Caso	**E1**	**E2**	**E3**	**E4**
c real	4	3	3	2

El *PCM* solo se utiliza con la variante *GK* ya que con distancias euclidianas produce demasiados errores sobre algunos casos. Los métodos *CFCM* y *FPCM* se aplican tanto con distancias euclidianas como con la variante GK. Para evaluar o validar las particiones obtenidas en cada caso se utilizan gráficos de pertenencia y del espacio de vectores característicos más los índices de pureza y eficiencia (ver sección 5.3.3) y el índice *PC* o coeficiente de partición (ver sección 5.3.1.1).

Los resultados se resumen en las tablas 5.4 a 5.7 y en las figuras que se muestran a lo largo de esta sección. En cada tabla, los Pr_1, Pr_2,…, representan las purezas de cada *cluster* individual (5.3.3.1), mientras que las ξ_1, ξ_2,…, representan las eficiencias de cada *cluster* individual (5.3.3.2). Pr_m representa la pureza promedio entre todos los *clusters* y ξ_m representa la eficiencia promedio entre todos los *clusters*. En cuanto a las figuras, contienen gráficos de *scores* con los objetos etiquetados según los resultados de una técnica *clustering* tal que C_k indica el grupo mientras que CC_k indica el prototipo del *cluster k* (el centro del grupo).

Corrección del *PC*

En la definición del cálculo del *PC*, que se presenta en la sección 5.3.1.1, se puede ver que este índice depende de las μ_{ik} obtenidas mediante una técnica *CLD*. En el caso de que la técnica *CLD* utilizada sea o la *CFCM* o la *CFCM-GK*, el *PC* que se calcule será en magnitud más pequeño a los *PC* obtenidos tras usar la *FCM* o la *FPCM*. Esto se debe al hecho de que en el caso de la *CFCM* (y de la *CFCM-GK*) el valor de μ_{ik} viene modificado por los valores de credibilidad, lo que hace que la magnitud de dichas pertenencias (y del *PC* resultante), sea siempre mucho más pequeña que la magnitud de las μ_{ik} asociadas a otras *CLD*. Luego, para la comparación que se muestra en las secciones 5.5.2.1 a 5.5.2.4, se propone la siguiente corrección a las μ_{ik} obtenidas con la *CFCM* o la *CFCM-GK* y previo al cálculo del correspondiente *PC*:

$$\mu_{ik}^{cr} = \frac{\mu_{ik}}{\psi_i} \tag{5.46a}$$

El superíndice *cr* es para indicar que las pertenencias se han corregido según la ecuación 5.46a.

Para el caso de la *PCM-GK*, se presenta otro problema con el *PC*. Todo parte de la restricción del método impuesta por la ecuación 5.19. De acuerdo a ésta, la suma de las τ_{ik} por cada muestra i (donde $1 \leq i \leq m$) puede llegar a ser menor igual o mayor a 1. Luego, si para un conjunto de datos en particular la suma de las tipicalidades de muchísimas muestras es mayor a 1, entonces el *PC* puede superar el valor de 1. De esta manera el *PC* no puede usarse como medida de comparación entre la *PCM-GK* y otras técnicas *CLD*. En vista de esto se introduce la siguiente corrección para las tipicalidades:

$$\tau_{ik}^{cr} = \frac{\tau_{ik}}{\sum_{k=1}^{C} \tau_{ik}} \tag{5.46b}$$

Al aplicar la ecuación 5.46b anterior las tipicalidades resultantes siempre suman 1 por cada muestra i. Con esto se asegura que los *PC* se sitúen siempre en el rango [0,1] y sirvan para la comparación entre la *PCM-GK* y otras *CLD*.

5.5.2.1 Análisis del caso E1

Si se analizan los resultados para el caso 1 (ver la tabla 5.4 y los gráficos 5.7 a 5.10) se observa que, en cuanto a eficiencia y pureza, tanto por cada *cluster* como las promedio, todos los métodos conducen a buenos resultados con valores muy cercanos o iguales a 100 %. Lo anterior se cumple usando cualquiera de las *CLD* propuestas, excepto con la *PCM-GK*, ya que con esta las *Pr* y las ξ de varios grupos tienden a ser bajas (ver la tabla 5.4).

Tabla 5.4. Validación de resultados de *clustering* para el caso **E1**.

				Pureza				Eficiencia			
	Pr_m	ξ_m	*PC*	Pr_1	Pr_2	Pr_3	Pr_4	ξ_1	ξ_2	ξ_3	ξ_4
FCM	99	99	0.78	97	97	100	100	100	98	98	100
FCM-GK	99	99	0.79	97	100	100	100	100	98	100	100
PCM-GK	74	84	0.60	63	62	100	72	58	88	100	92
CFCM	98	98	0.78	98	100	97	98	100	98	96	100
CFCM-GK	99	99	0.79	100	100	97	100	100	98	98	100
FPCM	99	99	0.78	100	100	97	98	100	98	98	100
FPCM-GK	99	99	0.79	100	100	97	100	100	98	100	100

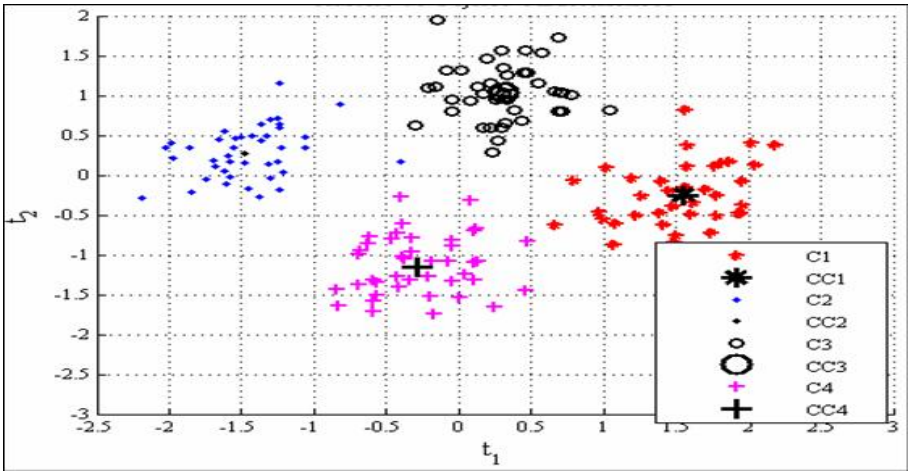

Figura 5.7. Partición de *clusters* mediante *FCM* – Caso **E1**

Los resultados obtenidos están indicando que prácticamente los grupos detectados sólo contienen muestras de un tipo de operación y las condiciones de operación forman regiones bien definidas y separadas entre sí. Esto se puede ver claramente en la figura 5.7 donde se muestran los diferentes grupos identificados según la salida del *FCM*. También, el gráfico de las μ_{ik} en la figura 5.8 muestra como las muestras asignadas a cada grupo tienen μ_{ik} significativamente mayores a las μ_{ik} de las mismas muestras al resto de los grupos. Por lo tanto, se pueden definir sin problemas las regiones de operación de cada producto y las reglas de control asociadas a cada región.

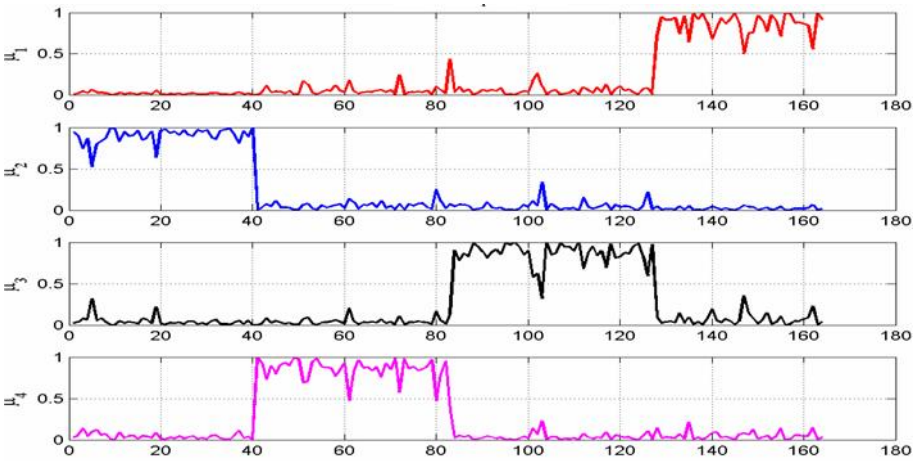

Figura 5.8. Pertenencias asociadas a cada grupo con *FCM* – Caso **E1**

En las particiones obtenidas mediante el *PCM-GK*, algunos grupos contienen muestras de diferentes condiciones de operación. En efecto, analizando la Figura 5.9, se observa que el grupo *C4* (etiquetado con 'o') esta formado por objetos que con el *FCM* se asignan a los grupos *C2* y *C3* (ver figura 5.7), mientras que *C3* (figura 5.9) es casi igual al grupo *C1* de la figura 5.7 pero con menos datos. Algo similar ocurre entre el *C2* de la figura 5.7 y el *C4* de la figura 5.9. Por último, el grupo *C1* aparece totalmente disperso entre *C2* y *C3*. Su centro ha quedado justo en una región media entre *C2* y *C3*. Todo esto está indicando que hay objetos coincidentes entre grupos y, consecuentemente, error en las particiones resultantes (ver gráfico de la figura 5.10), un problema que ya ha sido resaltado en la bibliografía sobre técnicas *CLD* (Barni *et al*., 1996; Pal *et al*., 1997; Amiri, 2003). Lo anterior, también se ve

reflejado en los valores más bajos de las Pr_i, ξ_i y el *PC*, del *PCM-GK* con respecto a las otras *CLD*. Por lo tanto, el método *PCM-GK* no es apropiado para manejar este caso de estudio.

Se debe notar que para obtener los resultados de la tabla 5.4, se aplicó cada técnica 5 veces sobre cada conjunto, variando en cada caso la estimación inicial de las pertenencias que se obtiene aleatoriamente (ver algoritmo del *FCM* en sección 5.2.2). Se observó que con cada método la variación en los resultados obtenidos por cada repetición eran nulos, excepto con el *PCM-GK* para el que los resultados nunca dieron igual (los resultados mostrados corresponden a la mejor prueba).

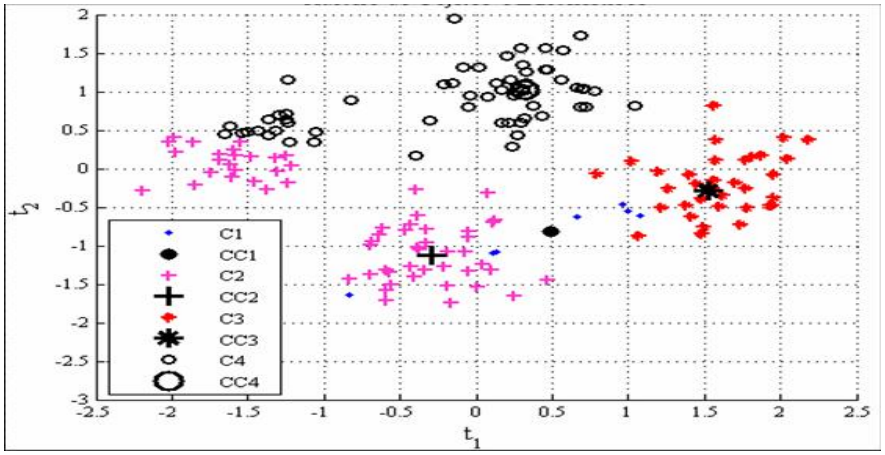

Figura 5.9. Partición de *clusters* mediante *PCM-GK* – Caso **E1**

En el análisis de este caso el *PC* de *CFCM* y el *CFCM-GK* se ha calculado según la corrección propuesta mediante la ecuación 5.46a (ver sección 5.5.2). Se observa que el *PC* así calculado es adecuado para hacer comparaciones entre *CFCM*, *CFCM-GK* y otros métodos de *CLD*. El valor de dicho *PC* al asociarse a los valores de las Pr_i y ξ_i, y en comparación a las otras *CLD*, indica muy buenos resultados de identificación.

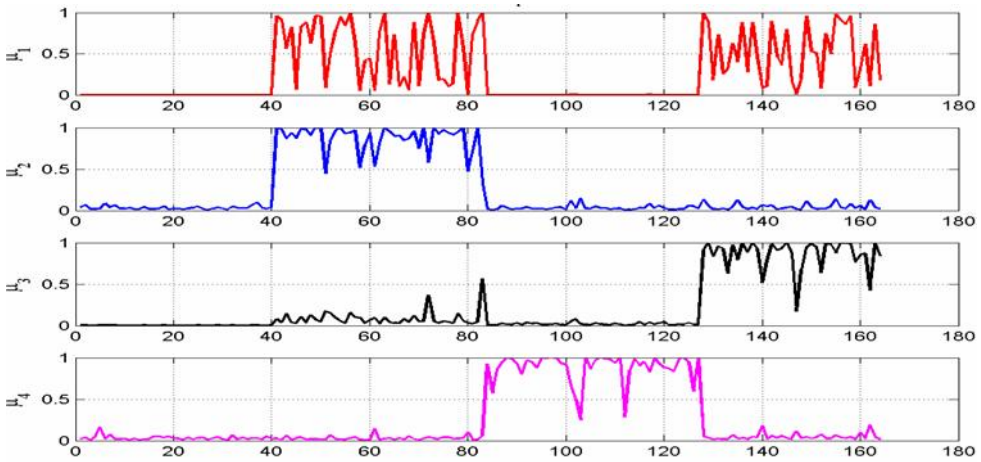

Figura 5.10. Pertenencias asociadas a cada grupo con *PCM-GK* – Caso **E1**

5.5.2.2 Análisis del caso E2

Como en el caso anterior, los resultados de la tabla se obtuvieron tras repetir la aplicación de cada técnica 5 veces, variando en cada caso la estimación inicial de las pertenencias que se obtiene aleatoriamente (ver algoritmo del *FCM* en sección 5.2.2). De la tabla 5.5 se desprende que en este caso todos los métodos logran obtener buenos resultados. Incluso el *PCM-GK*.

Solo en el caso de la técnica *CFCM*, algunas de las purezas y eficiencias calculadas son levemente menores al 100 %. Sin embargo, el resultado de esta misma técnica alcanza valores de 100 % tanto en Pr_m como en ξ_m tras utilizarse con la modificación *GK* (*CFCM-GK*).

Tabla 5.5. Validación de resultados de *clustering* para el caso **E2**.

				Pureza			Eficiencia		
	Pr_m	ξ_m	*PC*	Pr_1	Pr_2	Pr_3	ξ_1	ξ_2	ξ_3
FCM	100	100	0.97	100	100	100	100	100	100
FCM-GK	100	100	0.98	100	100	100	100	100	100
PCM-GK	100	100	0.99	100	100	100	100	100	100
CFCM	97	97	0.94	91	100	100	100	100	90
CFCM-GK	100	100	0.96	100	100	100	100	100	100
FPCM	100	100	0.97	100	100	100	100	100	100
FPCM-GK	100	100	0.98	100	100	100	100	100	100

5.5.2.3 Análisis del caso E3

Para este caso (ver la tabla 5.6), se observa que se tienen valores bajos de *Pr* y ξ con todos los métodos que se basan en la distancia euclidiana, mientras que los métodos que integran la estimación recursiva de la distancia mahalanobis (ver sección 5.2.3) obtienen muy buenos valores de *Pr* y ξ, excepto el *PCM-GK*. En la Figura 5.11 se muestra el gráfico de los *scores* agrupados según el *FCM*[a]. A diferencia del caso 1, los *scores* obtenidos tienden a agruparse en conjuntos elípticos. No obstante, el *FCM* produce una partición inadecuada para los grupos *C1* y *C2*. El uso de la distancia euclidiana fuerza al *FCM* a detectar *clusters* de forma esférica (ver comentario en la sección 5.2.2) y cuando los datos no se adaptan a esto, genera resultados erróneos como los de este caso. La Figura 5.12 contiene el gráfico de las μ_{ik} de cada muestra a cada grupo. Se observa que las μ_{ik} de los grupos 1 y 2 prácticamente se superponen entre sí, haciendo más patente el error en la partición.

Tabla 5.6. Validación de resultados de *clustering* para el caso **E3**.

				Pureza			Eficiencia		
	Pr_m	ξ_m	*PC*	Pr_1	Pr_2	Pr_3	ξ_1	ξ_2	ξ_3
FCM	87	69	0.62	83	78	99	58	51	99
FCM-GK	99	100	0.96	99	98	100	100	100	100
PCM-GK	*NaN*	97	0.59	100	54	*NaN*	91	100	100
CFCM	86	69	0.62	98	83	78	58	51	99
CFCM-GK	99	100	0.96	98	100	99	100	100	100
FPCM	87	69	0.62	83	99	78	58	51	99
FPCM-GK	99	100	0.96	99	100	98	100	100	100

En la Figura 5.13 se muestra el resultado de las particiones con el método *FCM-GK*, los grupos detectados se encuentran perfectamente diferenciados con valores de pureza y eficiencias sobre el 98 % (ver la tabla 5.5). Asimismo, el gráfico de la Figura 5.13 muestra como las observaciones se asignan correctamente a cada grupo mediante valores de pertenencia significativamente altos.

[a] Se muestra el gráfico tridimensional para una mejor visualización y debido a que en este caso multivariable el *ACP* genera 3 componentes principales y sus correspondientes *scores*.

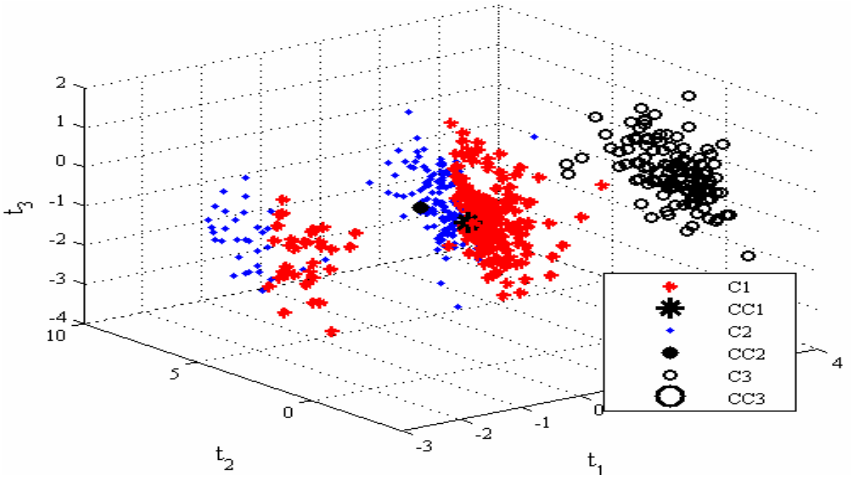

Figura 5.11. Partición de *clusters* mediante *FCM* – Caso **E3**

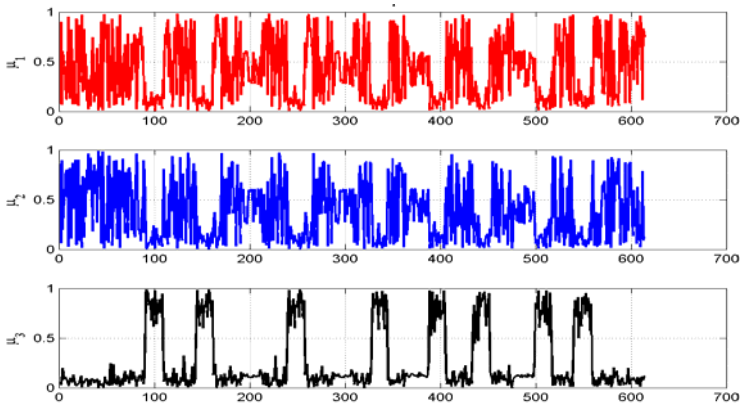

Figura 5.12. Pertenencia asociadas a cada grupo *FCM* – Caso **E3**

Es interesante ver que, a diferencia de los valores de pertenencia del caso 1 en que todas las muestras asociadas a un grupo aparecen juntas, las muestras que pertenecen a cada grupo aparecen en intervalos de muestreo separados. Esto se explica por el hecho de que, en este caso, se está analizando un proceso sometido a cambios frecuentes en las condiciones de operación. Pese a esto, el análisis permite estructurar correctamente el proceso, pudiéndose llegar a un diseño adecuado de un sistema de monitorización.

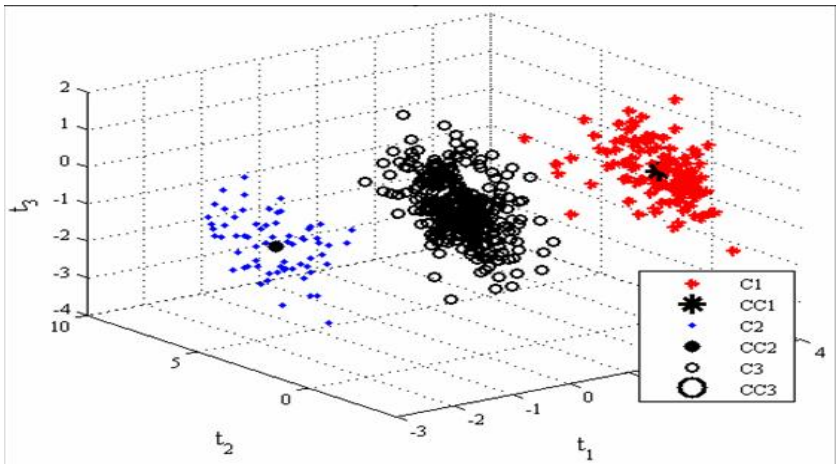

Figura 5.13. Partición de *clusters* mediante *FCM-GK* – Caso **E3**

El análisis anterior para el *FCM* se puede trasladar de manera semejante a la aplicación de las técnicas *CFCM* y *FPCM* sobre el caso de estudio actual. En efecto, los valores de las diferentes *Pr* y ξ con estas técnicas son semejantes a los obtenidos con el *FCM* (ver la tabla 5.6). Asimismo, la extensión de *CFCM* y *FPCM* mediante el procedimiento recursivo *GK* brinda mejoras significativas que se reflejan en valores de *Pr* y ξ por encima de 98 %. Aun cuando Teppola y Minkkinen ya recomendaron el uso de la distancia mahalanobis para mejorar la identificación (Teppola y Minkkinen, 1999), trabajos paralelos y posteriores no han tenido en cuenta esta recomendación por lo que muchas de las propuestas de estrategias de monitorización TEM-*CLD* que se presentan en la literatura fallarán en muchas situaciones prácticas debido a la limitación que provoca el uso de distancias euclidianas.

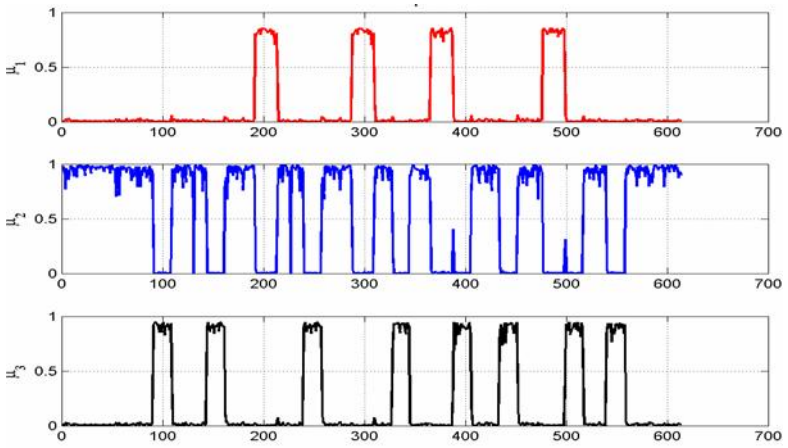

Figura 5.14. Pertenencia asociadas a cada grupo *FCM-GK* – Caso **E3**

Para este caso el *PCM-GK* vuelve a conducir a resultados erróneos. En el gráfico de la Figura 5.15 puede verse que los centros de los *clusters CC1* y *CC3* son totalmente coincidentes. Asimismo, el centro de *cluster* para *CC2* es muy cercano a los otros. Además, en el gráfico generado solo se observan etiquetas para 2 *clusters*: *C1* y *C2*, aun cuando antes de ejecutar la técnica el nº de *clusters c*=3 se especificó como parámetro de entrada al algoritmo del *PCM-GK*.

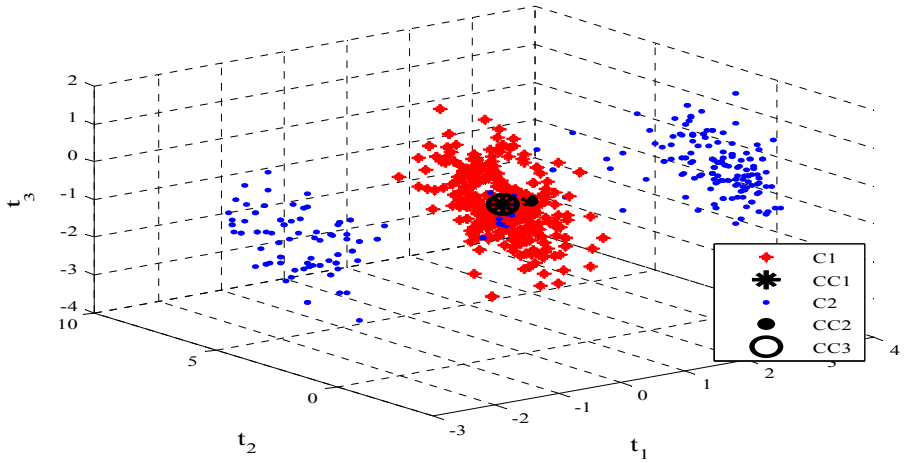

Figura 5.15. Partición de *clusters* mediante *PCM-GK* – Caso **E3**

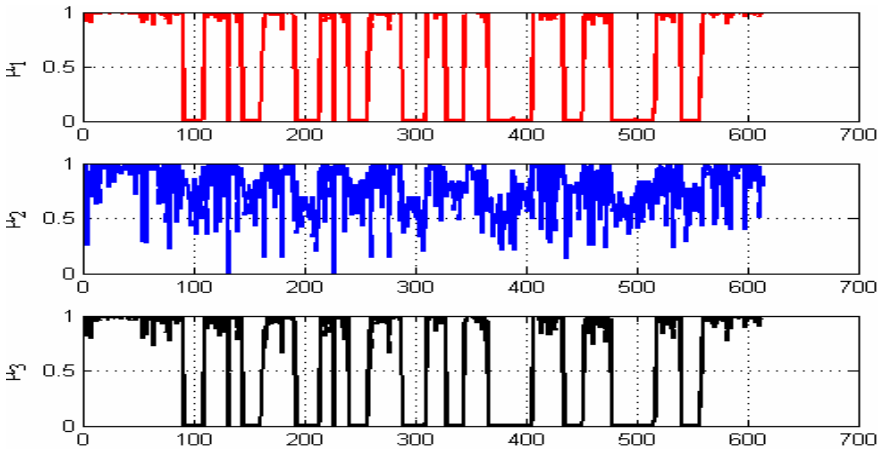

Figura 5.16. Pertenencia asociadas a cada grupo *PCM-GK* – Caso **E3**

En el gráfico de la figura 5.16 se puede observar que las pertenencias asociadas al grupo 1 (μ_1 en la figura 5.16) son siempre levemente mayores a las pertenencias del grupo 3 (μ_3). Esto esta indicando que de nuevo (como en el caso 1) se ha producido coincidencia de *clusters*. Debido a esto todos los datos de los grupos *C1* y *C2* que se observan en la figura 5.13 se han identificado como pertenecientes a un solo grupo *C2* en la figura 5.15, esto es, se han integrado los datos de 2 regiones de operación distintas. Luego, la pureza del tercer *cluster* no identificado no se puede calcular ya que el correspondiente Np_k (ver ecuación 5.43) es 0. Las casillas correspondientes a estos casos se identifican en las tablas de resultados como *NaN*. Así, al igual que en el caso 1, el *PCM-GK* no es útil para analizar los datos.

5.5.2.4 Análisis del caso E4

Este caso muestra resultados similares a los del caso anterior. Si se analizan los valores de *Pr* y ξ (ver tabla 5.7), se ve rápidamente que los *FCM*, *CFCM* y *FPCM* producen particiones inadecuadas, mientras que sus variantes basadas en el procedimiento recursivo *GK* mejoran significativamente los resultados. Fijándose en la Figura 5.17 se observa que las nubes de puntos de los *scores* correspondientes a los 3 primeros componentes principales obtenidos no son esféricas. Con esto queda una vez más en evidencia la utilidad de la modificación *GK* para aplicar *CLD* sobre datos con grupos de formas diversas. También se observa que, en este caso, la separación entre las nubes de puntos de los 2 grupos presentes es muy pequeña. No obstante, los *clusters* obtenidos tras usar *FCM-GK*, *CFCM-GK* o *FPCM-GK* logran identificar las 2 regiones de operación asociadas a los 2 estados existentes en este caso **E4**.

Tabla 5.7. Validación de resultados de *clustering* para el caso **E4**.

				Pureza		Eficiencia	
	Pr_m	ξ_m	*PC*	Pr_1	Pr_2	ξ_1	ξ_2
FCM	67	55	0.55	62	72	53	58
FCM-GK	100	100	1.00	100	100	100	100
PCM-GK	99	99	0.68	97	100	98	100
CFCM	67	56	0.53	72	62	53	58
CFCM-GK	100	100	1.00	100	100	100	100
FPCM	67	55	0.55	72	62	53	58
FPCM-GK	100	100	1.00	100	100	100	100

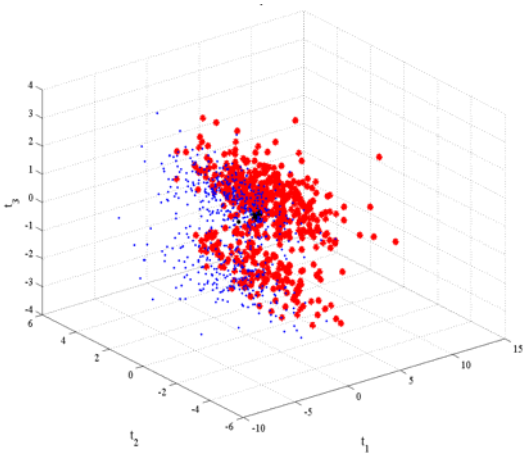

Figura 5.17. Partición de *clusters* mediante *CFCM* – Caso **E4**.

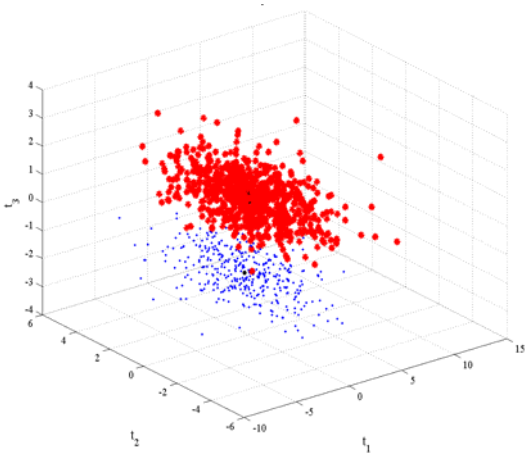

Figura 5.18. Partición de *clusters* mediante *CFCM-GK* – Caso **E4**.

En este caso también se observa que el *PCM-GK* da buenos resultados pero, como en el caso 2 éstos se obtuvieron tras hacer varias iteraciones del algoritmo correspondiente, en las que se varía la matriz inicial de pertenencias. La mayoría de estas iteraciones condujo a resultados erróneos de partición. Así, para obtener buenos resultados con el *PCM-GK* es necesario iterar varias veces hasta encontrar una matriz inicial de pertenencias apropiada que conduzca a una buena solución, con la desventaja adicional de que debe contarse con algún procedimiento paralelo que indique de alguna forma si la aplicación actual del *PCM-GK* es correcta o no lo es, lo que es innecesario con los otros métodos.

5.5.2.5 Observaciones sobre la comparación anterior

El análisis anterior deja clara varias cosas. Primero, que el uso de los métodos *FCM*, *FPCM* y *CFCM* modificados mediante la variante al cálculo de las distancias *GK* conduce a mejores estimaciones de las particiones en un conjunto de datos, que los mismos métodos *FCM*, *FPCM* y *CFCM* cuando se utilizan con distancias euclidianas. Esto se debe a que la restricción de solo identificar *clusters* de formas esféricas que impone el uso de la distancia euclidiana desaparece al usar la distancia mahalanobis estimada según la modificación *GK* (ver sección 5.2.3).

Por otro lado, se ha visto que el *PCM-GK* tiende a producir particiones erróneas y resultados inapropiados en muchos casos. La observación sobre el rendimiento del *PCM-GK* es importante ya que en el trabajo de Teppola y Mikkinen se había reportado la utilidad de esta técnica para monitorizar un caso específico sujeto a cambios frecuentes en las condiciones de operación (Teppola y Minkkinen, 1999). No obstante, si en la etapa de diseño del sistema de monitorización no se alcanza una buena definición de las regiones de operación (lo que ha ocurrido en los casos analizados), es inútil pensar en usar el sistema resultante para monitorizar. Por otro lado, dado el buen rendimiento de *FCM-GK*, *CFCM-GK* y *FPCM-GK*, se recomienda el uso de estas técnicas para el diseño de sistemas de monitorización de procesos sujetos a cambios frecuentes en las condiciones de operación.

5.5.3 Comparación de estrategias TEM-CLD en el manejo de outliers

En las estrategias que se discuten en la sección 5.4 y para la etapa de análisis se asume que los datos históricos utilizados solo contienen información de operaciones normales o bien no están contaminados por valores atípicos (*outliers*). En la comparación de la sección anterior se trabajó bajo esta suposición. En esta sección se vuelven a analizar los mismos casos de

estudio pero ahora con datos que contienen algunos *outliers*. Para esta comparación solo se utilizan las estrategias que en la sección precedente arrojaron resultados satisfactorios, esto es, las que utilizan *FCM* con *GK*, *CFCM* con *GK* y *FPCM* con *GK*.

5.5.3.1 La identificación de los outliers

En los trabajos de Yoo *et al.*, se hace una breve discusión teórica sobre la ventaja de utilizar el *CFCM* en caso de presencia de *outliers* en los datos (Choi *et al*., 2003; Yoo *et al*., 2003). No obstante, ni describe como hacer el tratamiento de éstos ni mucho menos presenta casos afectados por este tipo de anormalidades. Para la comparación que se muestra en esta sección se propone el siguiente procedimiento de identificación de *outliers* (etiquetado como *OutM1*):

1. Se toma la matriz de datos de proceso **Y**. Se calcula el correspondiente modelo ACP.
2. Se aplica el *clustering* sobre los *scores* del modelo ACP que se obtuvo en el paso anterior. Se utiliza como técnica *clustering* o el *FCM* con *GK*, el *CFCM* con *GK* o el *FPCM* con *GK*.
3. Se toman los diferentes valores de $\boldsymbol{\mu}_i$, y se calcula la siguiente medición (Choi *et al*., 2003):
$$up_i = \mu_{i,1} \cdot \mu_{i,2} \cdot \ldots \cdot \mu_{i,c} \qquad (5.47)$$
donde *up* es el producto de las pertenencias asignadas a la muestra *i*.
4. Se calcula un límite para *up* ($lim_{\boldsymbol{up}}$) de manera similar a como se calculan los límites para los *SPE* y T^2 (ver sección 4.3.1.1), esto es, se calcula un límite al 99 % de confianza donde dicho límite se basa en la distribución empírica de los *up*.
5. Se evalúan los valores de *up* respecto de $lim_{\mathbf{up}}$. Las correspondientes observaciones a las que *up* supera a $lim_{\mathbf{up}}$ se consideran *outliers* y se eliminan de los datos.

En el procedimiento anterior el uso de $\boldsymbol{\mu}_{i,k}$, obedece a que con ésta se reduce el análisis de *c* vectores ($\boldsymbol{\mu}_{i,k}$) a solo uno (*up*) conteniendo la misma información. Asimismo, ya se ha visto en trabajos precedentes de la literatura (Choi *et al*., 2003) que el análisis del *up* puede arrojar información valiosa sobre desviaciones del proceso (*outliers*, transiciones por cambios de operación, etc.).

En el presente trabajo se ha comprobado que dicha medición (*up*), dentro del procedimiento anterior, puede llegar a ser útil para la detección de *outliers*. No obstante, también se vio que la medición anterior estimada según la ecuación 5.47 solo es útil para los casos de las técnicas *FCM-GK* y *FPCM-GK*. Para el caso del método *CFCM-GK* se estableció mediante experimentación que era más útil la siguiente medición:

$$up_i = 1/(\psi_i + 0.01) \qquad (5.48)$$

En este caso, el *up* es función inversa de los valores de la credibilidad ψ_i de cada muestra *i*. En los valores de ψ_i los *outliers* se reflejan como puntos tendientes a 0. Si se hace el inverso, la medición resultante para el caso de los *outliers* se podría ver como un pico hacia arriba muy pronunciado. Al sumarle a la ψ_i una constante tan pequeña como 0.01 se busca que cuando se tenga un *outlier*, el valor de up sea mucho mayor al resto de los datos no *outliers*.

Asimismo, se llegó a ver que utilizando el vector de distancias d_{ik}, que se calcula una vez se tiene definido el modelo de *cluster* mediante cualquiera de las técnicas *CLD* que se consideran en esta sección, se pueden generar mediciones similares a las anteriores *up* que pueden llegar a conducir a tanta o más eficiencia en identificación de *outliers* mediante la estrategia descrita al inicio de esta sección. Así, se proponen las siguientes variantes de *up*:
- Para el caso de los métodos *FCM-GK* y *FPCM-GK*:

$$up_i = d_{i,1} \cdot d_{i,2} \cdot \ldots \cdot d_{i,c} \tag{5.49}$$

- Para el caso de los métodos *CFCM-GK*:

$$up_i = 1 \Big/ \left(\sum_{k=1}^{c} d_{i,k} + 0.01\right) \tag{5.50}$$

Con esto se tienen los 6 procedimientos de identificación de *outliers* que se listan en la tabla 5.8.

Tabla 5.8. Métodos de identificación de *outliers*.

Método	Estrategia TEM-*CLD*	Ecuación para $\mathbf{u}_p$
OutM1	ACP – *FCM-GK*	$up_i = \mu_{i,1} \cdot \mu_{i,2} \cdot \ldots \cdot \mu_{i,c}$
OutM2	ACP – *CFCM-GK*	$up_i = (\psi_i + 0.01)^{-1}$
OutM3	ACP – *FPCM-GK*	$up_i = \mu_{i,1} \cdot \mu_{i,2} \cdot \ldots \cdot \mu_{i,c}$
OutM4	ACP – *FCM-GK*	$up_i = d_{i,1} \cdot d_{i,2} \cdot \ldots \cdot d_{i,c}$
OutM5	ACP – *CFCM-GK*	$up_i = \left(\sum_{k=1}^{c} d_{i,k} + 0.01\right)^{-1}$
OutM6	ACP – *FPCM-GK*	$up_i = d_{i,1} \cdot d_{i,2} \cdot \ldots \cdot d_{i,c}$

5.5.3.2 Medidas de evaluación de la comparación

Para poder hacer una comparación se propone primeramente tomar un método de referencia. Para tal fin se selecciona al método basado en ACP que se describe en el anexo D y se etiqueta como *OutMr*. Adicionalmente, se propone utilizar las siguientes mediciones:

Eficiencia de Detección de *Outliers* (EDO):

$$EDO(\%) = \left(\frac{Nodr}{Not}\right) * 100\% \tag{5.51}$$

Porcentaje de Datos Normales Eliminados (DNE):

$$DNE(\%) = \left(\frac{Nod - Nodr}{m}\right) * 100\% \tag{5.52}$$

En las anteriores expresiones *Nodr* representa el número de *outliers* identificados mediante un método aplicado y *Not* representa al número real (total) de *outliers* en el caso considerado (ver tabla 5.9). Si el método detecta todos los *outliers* la eficiencia *EDO* será máxima (100%). Para el caso del *DNE*, *Nod* representa el número de observaciones detectadas como *atípicas* mediante un método aplicado. Dicho número puede ser mayor a *Nodr* debido a identificaciones erróneas, esto es, una técnica dada podría identificar valores normales como atípicos. Así, cuando *Nod>Nodr*, el método aplicado ha detectado observaciones normales como atípicas llegando así a eliminarse datos que no se deberían eliminar. Luego, el *DNE* se mide en relación al porcentaje real de *outliers Pot* en los datos:

- Si *DNE* > *Pot*, la técnica en uso erróneamente esta identificando datos correspondientes a operaciones normales como *outliers*.
- Si *DNE* = *Pot*, entonces la técnica en uso solo detectan *outliers* y por tanto no habrá peligro de eliminar datos correspondientes a operaciones normales.

5.5.3.3 Los casos de estudio

Como se indicó al inicio de esta sección, los casos que se utilizan son los mismos de la sección 5.5.1. No obstante, en este caso se añaden un número variable de *outliers* a cada caso para poder establecer la comparación en presencia de valores atípicos (ver tabla 5.9).

Tabla 5.9. Valores atípicos (*outliers*) añadidos a cada caso

Caso	*Pot*	Nº total de Observaciones	*Pot*
E1	*3*	176	1.7
E2	*1*	31	3.3
E3	*4*	614	0.7
E4	*2*	1123	0.2

5.5.3.4 Comparación de estrategias

Aquí se comparan las estrategias descritas en la sección 5.5.3.1, más la *OutMr* que se describe en el anexo D. Para cada caso se aplican las estrategias anteriores y como resultado se obtiene la tabla 5.10.

Tabla 5.10. Métodos de identificación de *outliers*.

Método	**E1**		**E2**		**E3**		**E4**	
	EDO (%)	*DNE (%)*	*EDO (%)*	*DNE (%)*	*EDO (%)*	*DNE (%)*	*EDO (%)*	*DNE (%)*
OutMr	100	1.1	100	6.5	50	1.5	100	1.7
OutM1	33	1.1	0	0	25	2.3	100	1.6
OutM2	66	0.6	100	0	25	0.8	50	0.1
OutM3	66	1.1	100	0	50	2	100	1.7
OutM4	100	0.6	0	3.3	100	0	100	0.1
OutM5	0	1.7	0	6.5	0	0.8	0	1.5
OutM6	100	0.6	0	3.3	100	0	100	0.1

Se puede ver que ninguno de los métodos logra identificar todos los *outliers* presentes en cada conjunto de datos, esto es, ningún método *OutMi* alcanza valores de *EDO* iguales al 100 % para todos los casos de estudio. En este sentido el *OutMr* es el que obtiene mejores resultados ya que en 3 casos (**E1**, **E2** y **E4**) logra identificar todos los *outliers* presentes con valores de *EDO* = 100%, y en el caso de **E3** identifica la mitad de los *outliers* presentes. No obstante, de la tabla se observa que por los valores de *DNE* asociados, el *OutMr* elimina en todos los casos datos normales tras haberlos identificado incorrectamente como *outliers*. Incluso, en 2 casos (**E2** y **E4**) es el que elimina más datos erróneos por lo que debe cuidarse este aspecto a la hora de usar esta técnica. Luego, si el sistema resultante se utiliza para monitorizar, la considerable eliminación de datos normales podría provocar valores más bajos para los límites de control del *SPE* y el T^2 en los casos en los que estos se quieran utilizar junto con las pertenencias. Esto indudablemente conducirá a generar más falsas alarmas de las que se producirían si se hubiesen descargado menos datos normales.

Otro de los resultados que más saltan a la vista, tras un primer vistazo a la tabla 5.10, es el hecho de que el método *OutM5* produce valores de *EDO* = 0 en todos los casos, mientras los *DNE* son siempre mayores a 0. Esto indica que dicho método no es capaz de identificar ningún *outlier* y, por el contrario, llega a clasificar valores normales como atípicos. En consecuencia el *up* utilizado, basado en un inverso de la suma de las distancias que se obtienen de la técnica *clustering* en uso, es un criterio totalmente inaceptable para intentar identificar *outliers*.

El método *OutM3* alcanza valores altos de *EDO* en todos los casos con 100 % para los casos de **E2** y **E4**. No obstante, al igual que el *OutMr* obtiene valores muy significativos de *DNE* en todos los casos lo que se traduce en la eliminación de muchos valores normales de cada conjunto de datos.

De los métodos restantes, lo más destacables es que el *OutM4* y el *OutM6* obtienen valores de *EDO* = 100 % en 3 de los 4 casos (**E1**, **E3** y **E4**), siendo los únicos que logran identificar todos los *outliers* en el caso con más atípicos (**E4**). Además, comparativamente son los que brindan valores más bajos de *DNE*. No obstante, en el caso con menos muestras disponibles y menos outliers (**E2**) no logran identificar el *outlier* presente. Esto indica que pese a la buena relación *EDO – DNE* que brindan para la mayoría de los casos, se debe tener cautela si se llega a adoptar alguno de estos métodos (el *OutM4* o el *OutM6*) para el análisis de un conjunto de datos.

Por último, se observa que el *OutMr* podría complementarse con el *OutM4* o el *OutM6* de cara a asegurar detección en cada caso y con *DNE* no significativamente altos lo cual sería una opción valida de tratamiento de *outliers* cuando ellos estén presentes. Algo similar se podría deducir para *OutMr* junto con *OutM3* o *OutM5*. Sin embargo, *OutM3* o *OutM5* no son buenos en la detección del caso **E3**, que es el mismo donde falla *OutMr* de modo que la combinación probablemente no supere esta carencia. Así, se hace la prueba de detectar combinando *OutMr* con *Outm4* (etiquetada como *OutMC1*) y *OutMr* con *OutM6* (etiquetada como *OutMC2*). Los resultados se muestran en la tabla 5.11.

Tabla 5.11. Métodos de identificación de *outliers*.

Método	*Caso* **E1**		*Caso* **E2**		*Caso* **E3**		*Caso* **E4**	
	EDO (%)	*DNE (%)*	*EDO (%)*	*DNE (%)*	*EDO (%)*	*DNE (%)*	*EDO (%)*	*DNE (%)*
OutMC1	100	1.1	100	6.5	100	1.3	100	1.4
OutMC2	100	1.1	100	6.5	100	1.3	100	1.4

Se muestra que dichas combinaciones logran acertar en la detección de *outliers* de todos los casos junto valores de *DNE* aceptables en todos los casos, por lo cual se propone que para aplicar estrategias de análisis de datos basadas en TEM-*CLD*, lo mejor sería usar cualquiera de las 2 combinaciones ACP-*FCMGK* o ACP-*FPCM-GK* y utilizar para el manejo de *outliers* o bien la estrategia *OutMC1* o bien *OutMC2*. La técnica *CFCM-GK* no se recomienda ya que con los up asociados (ver resultados de *OuMr2* y *OutMr5* en la tabla 5.10) no se obtienen buenos resultados de detección de *outliers*.

5.5.4 Análisis de estrategias de estimación del número de clusters

En las comparaciones anteriores siempre se asumió que el valor del número de regiones de operación *c* era conocido. Esto se cumple en muchísimas situaciones reales. Sin embargo, siempre puede haber situaciones en las que no se conozca el valor de *c*. En esta sección se evalúan las estrategias de estimación de *c* que se describen a lo largo de la sección 5.3. Se utilizan todos los casos descritos en la sección 5.5.1. Para cada uno de estos casos ya se conoce su valor de *c* (ver la tabla 5.3) por lo que los métodos deberían verificar este valor.

En primer lugar se evalúa el *MSCl* según la modificación propuesta en la sección 5.3.2.1, y en 3 maneras distintas:

- *MSCl-1*: En esta primera opción, se toma la matriz de datos **Y** disponible y se procesa directamente con el *MSCl*.
- *MSCl-2*: En esta segunda opción, se toma la matriz de datos **Y** disponible y se preprocesa con ACP. Luego, se trabaja sobre los *scores* obtenidos del modelo ACP.
- *MSCl-3*: Esta opción, es similar a la anterior con la diferencia de que primero se obtiene una aproximación levemente suavizada de cada variable individual en **Y**, mediante la *wavelets db1* con nivel de descomposición L=2. Luego, se procede como en *MSCl-2*.

Cada una de las variantes anteriores se aplica para distintos valores de α. De esta manera se podrá ver si existe algún valor típico de α para el cual el $\mathbf{r_a}$ resultante conduzca a la buena estimación de c para la mayoría de los casos (ver ecuación 5.37, sección 5.3.2). En la tabla 5.12 y las figuras 5.19 a 5.22 se muestra el resultado de la aplicación de cada uno de los *MSCl* a los casos de estudio considerados.

Tabla 5.12. Estimados de c con distintas variantes de *MSCl* y valores de α.

CASO	Método	α													
		0.70	0.75	0.80	0.81	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89	0.90	0.95
E1	*MSCl-1*	4	4	4	4	4	4	4	4	4	4	4	4	4	3
	MSCl-2	4	4	4	4	4	4	4	4	4	4	4	4	4	3
	MSCl-3	4	4	4	4	4	4	4	4	4	4	4	4	4	2
E2	*MSCl-1*	3	3	3	3	3	3	3	3	3	2	2	2	2	2
	MSCl-2	5	4	3	3	3	3	3	3	2	2	2	1	1	1
	MSCl-3	4	4	3	3	3	3	3	3	2	2	2	2	2	2
E3	*MSCl-1*	14	13	10	9	9	9	9	9	7	7	7	7	6	5
	MSCl-2	7	6	6	5	5	5	4	4	4	4	4	3	3	1
	MSCl-3	6	5	3	3	3	3	3	3	3	3	3	3	3	1
E4	*MSCl-1*	230	85	43	37	34	31	27	24	19	18	15	11	9	1
	MSCl-2	6	4	4	4	4	4	3	2	2	2	1	1	1	1
	MSCl-3	5	4	3	2	2	2	2	2	2	2	2	2	2	1

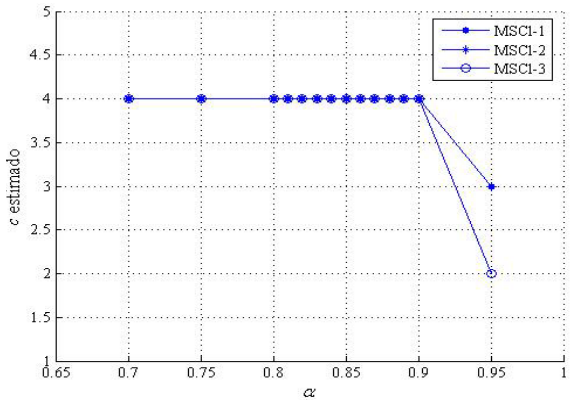

Figura 5.19. Estimados de c. Caso **E1**.

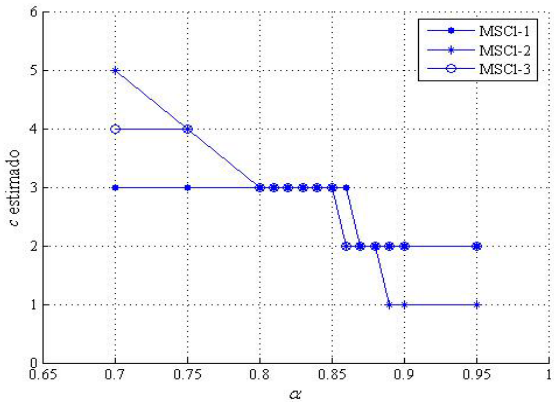

Figura 5.20. Estimados de c. Caso **E2**.

Al comparar los valores de c en la tabla 5.3 con los diferentes estimados de la tabla 5.12 se puede ver que, para un valor de $\alpha = 0.85$, los estimado mediante el método *MSCl-2* solo fallan en una ocasión (ver la tabla 5.10), mientras que los estimados con *MSCl-3* son siempre correctos. En efecto, los estimados con *MSCl-2* solo son erróneos para el caso **E3** en el que estiman 4 *clusters* en lugar de 3 (véase también la figura 5.20). De resto, todos los estimados

son precisos para el caso de $\alpha = 0.85$. También se observa que para este mismo valor de α, el método *MSCl-1* produce resultados demasiado erróneos en al menos 2 casos: **E3** (estima 9 *clusters* en lugar de 3) y **E4** (estima 24 *clusters* en lugar de 3).

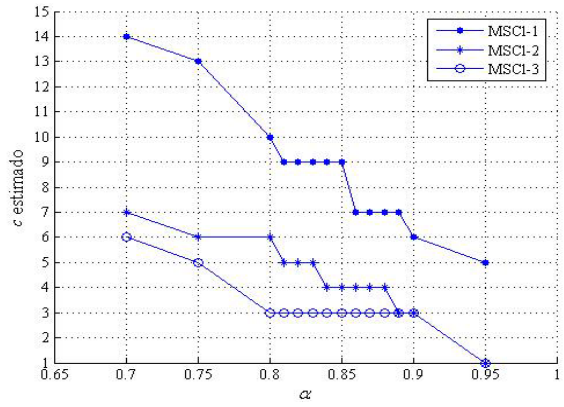

Figura 5.21. Estimados de *c*. Caso **E3**.

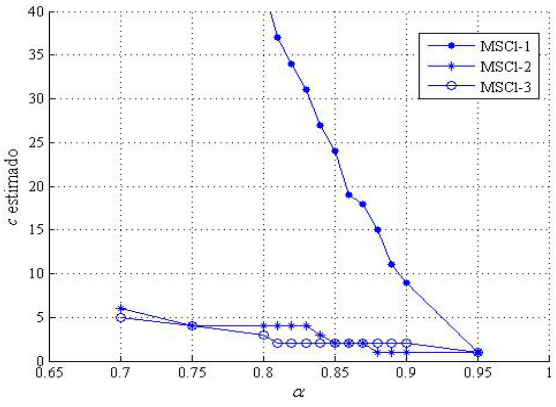

Figura 5.22. Estimados de *c*. Caso **E4**.

Si se utilizan otros valores distintos de $\alpha = 0.85$, el número de estimaciones cambia continuamente y de manera distinta para cada caso. En la figura 5.19 se observa que para el caso **E1** la variación de α no influye en los buenos resultados de ninguna de las alternativas planteadas para el *MSCl*, salvo para cuando $\alpha = 0.95$. En el caso **E2** todas las alternativas ofrecen buenos resultados cuando los valores de α están entre 0.8 y 0.85 (ver figura 5.20). No obstante, si se miran las figuras 5.21 y 5.22 se verá que para los casos considerados (**E3** y **E4**) α influye significativamente en los resultados. En especial, se puede ver con claridad que los estimados de *c* mediante *MSCl-1* son extremadamente incorrectos para los casos **E3** y **E4**.

A continuación, se pasa al análisis de los resultados de la estimación de *c* utilizando las estrategias basadas en índices de validación, que se discutieron en la sección 5.3. En este caso se hizo lo siguiente: se implementó el procedimiento descrito al inicio de la sección 5.3.1. Dicho procedimiento se aplicó 3 veces sobre cada caso: una vez utilizando $I_{VAL} = PC$, otra utilizando $I_{VAL} = CE$ y una última utilizando $I_{VAL} = XB$. Los resultados obtenidos se muestran en las figuras 5.23 a 5.26.

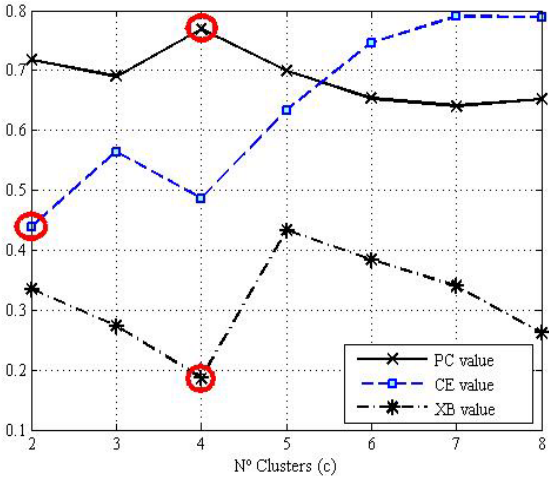

Figura 5.233. Estimados de *c*. Caso **E1**.

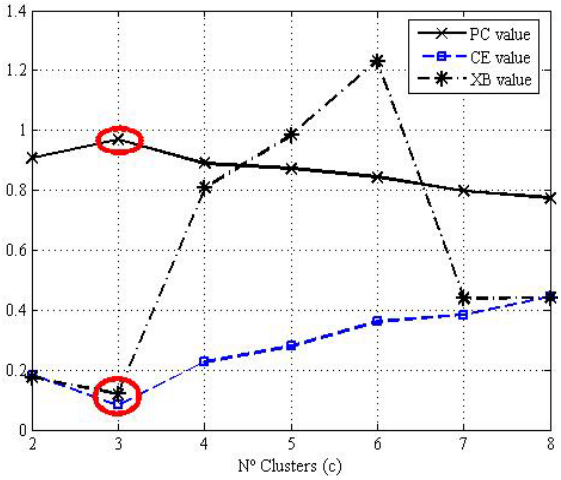

Figura 5.244. Estimados de *c*. Caso **E2**.

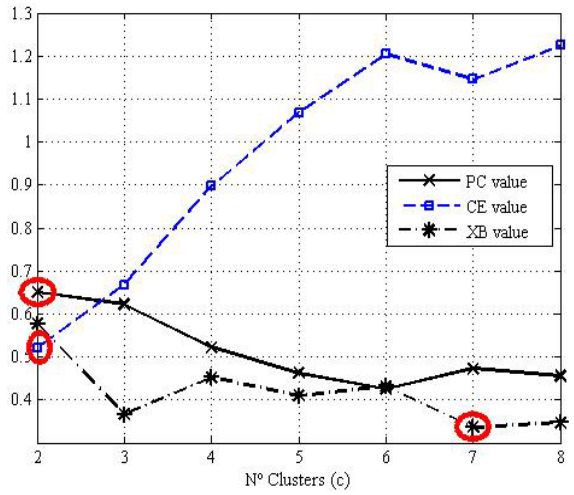

Figura 5.25. Estimados de *c*. Caso **E3**.

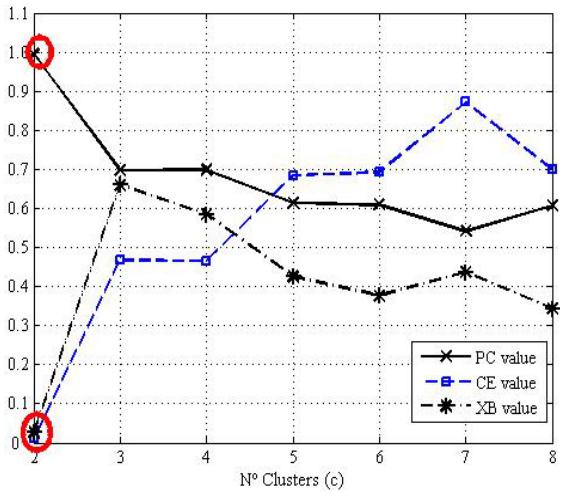

Figura 5.26. Estimados de *c*. Caso **E4**.

De acuerdo a la teoría expuesta a lo largo de la sección 5.3.1, se espera que la identificación adecuada de *c* se vea claramente a través de un máximo en la curva para el caso del *PC* y de un mínimo para el caso de los índices *CE* y *XB*. Sujetándose a estos criterios y explorando cada una de las curvas se identifican los correspondientes máximos y mínimos (puntos encerrados en un círculo) y se listan en la tabla 5.13.

Tabla 5.13. Estimados de *c* basados en índices *PC*, *CE* y *XB*.

Índice	**E1**	**E2**	**E3**	**E4**
PC	4	3	2	2
CE	2	3	2	2
XB	4	3	7	2

Se observa que tanto con el índice *PC* como con el índice *XB* se alcanza una identificación correcta en casi todos los casos (**E1**, **E2** y **E4**), aunque llega a fallar en el caso **E3**. No obstante, si se observa con atención la curva de *XB* en la figura 5.25 se puede ver que esta muestra un primer mínimo de la curva que coincide con el valor esperado de *c*. Si eventualmente se considera el primer mínimo de *XB* como criterio para determinar el c, y se reanalizan los gráficos éste nuevo criterio (primer mínimo de *XB*) inequívocamente podría conducir a los mismos buenos resultados para todos los casos con el añadido de un buen resultado para **E3**. En cuanto al índice *CE*, éste solo acierta en los casos **E2** y **E4**. Esto conduce a concluir que la estimación de c utilizando índices *PC* y *XB* no siempre conducirá a buenos resultados para algún caso particular, salvo que se verifique el criterio del primer mínimo de *XB* en los casos que se estudien.

Por otro lado, si se comparan los resultados entre los índices *PC* y *XB*, y la variante al método *Subtractive Clustering MSCl-2*, se deduce que es más seguro trabajar con la opción basada en el *MSCl* ya que esta logra la identificación correcta en todos los casos estudiados sin tener que llegar a hacer posibles cambios en el criterio utilizado como ha sido el caso del *XB*.

5.6 Conclusiones

En este capítulo se ha presentado un análisis comparativo de estrategias utilizadas para el análisis y monitorización de procesos multioperacionales. En una primera parte se discute teóricamente sobre las ventajas y desventajas de las distintas estrategias de este tipo existentes en la literatura. De este análisis se desprende claramente las ventajas y el atractivo que ofrecen un tipo de estrategias que combinan técnicas TEM con técnicas de *clustering* basadas en lógica difusa (TEM –*CLD*).

Se hace un estudio comparativo entre las distintas técnicas TEM-*CLD* existentes en la literatura y junto con otras variantes propuestas en este trabajo. Las comparaciones se hacen en términos de analizar e identificar correctamente la estructura de agrupamiento de datos procedentes de diversos casos de estudio y donde dichas estructuras responden a distintas regiones a las que opera cada proceso. En una primera comparación se verifica que el uso de los métodos *FCM*, *CFCM* y *FPCM* modificados mediante la variante al cálculo de las distancias *GK* conduce a mejores resultados en las estrategias TEM-*CLD* resultante que las propuestas existentes en la literatura. En efecto, la modificación GK permite a cada una de estas técnicas una correcta identificación de los diferentes grupos presentes en los datos. Hay que resaltar que en la literatura precedente sobre *CLD* y estrategias TEM-*CLD* no se había utilizado el *FPCM* mediante la variante al cálculo de las distancias *GK* por lo que la comparación permite descubrir el potencial del FPCM-GK para el tipo de análisis que se describe en este capítulo.

Posteriormente, se demuestra que el uso del *FCM-GK* y/o del *FPCM-GK*, combinado con información de estadísticos provenientes del ACP, permite una buena identificación de grupos y un buen manejo de los datos en presencia de *outliers*. Con esto se asegura la fiabilidad de estrategias TEM-*CLD* en presencia de outliers, algo que se ha obviado en la literatura reciente sobre aplicación de estrategias TEM-*CLD* para el análisis y supervisión de procesos multioperacionales.

Finalmente, se establece un análisis comparativo entre diversas propuestas de la literatura para la estimación del número de *clusters* en un conjunto de datos. Se muestra como mediante el uso de una pequeña variante al método *Subtractive Clustering* que se proponen en este capítulo, se asegura una mejor estimación del número de clusters grupos a formar. También se muestra cómo redefiniendo levemente el criterio de búsqueda basado en el índice *XB* se puede llegar a tener buenas estimaciones del número de *clusters* en un conjunto de datos.

Todos estos resultados permiten mejorar el rendimiento de las estrategias TEM-*CLD* actuales para aplicaciones de análisis de procesos multioperacionales y mediante la introducción de modificaciones bastante simples a las estrategias actuales. En el siguiente capítulo se aprovechan todos los resultados alcanzados en este capítulo para proponer estrategias de supervisión aplicadas a distintas situaciones de proceso.

NOMENCLATURA

A Matriz que induce la norma (en la función objetivo de una técnica *CLD*).
c Nº de *clusters*.
CE Coeficiente de Entropía de la Partición.

d_{ik} Distancia entre una observación *i* y un *cluster k*.
D Matriz conteniendo las distancias de cada observación x_i al resto de observaciones en **Y** para la variante al método *MSCl* (sección 5.3.2.1).
EDO Eficiencia de Detección de *Outliers*.
$\mathbf{F}_k$ Matriz covarianza del *cluster k* en los métodos basados en la modificación *GK*.
I_{VAL} Índice genérico de validación.
$J()$ Función objetivo de la optimización en la formulación de una técnica *CLD*.
$lim_{\mathbf{u}p}$ Valor umbral para detectar *outliers* sobre el vector $\mathbf{u}_p$.
$\mathbf{M}_{hc}$ Espacio de la partición rígida de **Y**.
$\mathbf{M}_{fc}$ Espacio de la partición difusa de **Y**.
m Nº de observaciones, objetos o muestreos disponibles.
N_k Nº de observaciones en un *cluster k*.
N_{DBj} Nº total de objetos pertenecientes a la condición de operación *j*.
N_{jk} Nº de objetos con una condición de operación *j* que están presentes en el *cluster* k.
Nod Nº de observaciones detectadas como atípicas mediante un método aplicado.
Nodr Nº de *outliers* reales detectados.
Not Nº real de *outliers*.
Npr_k Nº de objetos en el *cluster k*.
n Nº de variables medidas.
PC Coeficiente de Partición.
DNE Porcentaje de Datos Normales Eliminados.
P_i Medida del Potencial del *i*-ésimo punto para el método *MSCl*.
$P_{i\text{-}corr}$ Medida de Potencial corregida respecto a un P_1^* para el método *MSCl*.
Pot Porcentaje real de *outliers*.
Pr_k Pureza del *cluster k*.
P_i^* Medida de Potencial para el punto i con máximo *P* en el método *MSCl*.
r_a, r_b Radios definiendo una vecindad para el método *MSCl*.,donde $r_b = f(r_a)$.
SPE Estadístico que representa el error de predicción al cuadrado (*Squared Predictive Error*) de un modelo ACP.
S_{ACP} Criterio de similaridad basado ACP.
t *Scores* obtenidos con un ACP.
T Matriz de tipicalidades de la técnica *FPCM*.
T^2 Estadístico de *Hotelling*.
up Medición para explorar *outliers* en los métodos *OutMi*.
U la matriz de pertenencias de una técnica *CLD*.
v prototipo de *cluster* o centro del *cluster*.
y_t Observación de una variable en el instante *t*.
Y Matriz de datos de proceso de dimensiones *m*x*n*.
XB Índice de Xie-Beni.
$\mathbb{R}$ Conjunto de los números reales.
$\mathbb{Z}$ Conjunto de los números enteros.

LETRAS GRIEGAS
α,β Constantes asociadas al cálculo de los P_i en el método *MSCl*.
γ Constante con valores entre [0,1] para el cálculo de las credibilidades.
$\underline{\varepsilon}$ Umbral para los P_i bajo el cual se rechaza (sin lugar a dudas) un punto como centro de *clusters* en el método *MSCl*.
$\overline{\varepsilon}$ Umbral para los P_i sobre el cual se acepta (sin lugar a dudas) un punto como centro de *clusters* en el método *MSCl*.

δ Índice de difusividad para las técnicas *CLD*.
η Constante definida por el usuario para la función objetivo de la técnica PCM.
κ_i Distancia media entre x_i y los σ objetos más cercanos.
μ_{ik} Pertenencia individual de un objeto (una observación) *i* a un grupo *k*.
$\mu_{j,lim}$ Valor límite o de control para las de pertenencias en la condición de operación *j*.
ξ_k Eficiencia del *cluster* k.
ρ_k Volumen del *cluster* k.
σ Elementos más próximos a y_i, en términos de alguna norma de distancia $||\ ||_C$.
τ Tipicalidades para la técnica *FPCM*.
ψ_i Variable de credibilidad que representa la tipicalidad de $\mathbf{y}_i$ dentro de $\mathbf{Y}$.

SUPERÍNDICES

* Indicativo de que P_i tiene el máximo valor en el método *MSCl*.

SUBÍNDICES

i i-ésima observación, objeto o muestreo.
j j-ésima condición de operación.
k k-ésimo *cluster* o grupo.
t tiempo de muestreo.

ACRÓNIMOS

ACP	Análisis de Componentes Principales.
ART2	Redes basadas en Teoría de Resonancia Adaptativa o *Adaptive Resonance Theory (ART2)*.
CFCM	*Credibilistic Fuzzy c-means*.
CLD	*Clustering* basado en Lógica Difusa.
CP	Componentes Principales.
DM	*Data Mining*.
MCP	Mínimos Cuadrados Parciales o *Partial Least Squares (PLS)*.
FCM	*Fuzzy c-means*. Técnica *CLD*.
FPCM	*Fuzzy Possibilistic c-means*. Técnica *CLD*.
KDD	*Knowledge Discovery in Databases*.
MCP	Mínimos Cuadrados Parciales.
MSCl	Método *Subtractive Clustering*.
PCM	*Possibilistic c-means*. Técnica *CLD*.
SOM	*Self Organizing Maps*.
TEM	Técnicas Estadísticas Multivariables.
TEM-*CLD*	Estrategias de análisis que combinan Técnicas TEM con Técnicas *CLD*.
GK	Etiqueta para designar la variante al cálculo de la distancia en los métodos *CLD*, tal como lo propusieron *Gustafson* y *Kessel*.
OutMi	Etiqueta para los procedimientos de identificación de *outliers*.

CAPÍTULO 6. APLICACIONES DE ESTRATEGIAS BASADAS EN CLUSTERING PARA LA SUPERVISIÓN DE PROCESOS

RESUMEN
En este capítulo, se resumen los resultados obtenidos cuando se aplican los conceptos y metodologías propuestos en el capítulo anterior para el desarrollo de estrategias de supervisión de procesos.
En una primera parte se describe una estrategia de supervisión para procesos multioperacionales que trabajan en régimen continuo. La estrategia toma ventaja del clustering combinado con técnicas estadísticas multivariables para afrontar de manera eficiente tanto la correcta identificación de grupos de datos asociados a distintas condiciones de operación como al manejo de los datos asociados a los periodos de transición entre operaciones. Para ilustrar las estrategias anteriores se presenta el caso de un reactor de polimerización industrial operado a múltiples estados operacionales y en régimen continuo. Se logra ver de manera efectiva el diseño y la exitosa aplicación en línea del sistema de monitorización que resulta de la estrategia aplicada sobre el reactor de polimerización. En una segunda parte, se exploran variantes de las estrategias anteriores para su aplicación en casos con procesos operados en régimen semi-continuo o discontinuo. Las estrategias se utilizan para el análisis de los ciclos de operación de un reactor operado en forma semi-continua que conducen a propuestas de mejora en la operación y para una mejor planificación del mantenimiento. Los resultados muestran el potencial de las estrategias desarrolladas para múltiples aplicaciones.

6.1 Supervisión de procesos multioperacionales continuos

6.1.1 Revisión Preliminar

En capítulos precedentes ya se resaltó el auge creciente de las estrategias basadas en Técnicas Estadísticas Multivariables (TEM), y en especial, en el Análisis de Componentes Principales (ACP). Una nota característica de las estrategias TEM propuestas en la literatura es que las mismas se orientan a la supervisión de procesos de producción de productos individuales (Martin *et al*., 2002). Por otro lado, muchas industrias trabajan bajo esquemas de producción de una amplia variedad de productos, algunos fabricados en pequeñas cantidades, lo que conduce a cambios continuos en las condiciones del proceso o procesos multioperacionales. En estos casos las alternativas de supervisión serían:

- Modelos locales por cada producto o grado de producto.
- Modelos globales.

6.1.1.1 Modelos locales por cada producto o grado de producto

En este caso por cada producto, o grado de producto, se desarrollaría un sistema de monitorización según se explica en la sección 4.1 y basado en el ACP clásico o en alguna variante. Si el número de productos o grados de productos elaborados es muy alto, esto requeriría un número igual de modelos y sus correspondientes gráficos de control para el *SPE* (*Squared Predictive Error*), T^2, *scores*, etc. No obstante, el número de datos disponibles para algunos de estos productos o grados de productos podría ser muy bajo por la poca demanda de los mismos con sus correspondientes muy infrecuentes producciones. Para tales casos los

modelos obtenidos serían estadísticamente poco fiables dada la poca cantidad de datos utilizados. Adicionalmente, si los cambios de producto son muy frecuentes se tendría que estar continuamente cambiando los gráficos a usar en cada equipo o línea de producción en que se esté trabajando. Aún cuando esto podría automatizarse con ayuda de un ordenador, bajo este escenario de cambios tan frecuentes y con múltiples gráficos siempre existirá el riesgo de errores por parte del operador que supervisa ante tanta información cambiante, con los consecuentes efectos en los niveles de producción y en las calidades de los productos obtenidos.

6.1.1.2 Modelos globales

En estos casos se propone crear un único modelo a partir de los datos de todas las producciones pasadas. Las propuestas orientadas a ello se pueden clasificar en:

Estrategias que combinan técnicas TEM con Clustering

Estas estrategias ya se estudiaron en el capítulo 5. Combinan una técnica TEM con *Clustering* basado en Lógica Difusa (*CLD*) o TEM-*CLD*. En trabajos precedentes (ver sección 5.1) se ha puesto de manifiesto el potencial de las TEM-*CLD* para asistir en el análisis de procesos tras largos periodos de operación o para desarrollar sistemas de monitorización para procesos multioperacionales. Sin embargo, al usarlos pueden presentar problemas de detección de anormalidades como el que se ilustra a continuación: Supóngase una planta donde se fabrican 3 productos A, B y C. Los cambios de producción son muy frecuentes. Se toman los datos históricos **Y** que incluyen operaciones normales recientes de todos los productos y a intervalos de muestreo de 10 minutos. Tras procesar la matriz **Y** con ACP se obtiene un modelo ACP Global Tradicional (ACPGT) de los mismos con sus correspondientes matrices de *scores* **t** y *loadings* **P** (ver sección 4.1). Luego, al tratar los *scores* **t** con una técnica *CLD* se obtiene la identificación de grupos que se muestra en el gráfico de *scores* del ACPGT (figura 6.1a).

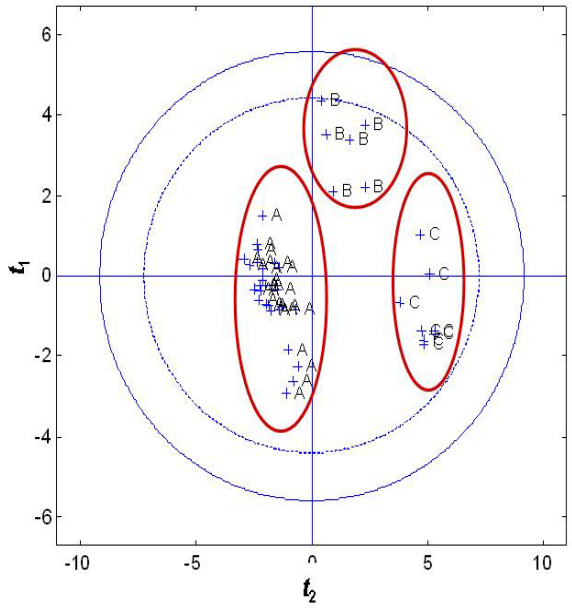

Figura 6.1a. *Scores* del ACPGT para **Y**.

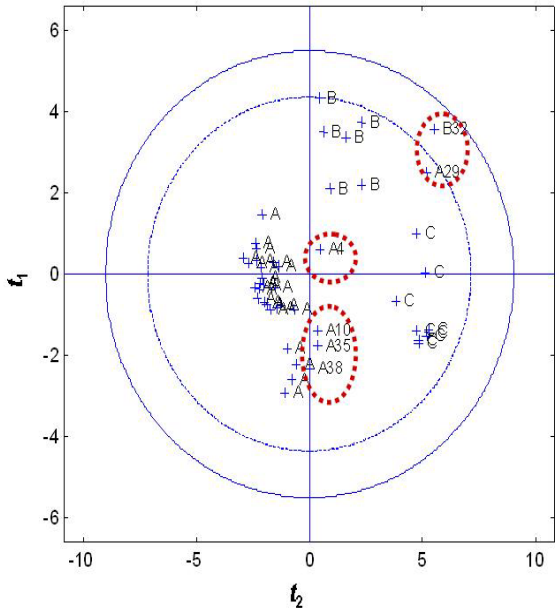

Figura 6.1b. *Scores* del ACPGT para $\mathbf{Y}_{nv}$.

En dicho gráfico, un punto representa un intervalo de operación de 10 minutos. Se observa que los intervalos en que se produce cada producto quedan claramente diferenciados formando grupos o regiones de operaciones normales para cada uno (A, B y C). Los círculos

concéntricos son los límites de confianza de los *scores* al 95 % y al 99% y dentro de los que se captan las variaciones de causa conocida durante la operación normal del proceso.

A continuación, se recoge un nuevo conjunto de datos de operaciones posteriores $\mathbf{Y}_{nv}$ que incluye seis intervalos de operación anormal. Al ir proyectando los datos sobre el gráfico de los 2 primeros *scores* de **Y** e identificarlos con el *CLD* (figura 6.1b), se observa que algunos intervalos de producción tanto para A como para B se alejan de las nubes de puntos de operación normal (ver puntos A4, A10, A29, A35, A38 y B32). No obstante, dichos puntos permanecen dentro de los límites de confianza del modelo global obtenido. Así, esto erróneamente indica que el proceso se mantiene operando sin problemas. Si se mira el resultado en los gráficos de control con *SPE* y T^2 (figura 6.2) se observa que la detección de periodos con operación anormal también es nula.

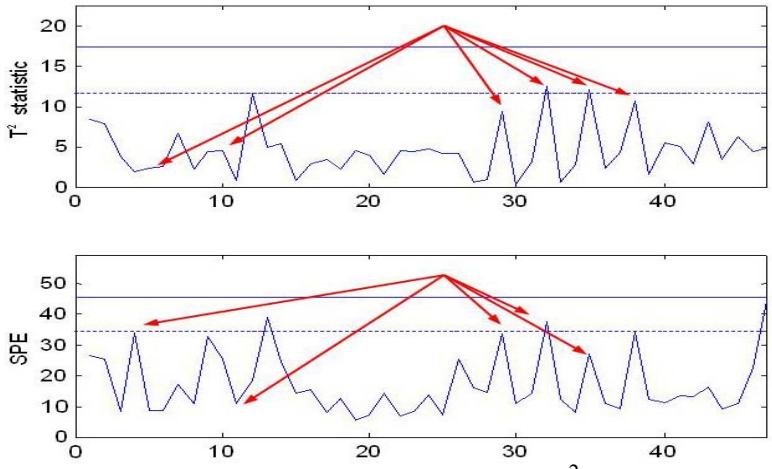

Figura 6.2. Detección con el *SPE* y T^2 del ACPGT.

Aun cuando a través del gráfico de *scores* (figura 6.1b) visualmente se podría rastrear una desviación en la operación, esto se iría complicando a medida que se trabaje con más productos y las zonas de separación entre regiones sea cada vez más pequeña.

Modelos ACP Multigrupos

En otros trabajos recientes (Hwang y Han, 1999; Martin *et al*., 2002) se proponen variantes a los modelos ACPGT para procesos multioperacionales. Un primer trabajo propone una estrategia que se puede resumir como sigue (Hwang y Han, 1999):

- Se toma la matriz de datos de operación **Y** y se divide por regiones de operación obteniéndose una matriz $\mathbf{Y}_k$ por cada grupo, donde $k = 1,\ldots, c$ y c es el nº de grupos.
- Se estandariza cada $\mathbf{Y}_k$ por separado y a partir de las medias y varianzas de cada una de sus columnas. Como resultado se obtienen las correspondientes $\mathbf{Y}_k^e$.
- Se crea una matriz $\mathbf{Y}_{cn}$ a partir de la unión de todas las $\mathbf{Y}_k^e$.
- A partir de $\mathbf{Y}_{cn}$ se obtiene un modelo único llamado ACP Multigrupo o ACPMg que integra las diferentes condiciones de operación del proceso en una misma región.

¿Cuál es la ventaja de esta nueva representación? Para responder a ello, se toma de nuevo la matriz de datos **Y** utilizada para crear las figuras 6.1a a 6.2. Se obtiene el correspondiente modelo ACPMg y se construye el gráfico de *scores* (figura 6.3a). A continuación se toma la matriz $\mathbf{Y}_{nv}$ (datos anormales en las figuras 6.1 a 6.2) y con ayuda del modelo ACPMg se

obtiene la proyección en el gráfico de los *scores* (figura 6.3b) y en los gráficos del *SPE* y T^2 (figura 6.4).

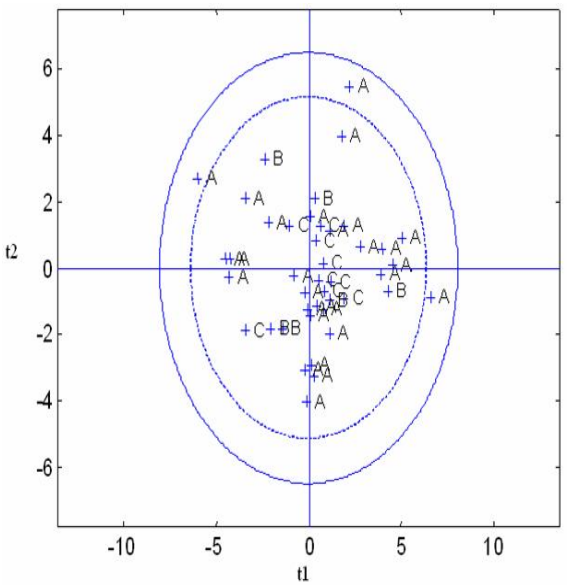

Figura 6.3a. *Scores* del ACPMg para **Y**.

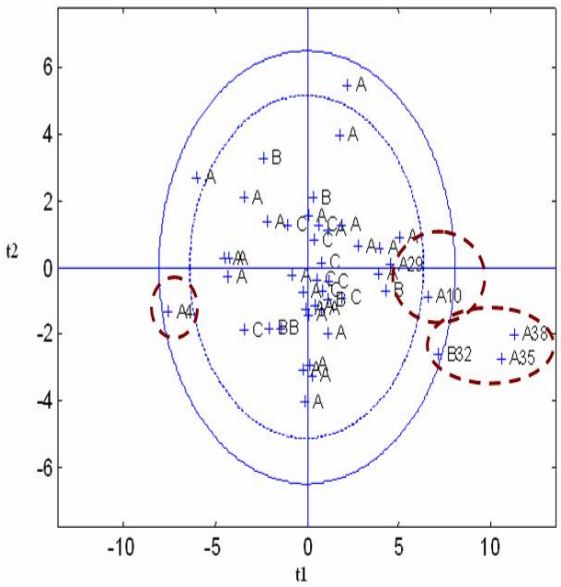

Figura 6.3b. *Scores* del ACPMg para $\mathbf{Y}_{nv}$.

A través del gráfico de *scores* (figura 6.3a) se puede ver que muchos de los puntos correspondientes a los intervalos de operaciones anormales se ven claramente desviados de la operación normal cosa que no sucede en el gráfico de *scores* obtenido con el ACPGT (figura 6.1a). Aún más, si se exploran los gráficos *SPE* y T^2 correspondientes al ACPMg las desviaciones quedan identificadas de forma muy clara (ver figura 6.4).

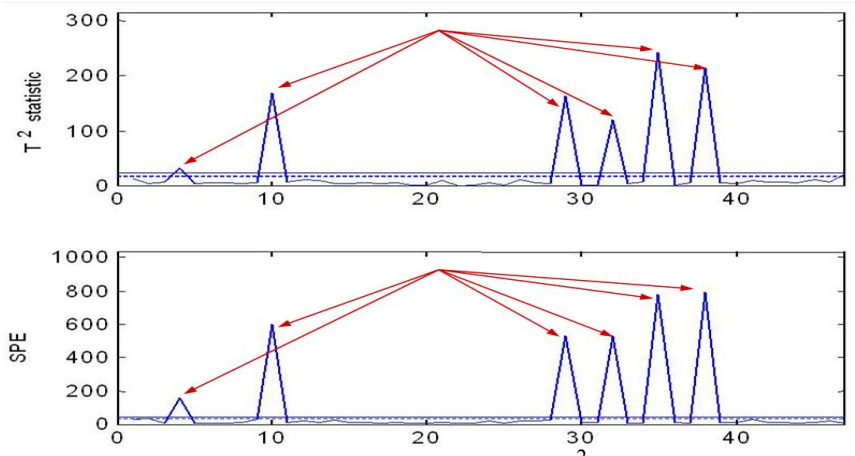

Figura 6.4. Detección con el *SPE* y T^2 del ACPMg.

En otro trabajo (Martin *et al*., 2002) se propone una estrategia ACPMg similar a la anterior (Hwang y Han, 1999). La diferencia en esta nueva propuesta ACPMg (Martin *et al*., 2002) se muestra a continuación:

- Se toman los datos de operación **Y** y se dividen por grupos o regiones de operación obteniéndose una matriz $\mathbf{Y}_k$ por cada grupo, donde k=1,…, c y c es el número de grupos.
- Para cada $\mathbf{Y}_k$ se calcula su correspondiente matriz covarianza $\mathbf{Q}_k$.
- A partir de las $\mathbf{Q}_k$ se calcula una matriz covarianza única denotada como $\mathbf{Q}_{cn}$.

- Se utiliza la matriz $\mathbf{Q}_{cn}$ dentro del procedimiento del ACP. Se obtiene el modelo ACPMg.

6.1.1.3 Observación general

La principal ventaja de las 2 estrategias basadas en modelos ACPMg (ver sección 6.1.1.2) radica en el hecho de disminuir significativamente el número de modelos a usar lo que puede ser muy práctico para los operadores en situaciones reales de procesos de fabricación flexible.

Por otro lado, dado que las estrategias están orientadas al caso de procesos multioperacionales en muchos casos se estarán aplicando durante la transición de operaciones. En estos casos, el sistema conducirá a detecciones de fallos cuando en realidad solo se esta produciendo la transición de una operación a otra (ver ejemplo de la sección 6.1.3). Por lo tanto, debería asegurarse que tanto el sistema de monitorización como el operador puedan discriminar entre una transición y un fallo. Adicionalmente, en las 2 estrategias descritas, se asume que los datos que se recogen del proceso ya se han procesado previamente de manera que los datos de transiciones ya han sido eliminados y, por tanto, se pueden utilizar directamente para construir el modelo. No obstante, en muchas situaciones reales no siempre se dispondrá ni será fácil hacer este tipo de tratamientos previos sobre los que la literatura no indica nada. Así, sería muy ventajoso que durante el diseño de sistemas de monitorización basados en modelos ACPMg, se pueda disponer de herramientas que asistan en el tratamiento de datos asociados a transiciones del proceso. También, se debe resaltar que en las propuestas de estrategias de supervisión tanto para procesos multioperacionales como procesos de fabricación de un solo producto (o una sola región de operación de trabajo) se suele asumir que al diseñar el sistema de supervisión los datos vienen libres de *outliers* (Hwang y Han, 1999; Martin *et al*., 2002). No obstante, sería muy útil que durante el proceso de diseño se contara con herramientas necesarias que aseguren la eliminación de datos atípicos en caso de no haberse podido eliminar previamente ya que los mismos pueden afectar negativamente el modelo resultante.

En las estrategias que se discuten en las secciones que siguen se toma ventaja de los resultados de capítulos precedentes para proponer estrategias efectivas de supervisión de procesos multioperacionales que den salidas alternativas a los inconvenientes mencionados anteriormente.

6.1.2 Propuesta de estrategias de supervisión para procesos multi-operacionales continuos

En esta sección se presenta un conjunto de desarrollos para asistir en la supervisión de procesos multioperacionales. Las estrategias que se proponen toman ventaja tanto de las estrategias TEM-*CLD* (discutidas en el capítulo 5) como del uso de los modelos ACPMg. Adicionalmente, el diseño del sistema de monitorización resultante se propone de forma sistemática lo que proporciona una guía fácil para posibles implementaciones reales. Este es un aspecto frecuentemente infravalorado en la literatura de análisis y monitorización de procesos.

6.1.2.1 Estrategia de diseño Mc

Se comienza por proponer una primera estrategia de diseño de sistemas de monitorización de procesos multioperacionales, etiquetada como Mc. En esta primera estrategia se asume que los datos históricos han sido previamente revisados y separados de forma que se pueden

recuperar los conjuntos que involucran solo operaciones normales del pasado sin problema. Esta es la suposición sobre la que basan todos los trabajos existentes en la literatura sobre estrategias TEM-*CLD*. La estrategia se describe a continuación:

Mc1 Se reciben los datos de proceso y se asignan a una matriz **Y**. Estos se estandarizan según la media y la desviación estándar de cada variable en **Y** como sigue:

$$\mathbf{Y}^{e} = \frac{(\mathbf{Y} - \bar{\mathbf{Y}})}{\mathbf{De}} \tag{6.1}$$

Donde $\mathbf{Y}^{e}$ es la matriz de datos estandarizada, $\bar{\mathbf{Y}}$ es el vector de las medias de cada variable en **Y** y **De** es la matriz de desviaciones estándar de **Y**.

Mc2 Se aplica ACP sobre $\mathbf{Y}^{e}$. Así, se obtienen las matrices **t** (*scores*) y **P** (*loadings*) y los vectores de valores para *SPE* y T^2.

Mc3 Se aplica el análisis *clustering* sobre **t**. Se obtiene la matriz **U** de pertenencias con sus μ_{ik} (pertenencias de cada punto dentro de la nube de puntos formada por los *scores* **t**), las coordenadas de los centros de *clusters* **v** ($\mathbf{v}_1,\ldots, \mathbf{v}_c$) y el vector **ds** con las distancias de cada muestra en $\mathbf{Y}^{e}$ a cada $\mathbf{v}_k$.

Mc4 Se analizan *outliers* mediante *SPE*, T^2 y **ds** (ver la sección 5.5.3).

Mc5 Se eliminan los *outliers* de los datos lo que resulta en la matriz de datos originales modificada como $\mathbf{Y}^{o}$. Con esta matriz se recalculan los modelos iniciales del ACP y del *clustering* obteniéndose $\bar{\mathbf{Y}}^{o}$, $\mathbf{De}^{o}$, $\mathbf{t}^{o}$, $\mathbf{P}^{o}$, $\mathbf{U}^{o}$, $\boldsymbol{\mu}^{o}$, $\mathbf{ds}^{o}$, $\mathbf{v}^{o}$, etc. Esta información se guarda para la monitorización en línea.

Mc6 Se divide $\mathbf{Y}^{o}$ en $[\mathbf{Y}_1^{o}, \mathbf{Y}_2^{o}, \ldots, \mathbf{Y}_c^{o}]$ usando para ello la información en $\mathbf{U}^{o}$.

Mc7 Se estandarizan $(\mathbf{Y}_1^{o}, \mathbf{Y}_2^{o}, \ldots, \mathbf{Y}_c^{o})$ separadamente. Esto conduce a los conjuntos $[\mathbf{Y}_1^{oe}, \mathbf{Y}_2^{oe}, \ldots, \mathbf{Y}_c^{oe}]$, $[\bar{\mathbf{Y}}_1^{oe}, \bar{\mathbf{Y}}_2^{oe}, \ldots, \bar{\mathbf{Y}}_c^{oe}]$ y $[\mathbf{De}_1^{oe}, \mathbf{De}_2^{oe}, \ldots, \mathbf{De}_c^{oe}]$.

Mc8 Se crea la matriz conjunta:

$$\mathbf{Y}_{cn}^{o} = \left[\mathbf{Y}_1^{oe}, \mathbf{Y}_2^{oe}, \ldots, \mathbf{Y}_c^{oe}\right] \tag{6.2}$$

Mc9 Se aplica ACP sobre $\mathbf{Y}_{cn}^{o}$. Así, se obtiene el modelo ACPMg para $\mathbf{Y}_{cn}^{o}$ con sus correspondientes $\mathbf{t}_{cn}$, $\mathbf{P}_{cn}$, SPE_{cn} y T_{cn}^2.

Mc10 Se fijan los límites para monitorizar en línea con $\boldsymbol{\mu}^{o}$, SPE_{cn} y T_{cn}^2.

La estrategia anterior se muestra de forma esquemática en la figura 6.5. En esencia, la estrategia es igual a las estrategias existentes en la literatura para modelos ACPMg con la diferencia de añadir un paso de detección y tratamiento de *outliers* de manera de asegurar un sistema no afectado por este tipo de datos anormales. En el paso Mc3 (bloque "Identificación de *clusters*") se cuenta con el número de productos ú operaciones pasadas como valor de *c* (número de *clusters*) para el algoritmo de *clustering* a utilizar. Si este valor no se conociese con precisión, se podría estimar según se plantea en las secciones 5.3 y 5.5.4. En cuanto a la técnica de *clustering* para identificar los puntos que pertenecen a cada grupo se tienen 2 opciones: La *Fuzzy c-means* con la variante de *Gustaffson-Kessel* o *FCM-GK* y la *Fuzzy Possibilistic c-means* también con la variante *Gustaffson-Kessel* o *FPCM-GK* (Ver conclusiones del capítulo 5). En pruebas hechas durante el desarrollo de este trabajo se vio que la selección de una ú otra no afecta significativamente en el resultado de las mismas. En lo que sigue siempre se utilizará la técnica *FCM-GK*.

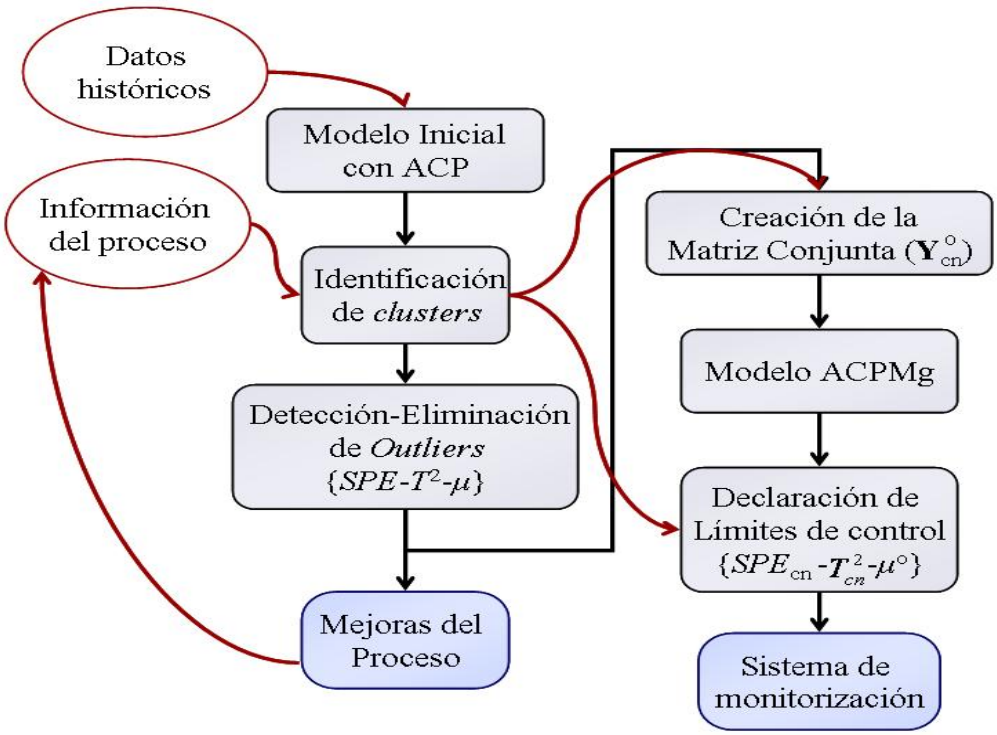

Figura 6.5. Estrategia de diseño Mc, para procesos continuos multioperacionales.

Los pasos Mc4 y Mc5 de la estrategia se agrupan en un solo bloque "Detección-Eliminación de *Outliers*". El tratamiento de *outliers* ya se vio que era muy importante en el capítulo anterior. Al integrarlo dentro de la estrategia Mc y según procedimientos bastantes simples y efectivos (ver sección 5.5.3) se asegura la eficiencia de la estrategia propuesta en presencia de datos anormales, aspecto este que en la literatura existente siempre se evade.

Tras el bloque "detección-eliminación de *outliers*" (pasos Mc4 y Mc5) se coloca el bloque "Mejoras del Proceso". Esto se hace para indicar que una vez aplicada la estrategia hasta este punto, los resultados que se tienen pueden servir para analizar el agrupamiento de operaciones pasadas, el número de instantes de operación anormales que han ocurrido, la contribución de cada variable a las diferentes operaciones normales y anormales, etc., todo lo cual puede servir para introducir distintas mejoras en el proceso como actualización de las condiciones de operación para cada tipo de producto o asignar un mayor control a ciertas variables influyentes en la operación de algunos productos.

Los pasos Mc6 y Mc7 son prerrequisitos para el paso Mc8, es por eso que en la figura 6.5 aparecen agrupados bajo un solo bloque llamado "Creación de la Matriz Conjunta". Luego, la matriz conjunta $\mathbf{Y}_{cn}^{o}$ se utiliza como entrada para obtener el modelo ACPMg. Para obtener dicho modelo, se ha adoptado la estrategia descrita al inicio de la sección 06.1.1.2 (Hwang y Han, 1999) evitando así el problema de la restricción de datos que conlleva el método propuesto en el trabajo de Martin et al., (2002). En el último paso de la metodología (paso Mc10) se propone el cálculo de los límites de control para las mediciones utilizadas con tal fin. Los límites de *SPE* y T^2 se calculan a partir de la información del ACPMg obtenido y según se discute en la sección 4.3.1. En el caso de los límites para los $\boldsymbol{\mu}^{o}$, estos se calculan mediante la prueba *t* de *Student* (Yoo *et al*., 2003). Con esto queda completada la metodología de diseño Mc.

6.1.2.2 Estrategia de diseño Mt

En esta segunda propuesta los datos pueden no solo contener datos de operaciones normales sino también datos de transiciones entre operaciones (cambio de productos), por lo que la estrategia debería ser capaz de poder extraer los datos de operación normal y obtener el sistema de monitorización a partir de estos. Así se propone el siguiente procedimiento para la estrategia Mt:

- Mt1 Se reciben los datos de proceso y se asignan a una matriz **Y**. Se estandarizan los datos obteniéndose $\mathbf{Y}^{e}$.
- Mt2 Se aplica ACP sobre $\mathbf{Y}^{e}$. Así, se obtiene la matriz de **t** (*scores*) y **P** (*loadings*) y los vectores de valores para *SPE* y T^2 (ver sección 4.1).
- Mt3 Se aplica el análisis *clustering* sobre **t**. Se obtiene la matriz **U** con todas las pertenencias μ_{ik}, las coordenadas de los centros de *clusters* **v** y el vector **ds** con las distancias de cada muestra en $\mathbf{Y}^{e}$ a cada $\mathbf{v}_k$.
- Mt4 Se analizan *outliers* mediante *SPE*, T^2 y **ds** (ver la sección 5.5.3).
- Mt5 Se eliminan los *outliers* de los datos lo que resulta en las matrices $\mathbf{Y}^{o}$, $\mathbf{t}^{o}$ y $\mathbf{U}^{o}$.
- Mt6 Se analizan los datos de la matriz $\mathbf{U}^{o}$ para identificar las transiciones (ver sección 6.1.2.2.1). De esto resulta la matriz $\mathbf{Y}_o^{oo}$ libre de datos de transiciones y de outliers. A partir de esta matriz se recalculan los modelos iniciales de ACP y de *clustering* obteniéndose $\bar{\mathbf{Y}}^{oo}$, $\mathbf{De}^{oo}$, $\mathbf{t}^{oo}$, $\mathbf{P}^{oo}$, $\mathbf{U}^{oo}$, $\boldsymbol{\mu}^{oo}$, $\mathbf{ds}^{oo}$, $\mathbf{v}^{oo}$,…, etc. Esta información se guarda para la monitorización en línea.
- Mt7 Se divide $\mathbf{Y}_o^{oo}$ en $[\mathbf{Y}_1^{oo}, \mathbf{Y}_2^{oo}, \ldots, \mathbf{Y}_c^{oo}]$ usando para ello la información de los grupos en $\mathbf{U}^{oo}$.
- Mt8 Se estandarizan $[\mathbf{Y}_1^{oo}, \mathbf{Y}_2^{oo}, \ldots, \mathbf{Y}_c^{oo}]$ por separado. Esto conduce a los conjuntos $[\mathbf{Y}_1^{ooe}, \mathbf{Y}_2^{ooe}, \ldots, \mathbf{Y}_c^{ooe}]$, $[\bar{\mathbf{Y}}_1^{ooe}, \bar{\mathbf{Y}}_2^{ooe}, \ldots, \bar{\mathbf{Y}}_c^{ooe}]$ y $[\mathbf{De}_1^{ooe}, \mathbf{De}_2^{ooe}, \ldots, \mathbf{De}_c^{ooe}]$.
- Mt9 Se crea la matriz conjunta según se expresa en la ecuación 6.2, obteniendose $\mathbf{Y}_{cn}^{oo}$.
- Mt10 Se aplica ACP sobre $\mathbf{Y}_{cn}^{oo}$. Así, se obtiene el modelo global para $\mathbf{Y}_{cn}^{oo}$ con sus correspondientes $\mathbf{t}_{cn}$, $\mathbf{P}_{cn}$, SPE_{cn} y T_{cn}^2.
- Mt11 Se fijan los límites para monitorizar en línea con $\boldsymbol{\mu}^{oo}$, SPE_{cn} y T_{cn}^2.

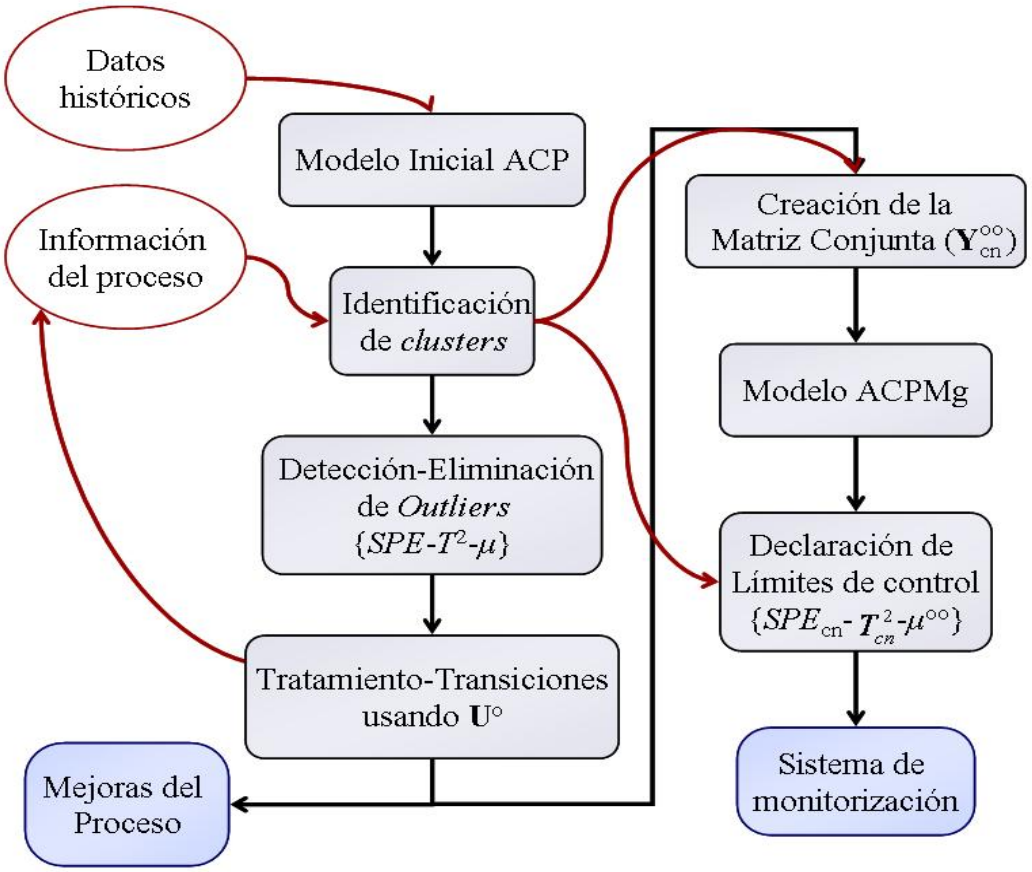

Figura 6.6. Estrategia de diseño Mt, para procesos continuos multioperacionales.

6.1.2.2.1 Tratamiento de las transiciones

La diferencia entre Mc y Mt radica en un paso adicional (paso Mt6) que aparece en Mt y que sirve para el tratamiento de las transiciones en los datos. A continuación se describe el procedimiento que se aplica en este paso:

Mt6a Se toma la matriz $\mathbf{U}^o$ y de ella se extraen *c* vectores de pertenencia $\boldsymbol{\mu}_k^o$. Cada uno de estos vectores contiene el grado de pertenencia de cada observación *i* en **Y** a cada grupo o región de operación *k*. Luego, de cada $\boldsymbol{\mu}_k^o$ se toma cada una de las μ_{ik}^o que indica a que grupo k pertenece la observación i. Con las μ_{ik}^o seleccionadas se crean los vectores $\boldsymbol{\mu}_k^{o+}$ que solo contienen las pertenencias de las *i* observaciones que forman parte de cada grupo *k*.

Mt6b Se toma cada vector $\boldsymbol{\mu}_k^{o+}$ y se filtra mediante la estrategia *levashrink* propuesta en el capítulo 2 y usando como función la *wavelet Daubechies db1*. Tras esto se tienen los vectores $\boldsymbol{\mu}_k^{of}$.

Mt6c Se toma cada vector $\boldsymbol{\mu}_k^{of}$ y se obtiene una aproximación suavizada del mismo. La aproximación suavizada se calcula según la ecuación 2.14 (ver sección 2.1.1.2) utilizando una *wavelet Daubechies* suavizada (*db4* ó *db8*) y un valor para el nivel de descomposición *L*=3. De esto resultan los vectores $\boldsymbol{\mu}_k^{off}$.

Mt6d Se calcula la media de cada vector $\boldsymbol{\mu}_k^{off}$. Luego, basándose en la prueba estadística *t* de *Student* se calculan los intervalos de confianza de la media IC_k para cada vector, a un nivel de confianza α, como sigue:

$$IC_k^s = \bar{\mu}_k^{off} + t_{1-\alpha/2,n-1}\sigma_k^{off} \quad (6.3)$$

$$IC_k^i = \bar{\mu}_k^{off} - t_{1-\alpha/2,n-1}\sigma_k^{off} \quad (6.4)$$

Donde IC_k^s e IC_k^i representan los límites superior e inferior del intervalo de confianza respectivamente, $\bar{\mu}_k^{off}$ es la media del vector $\boldsymbol{\mu}_k^{off}$, y σ_k^{off} es la varianza muestral de $\boldsymbol{\mu}_k^{off}$. El límite IC_k^s se toma como valor umbral para determinar las transiciones.

Mt6e Se comparan los valores de cada observación de cada $\boldsymbol{\mu}_k^{off}$ con el correspondiente IC_k^i. Todos los valores que estén por debajo del intervalo se consideran como transiciones. Luego, se eliminan las correspondientes observaciones de la matriz $\mathbf{Y}^o$, lo que conduce a la nueva matriz $\mathbf{Y}^{oo}$.

Mt6f

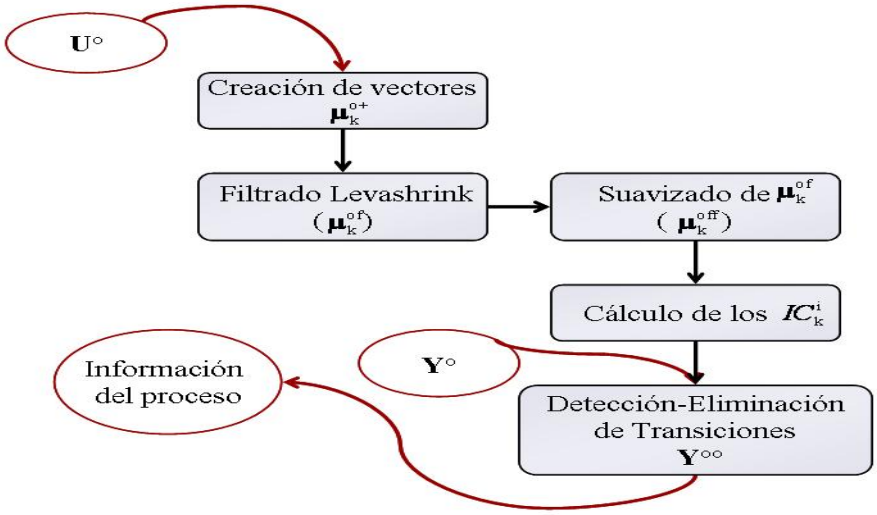

Figura 6.7. Procedimiento para el tratamiento de las transiciones.

En la figura 6.7 se muestra un esquema del procedimiento utilizado para el tratamiento de las transiciones. Para entender dicho procedimiento se propone un ejemplo sencillo que consiste en una planta hipotética de la cual se miden continuamente 5 variables y que en un instante de tiempo dado sufre un cambio en sus condiciones de operación debido a un cambio de producto. En la figura 6.8 se muestran datos medidos del proceso, incluyendo el punto donde sucede el cambio de operación.

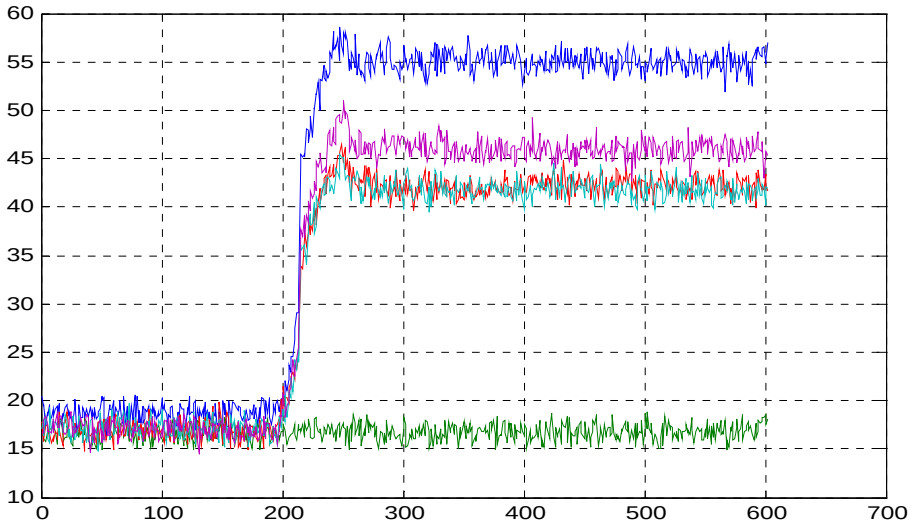

Figura 6.8. Variables del proceso de ejemplo.

Los datos se procesan mediante la metodología Mt. Al llegar al paso Mt6 se utiliza el procedimiento para el tratamiento de las transiciones. Los vectores contenidos en $\mathbf{U}^{o}$ ($\boldsymbol{\mu}_k^{o+}$) se muestran en el gráfico de la esquina superior izquierda de la figura 6.9. Al redefinirlos según el número de observaciones que pertenece a cada *cluster* se tienen los $\boldsymbol{\mu}_k^{o+}$ que se muestran en el gráfico de la esquina inferior derecha de la figura 6.9. Los valores que caen abruptamente en ambos vectores corresponden al intervalo donde se produce la transición de operaciones.

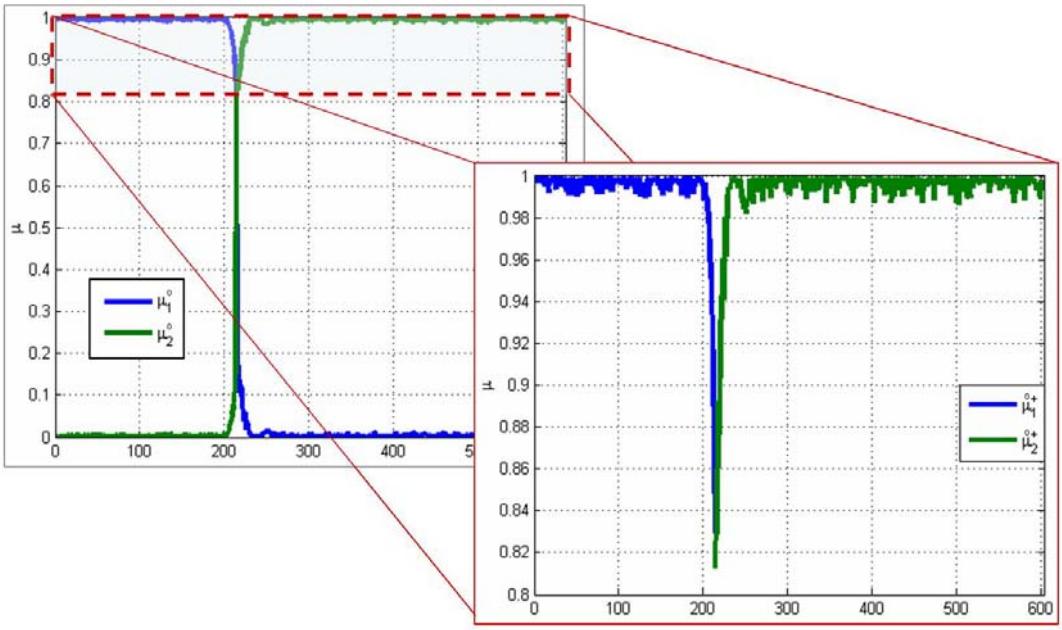

Figura 6.9. Gráfico de vectores de pertenencia $\boldsymbol{\mu}_k^{o}$ y $\boldsymbol{\mu}_k^{o+}$.

Como se puede observar en el gráfico de la esquina inferior derecha (figura 6.9), los vectores de pertenencia no son suaves sino que sufren de una oscilación aleatoria a lo largo de sus valores. Con los pasos Mt6b y Mt6c se obtienen versiones suavizadas de estos vectores de pertenencias. La ventaja de este suavizado queda más clara tras la aplicación de los pasos Mt6d y Mt6e. Primero se aplica Mt6d y se obtienen los correspondientes IC_k^i (ver figura 6.10, donde el gráfico inferior es una ampliación del gráfico superior). Para visualizar el efecto que

tiene el suavizado se repite el paso Mt6d pero esta vez se utilizan los vectores $\boldsymbol{\mu}_k^{o+}$ (y su correspondiente media y varianza) en lugar de los vectores suavizados $\boldsymbol{\mu}_k^{off}$, obteniéndose con ello los correspondientes IC_k^i (ver figura 6.11, donde el gráfico inferior es una ampliación del gráfico superior).

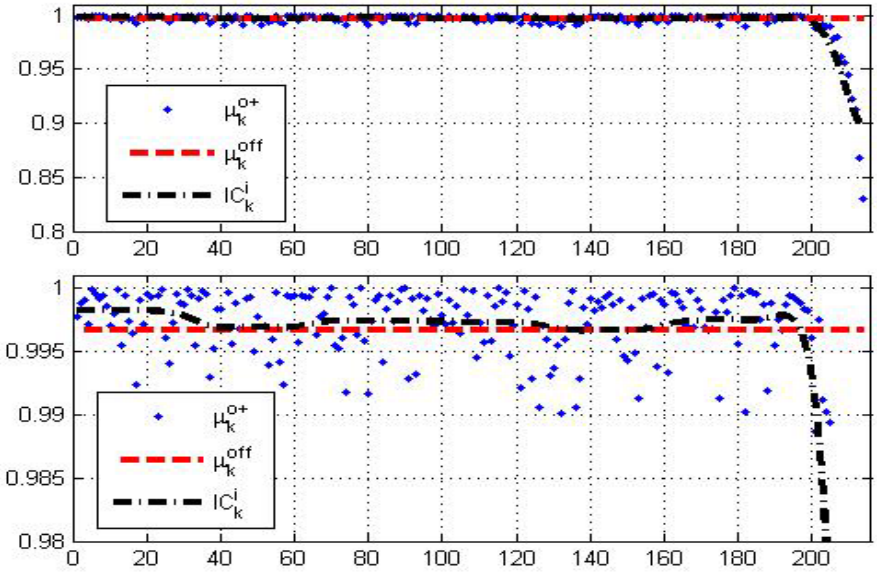

Figura 6.10. Límite de control para tratamiento de Transiciones basado en $\boldsymbol{\mu}_1^{off}$.

Tomando los resultados mostrados en la figura 6.10, y para el caso de las $\boldsymbol{\mu}_{k=1}^{o+}$ o $\boldsymbol{\mu}_1^{o+}$ (los resultados para el caso de $\boldsymbol{\mu}_{k=2}^{o+}$ son similares), se puede ver que el vector $\boldsymbol{\mu}_k^{off}$ siempre se mantiene por encima de IC_k^i, excepto en el final que es donde ocurre la transición. Así, si se aplica Mt6e sobre $\boldsymbol{\mu}_k^{off}$ se eliminaran solo las correspondientes observaciones asociadas a la transición.

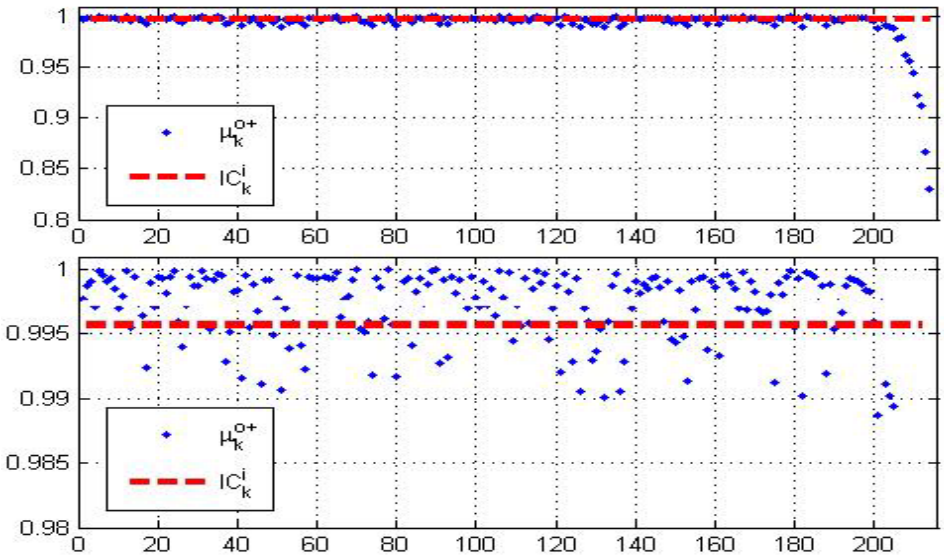

Figura 6.11. Límite de control para tratamiento de Transiciones basado en $\boldsymbol{\mu}_1^{o+}$.

Si se hace el mismo análisis sobre la figura 6.11, se puede ver rápidamente que al aplicar el paso Mt6e se eliminarían muchísimos datos que corresponderían a la operación normal. De ahí que es más ventajoso aplicar el procedimiento tal como se propone.

6.1.2.3 Monitorización

Tras aplicar cualquiera de las estrategias de diseño anteriores, se tiene un modelo ACP multigrupo que aporta las mismas reglas de control basadas en *SPE* y T^2 para todas las regiones de operación. Aparte, se calculan los límites para los μ_{ik} de cada *cluster*. En la mayoría de los trabajos precedentes, solo se utilizan los μ_{ik} y sus correspondientes límites para monitorizar tal que por cada *cluster* generan un gráfico de control (Choi *et al*., 2003). En este capítulo se propone un gráfico alternativo que consiste en graficar todos los valores de pertenencia de cada grupo en un mismo gráfico de barra diferenciado por colores. De esta forma, el operador en cada instante puede ver la operación dominante y durante la ocurrencia de una transición de operaciones puede llegar a visualizar fácilmente el momento en que el sistema ha alcanzado el nuevo estado deseado. Asimismo, el uso de este gráfico con todas las pertenencias reduce el número de gráficos de pertenencia individuales que, dependiendo del proceso considerado (si se fabrican muchos productos), podría ser muy alto. En paralelo, se propone el uso de los gráficos *SPE* y T^2 del ACP global de tal forma que con el gráfico conjunto de las pertenencias se controla el estado actual o su cambio y con los gráficos (*SPE* y T^2) se controla la aparición de posibles fallos. El algoritmo para la monitorización sería como sigue:

1. Se toman las nuevas mediciones del tiempo actual $\mathbf{Y}(t)$, donde t es el tiempo de muestreo. Se estandarizan según la ecuación 6.1 y utilizando como medias y varianzas a los $\bar{\mathbf{Y}}^{o}$ y $\mathbf{De}^{o}$ si se trabaja con Mc o a los $\bar{\mathbf{Y}}^{oo}$ y $\mathbf{De}^{oo}$ si se trabaja con Mt. Se obtiene $\mathbf{Y}^{e}(t)$.
2. Se calculan los *scores* del tiempo actual como sigue:
$$\mathbf{t}(t) = \mathbf{P}^{T}(t) \cdot \mathbf{Y}^{e}(t) \tag{6.5}$$
3. Con la ayuda de los $\mathbf{t}(t)$ y los $\mathbf{v}^{o}$ (o los $\mathbf{v}^{oo}$) se calculan las pertenencias del punto actual a cada *cluster* de operación (μ_{ik}) como sigue (ver también en sección 5.2) :

$$d_{ik}^2 = \left(\mathbf{t}_i - \mathbf{v}_k\right)^T \mathbf{A}_k \left(\mathbf{t}_i - \mathbf{v}_k\right), \quad 1 \le k \le c, \quad 1 \le i \le n, \tag{6.6}$$

$$\mu_{ik} = \frac{1}{\sum_{l=1}^{c} \left(d_{ik}^2 / d_{lk}^2\right)^{\frac{1}{(\delta-1)}}} \tag{6.7}$$

4. Se evalúa la regla de control para las μ_{ik} y se determina a que operación corresponde el punto actual.
5. Se estandariza nuevamente a $\mathbf{Y}(t)$, pero esta vez se utiliza como medias y desviaciones estándar a los $\bar{\mathbf{Y}}_{k}^{o}$ y $\mathbf{De}_{k}^{o}$ (o $\bar{\mathbf{Y}}_{k}^{o}$ y $\mathbf{De}_{k}^{o}$) que se correspondan con la operación a la que pertenece el punto actual. Así se obtiene $\mathbf{Y}_{k}^{e}(t)$.
6. Con la ayuda de los $\mathbf{t}_{cn}$, $\mathbf{P}_{cn}$ obtenidos durante la etapa de diseño se calculan los correspondientes $SPE_{cn}(t)$ y $T_{cn}^2(t)$ del instante t actual. Luego, se evalúa si estos $SPE_{cn}(t)$ y $T_{cn}^2(t)$ superan los correspondientes límites. En caso de que superen los límites se analiza el gráfico de las pertenencias para ver si corresponde a una transición. Si no corresponde a una transición, la observación corresponde a un fallo.

6.1.3 Caso de estudio: Monitorización de un reactor de polimerización industrial

Este caso ya fue utilizado en un capítulo previo (ver sección 4.3.2.2) y consiste en la simulación de un reactor co ntinuo industrial para la fabricación de Acetato de Polivinilo. Este ejemplo es particularmente atractivo ya que las dinámicas del modelo son muy complicadas y al hacer el cambio de operaciones el sistema tarda tiempo en alcanzar su nuevo estado estacionario. A continuación se propone el siguiente escenario de operación:

Se fabrica un producto *P*, a diferentes grados de calidad y para satisfacer distintas demandas de clientes. Los diferentes grados de producto se consiguen con solo ajustar la temperatura del reactor como se indica en la tabla 6.1. Se crean varios sub-escenarios, identificados como EOMi, y se describen a continuación:

Tabla 6.1. Diferentes Grados del producto A.

Producto	Temperatura
Pa	330 ºC
Pb	310 ºC
Pc	350 ºC

- EOM1: Se trabaja durante 5708 minutos. Cada minuto se recoge la medición de las variables del proceso (ver variables en sección 4.3.2.2). A lo largo de este tiempo el reactor se somete a cambios en la operación para completar la producción de los distintos productos. No se producen fallos durante la operación.

- EOM2: Se trabaja durante 600 minutos. A lo largo de este tiempo el reactor se somete a un solo cambio de operación para pasar de la producción de *Pa* a la de *Pb* y, luego, para volver a la de *Pa*. No se producen fallos durante la operación.

- EOM3: Se opera el reactor durante unos 650 minutos. A lo largo de este tiempo solo se produce *Pa*. Luego, se produce una pequeña descalibración en la válvula de entrada de alimentación de monómero al reactor.

- EOM4: Es similar al anterior. En este caso se produce una perturbación en el minuto 600 que consiste en una leve oscilación senoidal del caudal de iniciador en la alimentación y que crece escalonadamente. Tras un intervalo de 250-300 minutos se produce una leve rotura en la válvula de alimentación que provoca una pérdida continua y creciente de material por la válvula.

El primer escenario EOM1 se utiliza para diseñar el sistema de monitorización para el reactor de monitorización y los restantes (EOM2, EOM3 y EOM4) se utilizan para evaluar el sistema ya diseñado bajo diferentes situaciones de operación. A continuación se describe el resultado para cada uno.

Análisis inicial de los datos y diseño del sistema de monitorización con EOM1

Los datos de *EOM1* representan la operación histórica del proceso. Dado que por experiencia de las operaciones pasadas se ha visto que las transiciones pueden cubrir intervalos de duración significativos (entre 25 – 60 minutos) se decide aplicar la estrategia Mt.

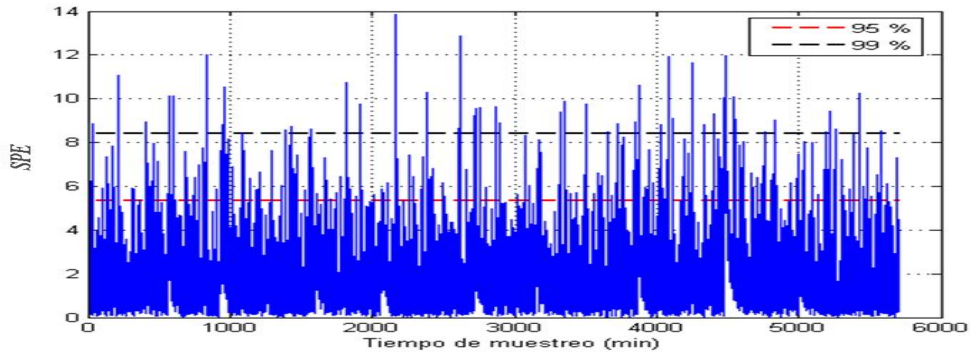

Figura 6.12. *SPE* del ACP sobre los datos originales. Diseño *Mt*.

En los gráficos de las figuras 6.12, 6.13 y 6.14 se muestran los *SPE*, T^2 y *scores* que se obtienen tras el paso 2 de Mt. En el gráfico de T^2 se observan intervalos que tienden a estar más cerca de los límites de control e incluso sobrepasan muchas veces estos límites (ver por ejemplo alrededor del minuto 1000 y del minuto 3000 en la figura 6.13). Estos intervalos corresponden a las condiciones de operación de los productos *Pb* y *Pc*. Así, si se utilizara este gráfico para monitorizar, con algunos productos se producirían muchas falsas.

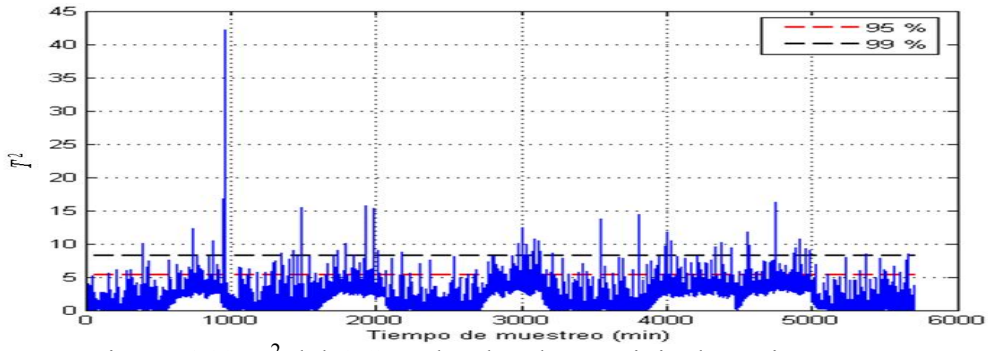

Figura 6.13. T^2 del ACP sobre los datos originales. Diseño Mt.

Tras aplicar la estrategia hasta el paso Mt6, las transiciones quedan eliminadas. El efecto de esto es muy notable en cuanto a que los *clusters* en el gráfico de los *scores* quedan mejor definidos (figura 6.15) que antes del tratamiento de los *outliers* y transiciones (figura 6.14).

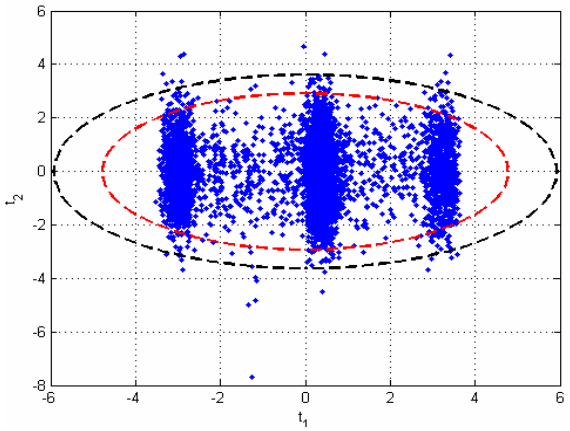

Figura 6.14. *Scores* de los datos hasta el paso Mt3.

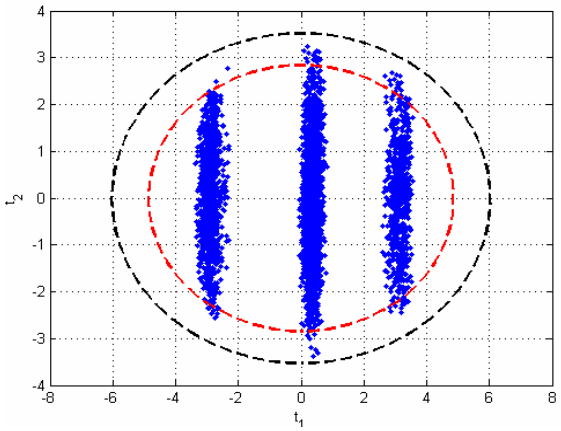

Figura 6.15. *Scores* de los datos hasta el paso Mt6.

Tras aplicar Mt hasta el paso final (Mt11) se obtiene el modelo ACPMg. En las figuras 6.16 y 6.17 se muestran los *SPE* y T^2 correspondientes. Dado que se han eliminado los datos de las transiciones, el número de muestras es ahora menor si se compara con los *SPE* y T^2 de las figuras 6.12 y 6.13. También, se observa que las señales resultantes son homogéneas a lo largo de todo el tiempo considerado. Esto conduce a que en el momento de una desviación bajo cualquier operación el sistema responderá de manera similar.

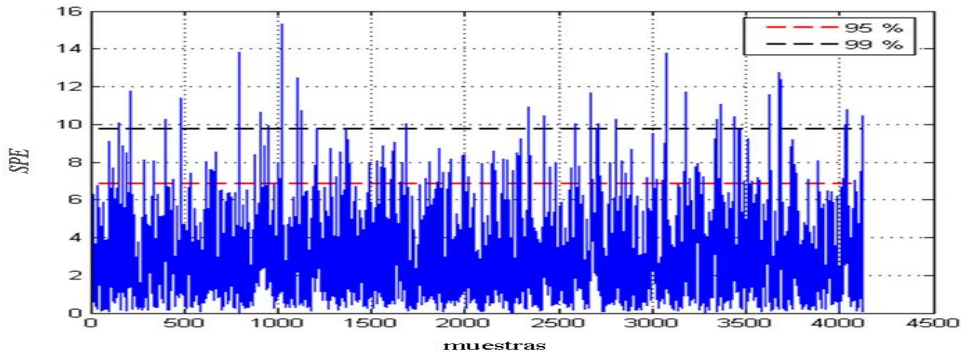

Figura 6.16. *SPE* del ACPMg. Estrategia Mt.

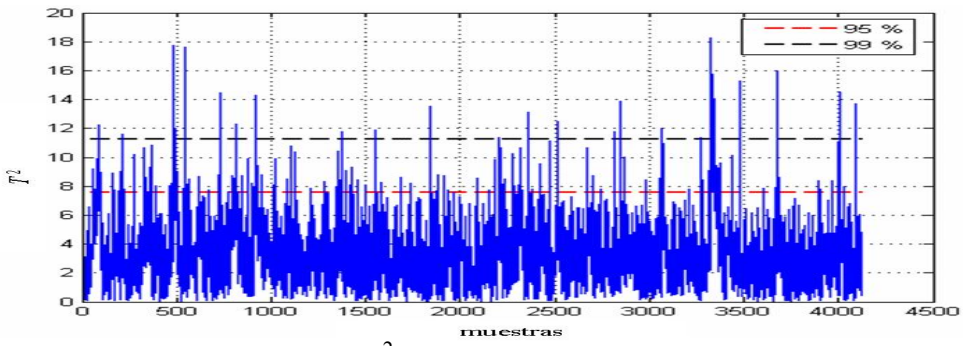

Figura 6.17. T^2 del ACPMg. Estrategia Mt.

Finalmente, se muestra el gráfico de *scores* correspondiente al modelo ACPMg obtenido (ver la figura 6.18). En este se muestra cuan homogéneo es el modelo resultante respecto de los modelos de las figuras 6.14 y 6.15. Las regiones de operación de todos los productos (*Pa*, *Pb* y *Pc*) quedan concentradas dentro de unos mismos límites de confianza del 95 y 99 %.

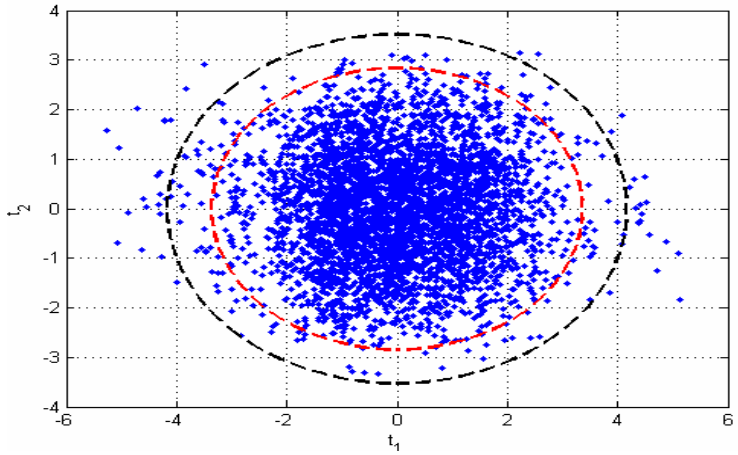

Figura 6.18. *Scores* del ACPMg obtenido.

Análisis de escenario de operación EOM2

Los datos de *EOM2* representan una operación posterior al diseño. Durante la simulación de los mismos se aplica paralelamente la estrategia de monitorización que se describe en la sección 6.1.2.3 y utilizando el modelo ACPMg obtenido en la sección anterior (ver Análisis inicial de los datos y diseño del sistema de monitorización con EOM1). El resultado de la monitorización se muestra en las figuras 6.19 a 6.22.

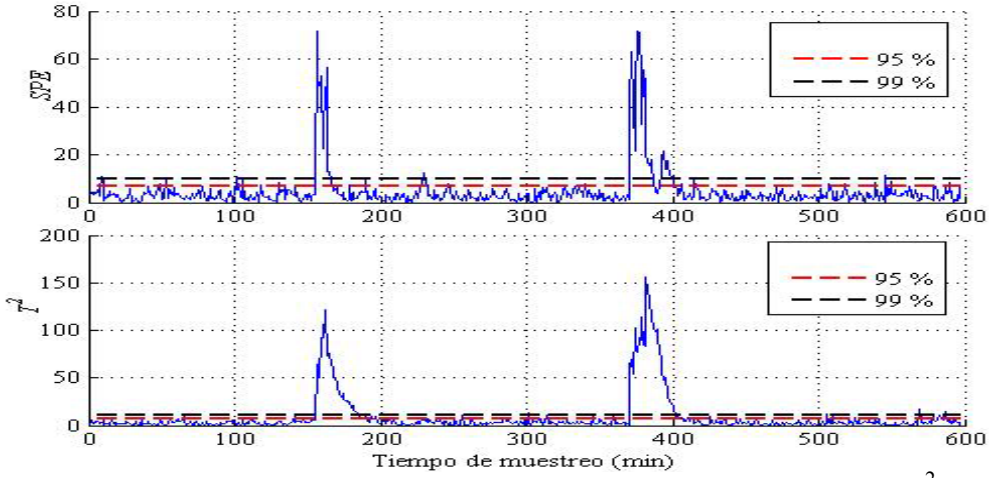

Figura 6.19. Monitorización de un cambio de operación con *SPE* y T^2.

Si se monitoriza solo con los gráficos aportados por el ACPMg (esto es con *SPE* y T^2) tal y como se propone en los trabajos de Hwang y Han (1999) y Morris *et al*, (Martin *et al*., 2002), se obtendrá el resultado de la figura 6.19. Los gráficos para ambos estadísticos muestran 2 desviaciones muy pronunciadas en 2 intervalos separados: un primer pico entre 100 y 200 y otro pico entre 350 y 400. Los picos corresponden al cambio de operación desde la producción de *Pa* a *Pb* (pico 1) y al cambio desde la producción de *Pb* a *Pa*. Aún cuando el operador ya este informado siempre puede haber el riesgo de que reaccione equivocadamente ante la señal que observa en los gráficos, lo que le llevaría a tomar "acciones correctivas" que podrían conducir a una verdadera situación anormal.

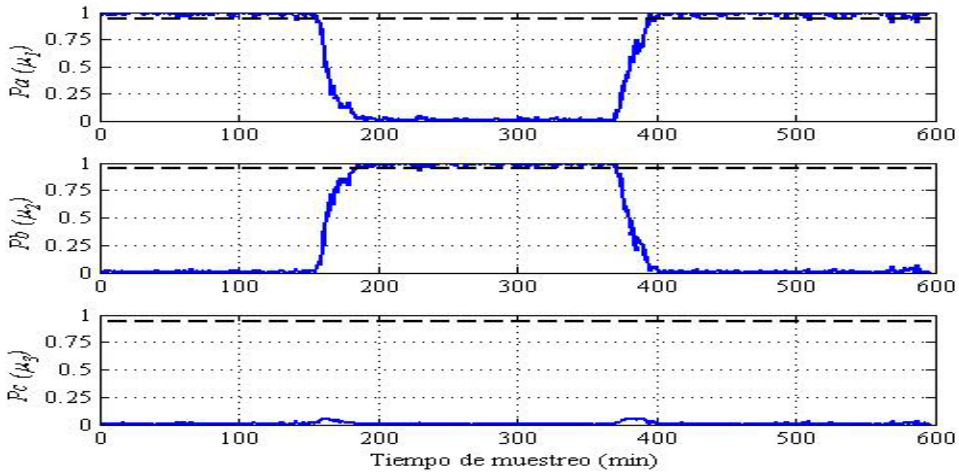

Figura 6.20. Gráficos de pertenencia – caso EOM2

Si por el contrario, solo se utilizan los gráficos de pertenencia a cada *cluster* (ver figura 6.20) tal y como se propone en diversos trabajos anteriores (Næs y Mevik, 1999; Teppola y

Minkkinen, 1999; Rosen, 2001), se vería con claridad los cambios de operación. En efecto, fijándose en la figura 6.20 se observa claramente que el valor de μ_1 (la pertenencia asociada al producto *Pa*) se mantiene en 1 mientras se esta produciendo *Pa*, mientras que las pertenencias de *Pb* y *Pc* (μ_2 y μ_3) respectivamente) permanecen prácticamente en cero. Luego, cuando se produce el cambio de operación entre los minutos 150-200, se observa como μ_1 cae progresivamente hasta cero mientras que el valor de μ_2 comienza a subir hasta estabilizarse en 1, lo que indica el transito y llegada a las nuevas condiciones de operación. Durante todo este intervalo μ_3 permanece en cero lo que indica que el sistema en transito nunca se ha desviado hacia la región de operación correspondiente al producto *Pc*.

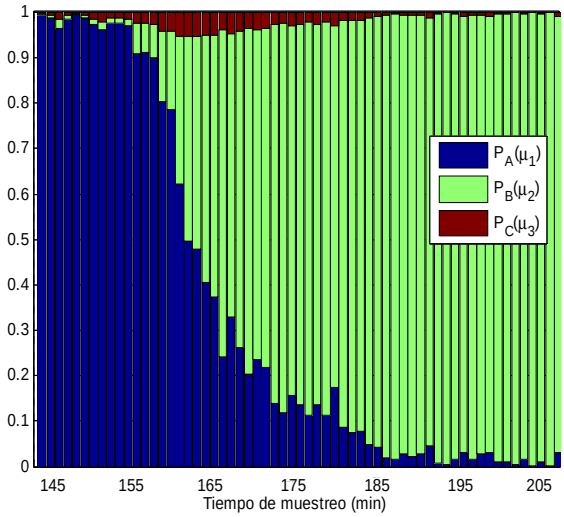

Figura 6.21a. μ conjuntas entre 145-205 min.

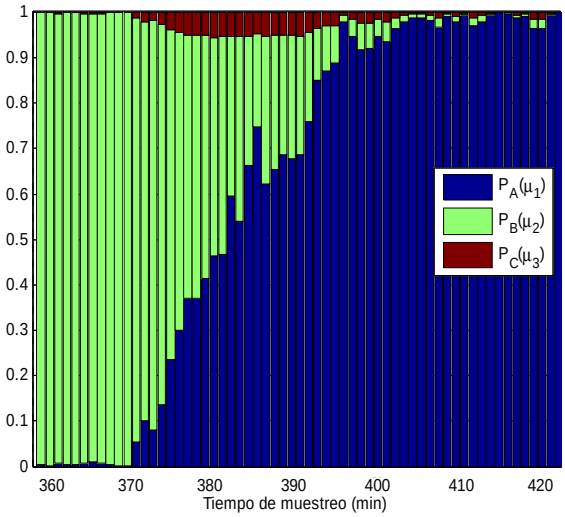

Figura 6.21b. μ conjuntas entre 360-420 min.

Alternativamente, en las figuras 6.21a y 6.21b se presenta el gráfico conjunto de las pertenencias. Cada barra, en cada instante de tiempo, muestra de forma superpuesta de pertenencia de cada muestra a los diferentes grupos. Para una mejor visualización el gráfico se define sobre un intervalo de 60 minutos. Con ayuda del gráfico se visualiza fácilmente el cambio que ocurre y cuando se llega a las condiciones del producto deseado. El resultado obtenido mediante este gráfico es similar a usar el gráfico de la figura 6.20. No obstante y como ya se dijo en una sección precedente, si el número de productos es muy alto, el número de gráficos de pertenencia individuales se incrementaría proporcionalmente con el consiguiente exceso de información para el operador.

Análisis de EOM3

Durante la simulación del caso EOM3 se vuelve a aplicar la estrategia de monitorización que se describe en la sección 6.1.2.3, y utilizando el modelo ACPMg obtenido anteriormente. El resultado de la monitorización se muestra en las figuras 6.22 a 6.24.

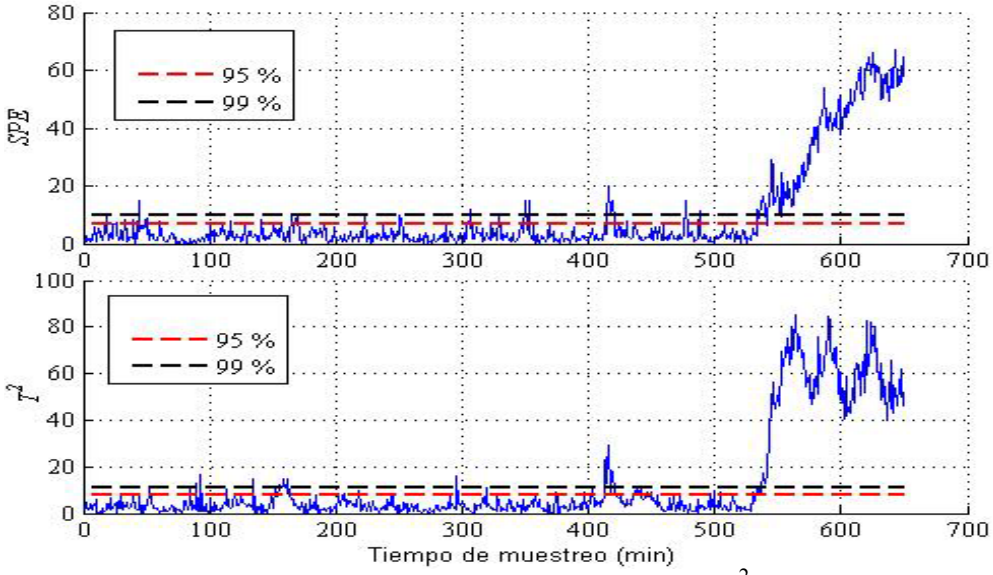

Figura 6.22. Monitorización de un fallo con *SPE* y T^2 - caso EOM3.

En este caso tanto el gráfico del *SPE* como el del T^2 son eficientes para detectar el cambio ocurrido aproximadamente en el minuto 530 (ver figura 6.22).

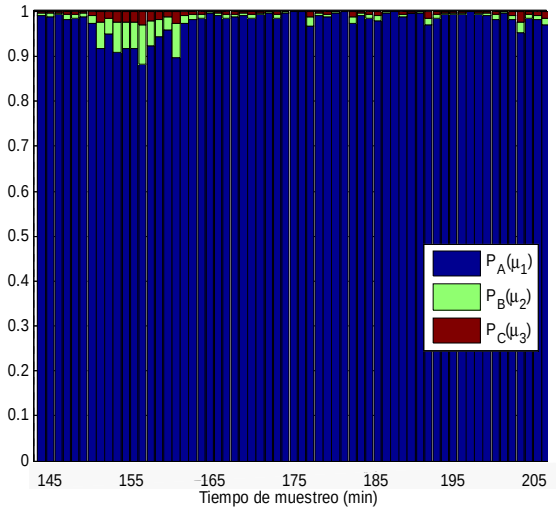

Figura 6.23. μ conjuntas entre 145-205 min.

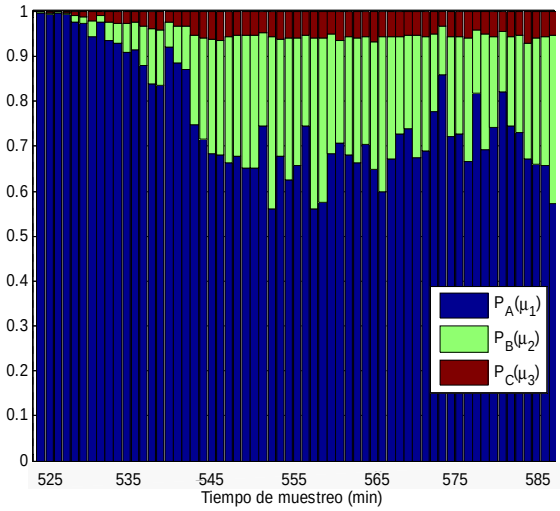

Figura 6.24. μ conjuntas entre 525-585 min.

En las figuras 6.23a y 6.23b se muestra el gráfico de pertenencia conjunta en 2 intervalos distintos. En un primer intervalo [*t*=145, *t*=205] el proceso opera sin problemas lo que se traduce en un predominio de la operación de *Pa* (Sector azul). Dentro del intervalo [t=525, t=585] el proceso sufre el fallo descrito anteriormente. En este caso se observa un decremento de los valores de μ_1 pero la operación de *Pa* sigue predominando (μ>0.6) minutos después que el fallo se ha producido. Como se ve, no es fácil discernir la ocurrencia de un fallo a través del gráfico conjunto pero, por otro lado, puede verse que el sistema permanece aún más cerca de la producción de *Pa*. Esta información puede servir al operador o a un sistema de diagnosis para relacionar el fallo actual con *Pa* y reducir así el número de causas de fallos a analizar.

Análisis de EOM4

Los datos de *EOM4* representan la operación del proceso para producir *Pa* y en la que se presenta un fallo en la válvula de alimentación al reactor. El fallo inicialmente se presenta

como una perturbación muy leve que luego crece. Durante la simulación de este caso se aplica la estrategia de monitorización que se describe en la sección 6.1.2.3. El resultado de la monitorización se muestra en las figuras 6.25 a 6.27.

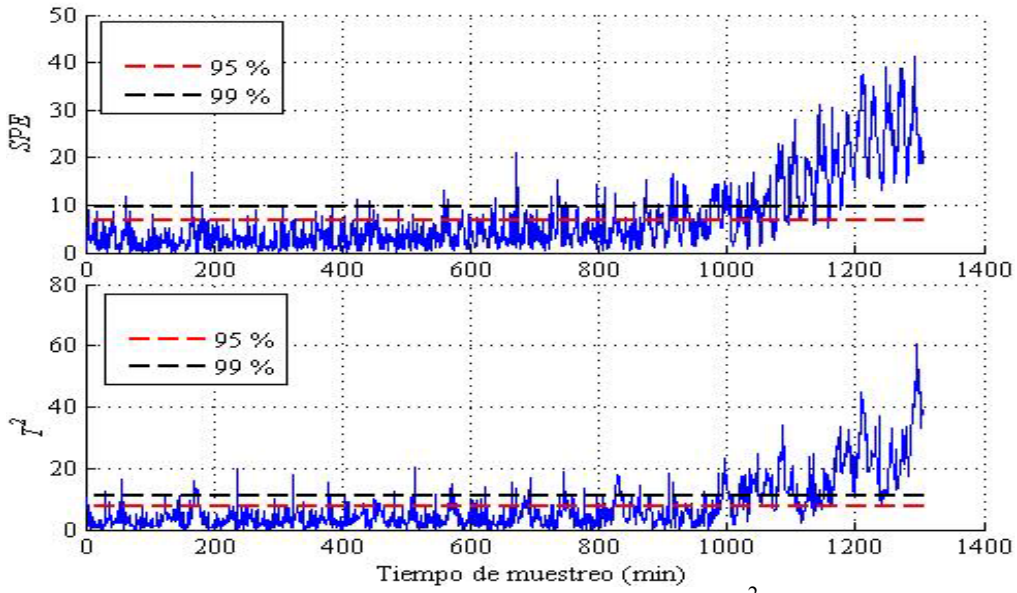

Figura 6.25. Monitorización de un fallo con *SPE* y T^2 - caso EOM4.

Al igual que en el caso EOM3, el fallo se detecta a través de los gráficos *SPE* y T^2, mientras que en el gráfico de pertenencias conjuntas el sistema mantiene como producto base a *Pa*. Gracias a esto los consiguientes esfuerzos del operador se dirigirán a tomar acciones dentro del ámbito de lo conocido sobre la operación de *Pa*.

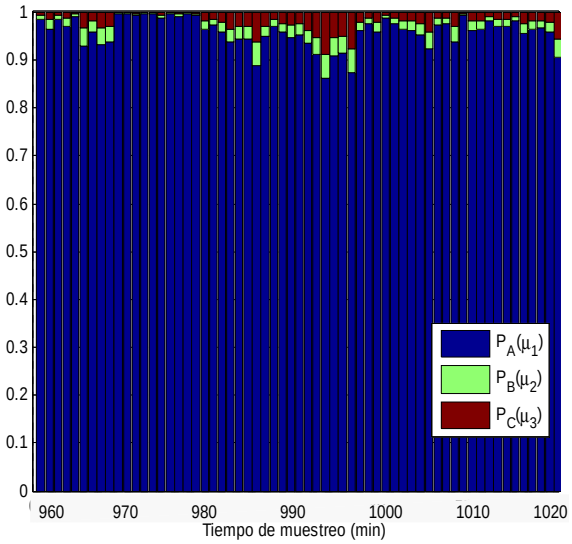

Figura 6.26. μ conjuntas entre 960-1020 min.

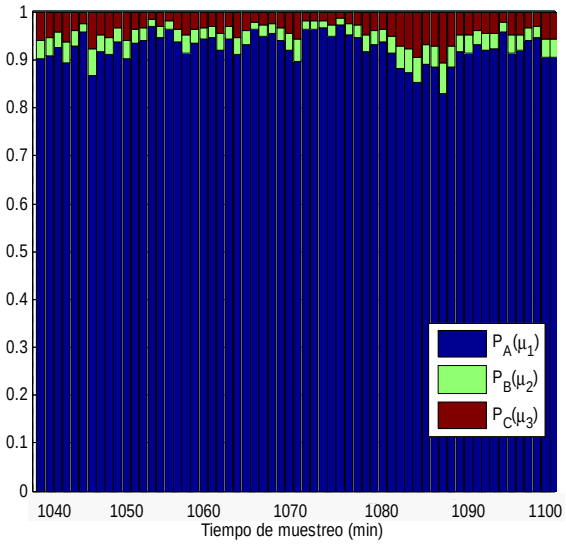

Figura 6.27. μ conjuntas entre 1040-1100 min.

6.2 Supervisión de procesos afectados por decaimiento de la operación

Se sabe que la eficiencia de unidades y equipos de muchos procesos decrece con el tiempo debido a efectos tan distintos como ensuciamiento, desactivación de catalizadores, formación de subproductos, etc. Esto ocasiona paradas frecuentes y periódicas para la limpieza o mantenimiento de equipos, siendo el objetivo de estas paradas el de restablecer las condiciones de operación originales con sus consecuentes beneficios por mejora de la productividad. En la literatura se pueden encontrar diversas propuestas en las que se intenta modelar el ciclo de caída en la eficiencia mediante métodos de políticas óptimas (Epstein, 1979) o formulaciones basadas en métodos de Programación No Lineal (*NLP* por sus siglas en ingles y Programación No Lineal Entero-Mixta (*MINLP* por sus siglas en ingles) para asistir en la programación de las paradas (Sanmartí *et al*., 1997; Jain y Grossman, 1998; Sequeira *et al*., 2003). Por otro lado no se ha explotado la información de los datos históricos del proceso para ayudar a la solución de estos problemas, ni mucho menos se ha estudiado la supervisión asociada a estos procesos.

En esta sección se presenta una estrategia que explota la información de los datos históricos para ayudar a la supervisión de procesos afectados por decaimiento de la operación. La estrategia propuesta es una adaptación de los desarrollos que se presentan a lo largo de la sección 6.1, para los casos específicos a tratar en esta sección.

6.2.1 Caso propuesto: Operación de un reactor afectada por ensuciamiento

En esta sección, se utiliza como caso de estudio un proceso basado en el reactor *CSTR* que se describe en la sección C.3 del anexo C. Con este modelo se propone y se simula el siguiente escenario de operación:

El equipo, luego de haber sido limpiado, se pone en operación. Tras poco tiempo de trabajo se comienza a ensuciar el sistema de intercambio de calor dentro del reactor pero no es sino poco más tarde cuando dicho ensuciamiento se ve reflejado en el aumento continuo de la cantidad requerida de flujo del refrigerante *qc* para intentar mantener la temperatura del reactor T en su valor de consigna. Aún cuando poco a poco el valor de T se estabiliza (por el trabajo continuo aunque lento del sistema de control el sistema tiende a estabilizarse) por siempre ocurre que se llega a un valor límite en el flujo de enfriamiento que entra a la camisa del reactor (límite en el costo de energía). Es por ello que se vuelve a detener el reactor para limpiarlo.

Se asume que durante la operación del reactor se dispone de mediciones de las siguientes variables: nivel de líquido en el reactor L (que se obtiene a partir de V), la concentración de reactivo en el reactor Ca, la concentración del producto a la salida Cb, la temperatura del reactor T, los flujos de entrada y salida del proceso, q_0 y q respectivamente, y el flujo del refrigerante q_C. Se pretende establecer un patrón de desarrollo del proceso, a través de la caracterización del ensuciamiento, y con ello obtener información útil para soportar la monitorización y el mantenimiento del proceso.

6.2.2 Estrategia de análisis del reactor con ensuciamiento

Para analizar el tipo de problemas que plantea el caso de estudio actual (decaimiento en la operación), se pretende utilizar las mismas estrategias desarrolladas en secciones anteriores

(ver sección 6.1.2) y que combinan Técnicas TEM, en este caso el ACP, con *clustering* tipo *CLD*. No obstante, el tipo de problemas que plantea el caso de estudio actual involucra una cierta particularidad en la forma en que se ordenan los datos históricos a lo largo del tiempo que exige una adaptación de las estrategias a utilizar. A continuación, se describen las adaptaciones y la forma final en que se aplica la estrategia.

6.2.2.1 Acondicionamiento de los datos

La operación del proceso en estudio (reactor de polimerización) es tal que continuamente se realizan paradas de mantenimiento y luego se reinicia la operación. Por lo tanto, aún cuando el reactor es de operación continua, se puede decir que tras cada parada se tiene un "nuevo lote de producción" al cual denotaremos como *Pr*. Luego, los datos formarán un arreglo o matriz tridimensional, *Mtr*, tal como se muestra en la figura 6.28 y en el que cada capa horizontal representa una producción *Pr*.

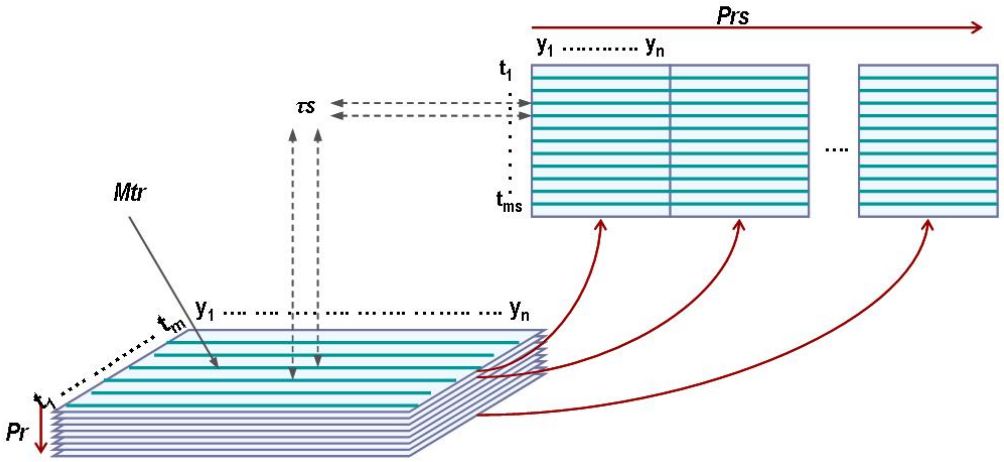

Figura 6.28. Vista temporal de los datos históricos.

En la matriz tridimensional *Mtr*, por cada *Pr* se tienen *n* variables de proceso $\mathbf{y}_i$, cada una de las cuales se ha muestreado *m* veces a intervalos de tiempo sucesivos y de longitud τ. Esta es la misma representación que se utiliza en el análisis y supervisión de procesos discontinuos (Nomikos y MacGregor, 1994; Nomikos y MacGregor, 1995) y que en este caso se adopta para describir los datos del caso estudiado. Como en el caso de los procesos discontinuos, para poder procesar esta matriz tridimensional *Mtr* se hace un desdoblamiento de la misma como se muestra en la figura 6.28. Se toma cada *Pr* y se coloca una al lado de otra tal que los tiempos de muestreo t_i de cada *Pr* representan las filas de la matriz desdoblada *Md* y las variables $\mathbf{y}_i$ de cada *Pr* forman las columnas de *Md*. Debido a que la cantidad de datos a procesar puede llegar a ser muy grande se propone lo siguiente:

- Se fija un valor τs tal que
  $$\tau s > \tau \quad , \quad \tau s \ll m \cdot \tau \tag{6.8}$$
- De cada *Pr* se toman muestras a intervalos sucesivos τs tal que se obtiene una matriz *Prs* con *n* variables y *ms* muestras.

En la figura 6.28 se ilustra lo anterior por añadir líneas sobre las matrices que indican que solo se recogen estas filas de datos para crear las correspondientes *Prs*. Luego, estas *Prs* son las que forman la matriz desdoblada *Mds*.

Lo anterior también es útil para casos en que el número de muestras *m* varíe en cada *Pr*. Tras el procesamiento explicado, todas las *Prs* tendrán la misma longitud *ms*. En los experimentos utilizados, las diferencias encontradas eran pequeñas (no mayor de 6 unidades de tiempo entre todas las *Pr*). Si las diferencias fuesen más significativas se tendría que utilizar algún método de sincronización utilizado en el tratamiento de procesos discontinuos (Kourti, 2003).

6.2.2.2 Análisis de los datos

Una vez los datos se han acondicionado según se ha discutido en la sección 6.2.2.1, éstos se procesan mediante la estrategia Mt desarrollada en la sección 6.1.2.2. Al final de la estrategia Mt, la matriz $\mathbf{U}^{oo}$ obtenida en el paso Mt6 (ver sección 6.1.2.2) se utiliza para determinar la longitud de las etapas que puedan aparecer durante la operación del proceso. También, si fuese requerido por el proceso en estudio, se puede obtener el correspondiente sistema de monitorización (ver sección 6.1.2.3).

6.2.3 Aplicación de la estrategia sobre el escenario propuesto

Se tienen datos de 5 operaciones normales basadas en las condiciones de operación C3 (ver anexo C). De acuerdo a la metodología planteada, lo primero que se hace es acondicionar los datos según se explica en la sección 6.2.2.1. El tiempo total de operación de los experimentos es de más de 9 días de duración (más de 13000 min.). Así, para reducir el volumen de datos que se miden cada 30 segundos, se toman valores a intervalos de cada 10 minutos. Luego, se obtienen los *Prs* y la matriz *Mds*. A continuación se analiza la matriz resultante mediante la estrategia Mt. Como no se tenía una información precisa acerca del número de *clusters c* en que se distribuirían los datos, antes de aplicar el paso Mt3 se estimó *c* usando el método *Subtractive Clustering MSCl-2* (ver sección 5.3.2). Luego, se aplicó Mt3 y se construyó el gráfico de los *scores* correspondientes al modelo ACP obtenido hasta ese momento (ver figura 6.29).

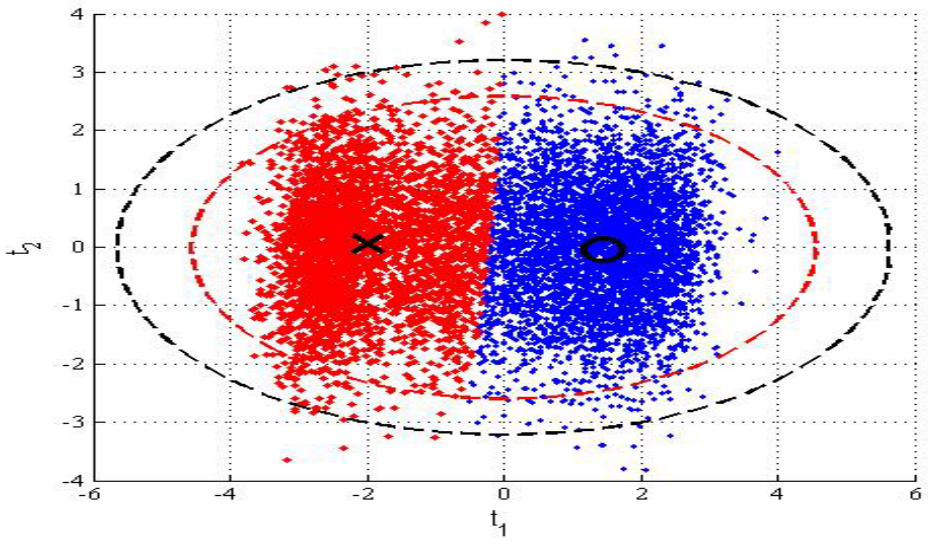

Figura 6.29. *Scores* de los datos hasta el paso Mt3.

El valor de *c*, estimado con la técnica *MSCl-2*. es igual a 2. De la figura 6.29 se puede ver que los datos se han agrupado en 2 regiones:

- Una región R1 a la izquierda con su centro identificado con una X. Esta región o grupo tiende a ser más densa, en cantidad de puntos, hacia su extremo izquierdo y a la vez menos densa a medida que se acerca al otro grupo.

- Una región R2 a la derecha con su centro identificado con una O. En este caso la región identificada tiende a ser menos densa a medida que se acerca al otro grupo y más densa hacia su extremo derecho. No obstante, se muestra levemente más homogénea, en densidad, que la región anterior.

Si se continúa la aplicación de la estrategia Mt hasta el paso Mt6, se puede construir el gráfico de *scores* que se muestra en la figura 6.30.

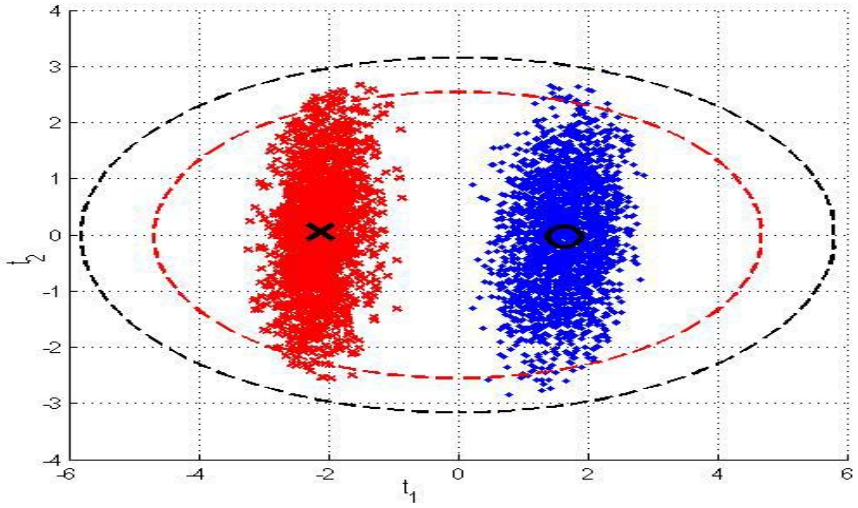

Figura 6.30. *Scores* de los datos hasta el paso Mt6.

En este nuevo gráfico se han eliminado muchos datos y solo quedan representadas 2 regiones muy separadas entre sí.

¿Por qué se han detectado 2 *clusters*?,¿Qué representan los datos eliminados? Aún cuando parece un poco ambigua la estructura de datos obtenida hasta el paso Mt3 (2 grupos identificados) y la subsiguiente eliminación de datos, la misma obedece a lo siguiente:
Como se puede recordar de la descripción hecha en la sección 6.2.1, el reactor se lava, se pone en funcionamiento y comienza a sufrir de un ensuciamiento continuo en el sistema de transferencia de calor. Tomando la temperatura del reactor *T* durante una de las producciones disponibles (*Pr*) se obtiene el gráfico de la figura 6.31.

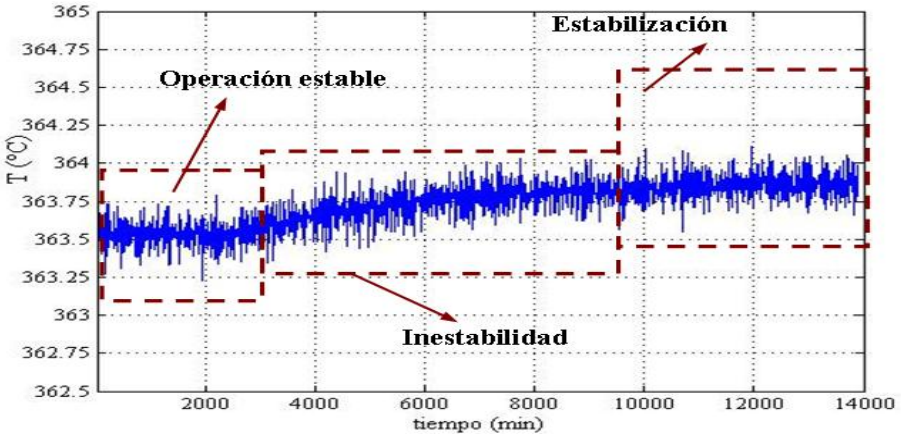

Figura 6.31. Curva de *T* durante una Producción *Pr*.

En el se observa que a lo largo de un intervalo de tiempo inicial (más o menos hasta t = 2000-3000 min.) la T se mantiene estable alrededor del valor 363.5 ºC. Luego, comienza a subir leve y lentamente hasta aproximadamente después de los 9000 min., en que T logra estabilizarse alrededor de un nuevo valor (aprox. 363.8 ºC). Esto esta reflejando que solo a partir de t = 2000-3000 min., el efecto de ensuciamiento comienza a hacerse notorio como para afectar los valores de las variables del sistema. Antes de eso, el sistema se mantiene estable en su operación. También está reflejando algo que se asumió durante el desarrollo del modelo del reactor: que el sistema de control no es capaz de mantener la temperatura en su valor inicial en el momento en que el efecto de ensuciamiento comienza a hacerse notorio y que luego de un largo tiempo (muchos minutos) se logra estabilizar el sistema a expensas de un gasto cada vez mayor de energía. Esto último puede verse con más claridad a través del gráfico del flujo de enfriamiento q_C (ver figura 6.32). Como respuesta del sistema de control, en el instante en que T comienza a variar el valor de q_C se incrementa y, aún cuando poco a poco se va consiguiendo estabilizar a T, q_C sigue aumentando hasta un valor límite de 15.6 lt/min., en el que, como ya se señalo en la sección 6.2.1, el costo de energía se hace prohibitivo (ver nuevamente la figura 6.32).

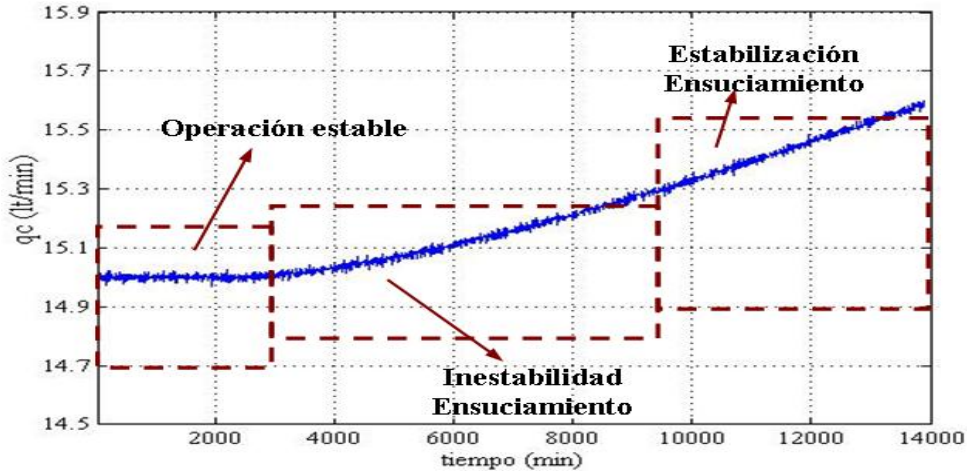

Figura 6.32. Curva de q_C durante una Producción *Pr*.

Las tres etapas identificadas se han etiquetado como "Operación estable", "Inestabilidad" y "Estabilización" tanto en la figura 6.31 como en la figura 6.32. Lo que identifican las técnicas utilizadas como grupos principales corresponde a las etapas "Operación estable" y "Estabilización" con la particularidad de que en ambas el sistema es estable (etapa inicial) o tiende cada vez más a ello (etapa final). Los puntos que se eliminan corresponden a la etapa de inestabilidad (la transición) entre las etapas estable inicial (grupo de la izquierda en la figura 6.30) y la etapa de estabilización final (grupo de la derecha). A través de los datos de $\mathbf{U}^{oo}$ se logra establecer de manera aproximada la longitud de dichas etapas en cada una de las 5 producciones conjuntamente analizadas. Los valores de estas etapas se muestran en la tabla 6.2.

Tabla 6.2. Tiempo identificado de las etapas del proceso.

Producción *Pr*	Etapa Inicial Estable (min.)	Etapa Inestabilidad (min.)	Etapa Estabilización (min.)
1	3000	6800	4010
2	3010	6820	4010
3	3020	6800	4030
4	3010	6780	4020
5	3020	6800	4040

El valor de los intervalos identificados es más o menos semejante. Esto es indicativo de que la identificación es buena ya que todas las producciones *Pr* analizadas corresponden a un mismo tipo de operación normal. Además, si se observa con atención el gráfico de otras variables como la concentración de reactivo *Ca* (figura 6.33) se puede ver que las etapas identificadas están reflejadas en el comportamiento de muchas de dichas variables.

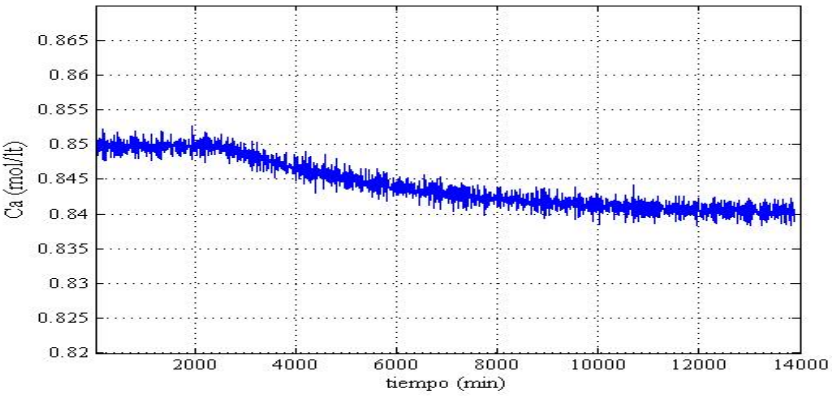

Figura 6.33. Curva de *Ca* durante una Producción *Pr*.

El promedio de los valores de cada etapa (ver tabla 6.2) puede servir como información para el mantenimiento del proceso. Por otro lado, durante la operación el gráfico 6.31b se puede utilizar para monitorizar el proceso y ayudar a identificar el inicio del ensuciamiento y cuando se esta en la etapa de estabilización. Así, al identificar el inicio de cada etapa y por los valores promedios de las mismas se puede establecer una sencilla proyección de cuando terminara el proceso actual.

6.3 Conclusiones

En este capítulo se ha presentado la aplicación de una serie de desarrollos para la supervisión de distintas situaciones que se pueden presentar en la operación de procesos químicos.

En una primera parte se estudia y se propone una forma de analizar procesos multioperacionales. A diferencia de las estrategias existentes en la literatura, se introduce el tratamiento de las transiciones durante la etapa de diseño del sistema de monitorización, como soporte al operador durante un cambio de operaciones y/o para ayudarle a reducir rápidamente el número de posibles causas de anormalidades tras la ocurrencia de un fallo.

En una segunda parte, las estrategias anteriores se extienden al análisis de procesos que se ven afectado por ciclos de decaimiento en la operación. La aplicación sobre el caso de estudio mostrado permite ver cómo se puede explotar la información de operaciones pasadas tanto para profundizar en el conocimiento del proceso como para asistir en la supervisión del proceso y en la planificación y mantenimiento del mismo.

NOMENCLATURA

c	Número de grupos o *clusters*.
ds	Vector de distancias de cada muestra en **Y** a cada centro de *cluster* $\mathbf{v}_i$.
De	Matriz de desviaciones estándar de **Y**.
IC_k	Intervalos de confianza asociados a la media de un vector $\boldsymbol{\mu}_{ff}$.
IC_k^s, IC_k^i	Valores límites superior (s) e inferior (i) de IC_k.
L	Nivel de descomposición.
m	Número de mediciones o muestras disponibles en una matriz *Pr*.
ms	Número de mediciones o muestras disponibles en una matriz *Prs*.
Md	Matriz desdoblada obtenida a partir de la matriz M_t.
Mds	Reducción de la matriz *Md* a solo *ms* mediciones.
Mtr	Arreglo tridimensional de datos tras realizar varias *Pr*.
n	Número de variables medidas.
P	*loadings* obtenidos con un ACP.
Pr	Ciclo o lote de producción entre 2 paradas de mantenimiento.
Prs	Reducción de la matriz Pr a solo *ms* mediciones.
$\mathbf{Q}_k$	Matriz covarianza de $\mathbf{Y}_k$.
$\mathbf{Q}_{cn}$	Matriz covarianza única calculada a partir de las $\mathbf{Q}_k$.
SPE	Estadístico que se obtiene del ACP y representa el error de predicción al cuadrado o *SPE* (*Squared Predictive Error*).
SPE_{cn}	*SPE* asociado a la matriz $\mathbf{Y}_{cn}$.
t	*Scores* obtenidos con un ACP.
T^2	Estadístico de *Hotelling* que se obtiene del ACP.
T_{cn}^2	T^2 asociado a la matriz $\mathbf{Y}_{cn}$.
U	Matriz **U** de pertenencias.
$\mathbf{v}_k$	prototipos del *cluster* o centros de *cluster*
$\mathbf{y}_i$	Variable de proceso
Y	Matriz de datos originales.
$\mathbf{Y}^e$	Matriz de datos estandarizados.
$\mathbf{Y}^o$	Matriz de datos originales tras eliminar *outliers*.
$\mathbf{Y}^{oe}$	Matriz de datos estandarizados tras eliminar *outliers*.
$\mathbf{Y}^{oo}$	Matriz de datos originales tras eliminar *outliers* y datos de transiciones.
$\mathbf{Y}^{ooe}$	Matriz de datos estandarizados tras eliminar *outliers* y datos de *transiciones*.
$\mathbf{Y}_{cn}^o, \mathbf{Y}_{cn}^{oo}$	Matrices conjuntas formadas por la unión de las matrices $\{\mathbf{Y}_k^{oe}\}, \{\mathbf{Y}_k^{ooe}\}$.
$\mathbf{Y}_k^o, \mathbf{Y}_k^{oo}$	Matrices con datos de $\mathbf{Y}^o$, $\mathbf{Y}^{oo}$ que corresponden al grupo *k*.
$\mathbf{Y}_k^{oe}, \mathbf{Y}_k^{ooe}$	Matrices de datos de $\mathbf{Y}_k^o, \mathbf{Y}_k^{oo}$ estandarizados.
$\bar{\mathbf{Y}}$	Vector de las medias de cada variable en **Y**.

LETRAS GRIEGAS

σ_{ff}	Varianza muestral de $\boldsymbol{\mu}_k^{ff}$.
μ_{ik}	Pertenencia individual de un objeto (una observación) *i* a un grupo *k*.
$\boldsymbol{\mu}_k^{o+}$	Vector de observaciones de $\boldsymbol{\mu}_k^o$ (o $\boldsymbol{\mu}_k^{oo}$) que pertenecen a cada grupo *k*.
$\boldsymbol{\mu}_k^{of}$	Filtrado del vector $\boldsymbol{\mu}_k^{o+}$ obtenido mediante la estrategia *levashrink*.
$\boldsymbol{\mu}_k^{off}$	Suavizado del vector $\boldsymbol{\mu}_k^{of}$ obtenido mediante ecuación 2.14 (capítulo 2).

$\bar{\mu}_k^{ff}$	Media del vector $\boldsymbol{\mu}_k^{ff}$
τ	Intervalo de tiempo entre cada muestreo.
τs	Intervalo de tiempo entre cada muestreo para obtener las matrices *Prs*.

SUPERÍNDICES

e	Variables estandarizadas.
o	Variables obtenidas tras eliminar *outliers* de **Y**.
oo	Variables obtenidas tras eliminar *outliers* y transiciones de **Y**.

SUBÍNDICES

i	i-ésima observación, objeto o muestreo.
k	k-ésimo grupo o *cluster*.
t	tiempo de muestreo.

ACRÓNIMOS

ACP	Análisis de Componentes Principales.
ACPGT	ACP Global Tradicional.
ACPMg	ACP Multigrupo.
CLD	*Clustering* basado en Lógica Difusa.
FCM-GK	Técnica *Fuzzy C-Means* modificada con la variante de *Gustaffson-Kessel*.
FPCM-GK	Técnica *Fuzzy Possibilistic C-Means* modificada con la variante *Gustaffson-Kessel*.
MINLP	Programación No Lineal Entero-Mixta o *Mixed Integer Nonlinear Programming*.
NLP	Programación No Lineal o *Non-linear Programming*.
TEM	Técnicas Estadísticas Multivariables
EOMi	Etiqueta para los escenarios del caso del reactor de polimerización.
Mc	Etiqueta para la estrategia de diseño de sistemas de monitorización sin tratamiento de transiciones.
Mt	Etiqueta para la estrategia de diseño de sistemas de monitorización con tratamiento de transiciones.
TEM-*CLD*	Estrategias de análisis que combinan Técnicas TEM con Técnicas *CLD*.

CAPÍTULO 7. CONCLUSIONES Y TRABAJO FUTURO

La motivación del trabajo desarrollado en esta tesis ha sido el reto de la extracción y el aprovechamiento de la información subyacente en los datos históricos de proceso para su utilización como soporte en la toma de decisiones de múltiples tareas asociadas a la Industria Química y de Procesos (IQP). Una revisión atenta de la literatura sobre Ingeniería Química y de Procesos revela el gran interés de académicos e industriales en el desarrollo de estrategias de análisis de datos que den respuesta a dicho reto. El interés ha sido tal que el número de técnicas y estrategias propuestas en la literatura actual es muy amplio. Basta con adentrarse en el aun reciente campo de Knowledge Discovery in Databases (*KDD*) para darse cuenta de ello. Pese al esfuerzo hecho, tras la revisión atenta de los desarrollos realizados se observa que:

- Gran parte de las propuestas se han desviado hacia perfeccionamientos teóricos de las técnicas utilizadas más que a dar respuestas efectivas a los problemas que se pretende resolver.
- Lo anterior ha conducido a que, para algunos problemas, existe un alto número de propuestas. Esto as u vez ha provocado la necesidad de comparación entre las propuestas existentes para lograr definir cual puede dar mejor respuesta a lo que se busca resolver o soportar.

Así, a lo largo de esta tesis se ha hecho hincapié en estudios comparativos de las estrategias existentes más utilizadas para cada una de las situaciones abordadas, siendo los resultados de dichos estudios una de las contribuciones primarias del presente trabajo de tesis.

Por otro lado, el abanico de problemas de toma de decisión en la IQP es muy amplio guardando todos ellos una estrecha relación con el tratamiento de la información proveniente de los datos históricos de proceso. Dentro de este amplio espectro de problemas, esta tesis trata esencialmente de aquellos relacionados con el soporte de tareas operacionales en los niveles de decisión más bajos en la típica jerarquía de control de una IQP. Dentro de este contexto, los esfuerzos se han orientado principalmente a asegurar la calidad de los datos de mediciones de proceso a través de los métodos de rectificación y a asistir en la supervisión de procesos.

Contribuciones en el área de Rectificación de datos

Los métodos de rectificación se han estudiado durante más de 3 décadas y, no obstante, se siguen explorando nuevos desarrollos. Ello es debido, por un lado, al hecho de que datos incorrectos (contaminados por ruidos) pueden conducir a obtener conclusiones erróneas en las tareas operacionales que los utilizan, mientras que el uso de datos verificados o rectificados minimizará el riesgo de errores en su aplicación. Por otro lado, el número de tareas operacionales que utilizan estos datos es muy amplia (control regulador, monitorización, diagnosis de fallos, control de calidad, optimización en tiempo real,…etc.).

En un primer enfoque, se analizan las técnicas de filtración univariable (capítulo 2). En esta área, estudios de años recientes han puesto de manifiesto la superioridad de las técnicas *wavelets* para filtrado frente a otras técnicas más clásicas. En el desarrollo de la tesis, se logra llenar vacíos en las propuestas con *wavelets* existentes como:

- Aportar evidencia experimental de cuales son aquellas funciones *wavelets* más apropiadas para la filtración de diversos patrones de curvas y proponer guías concretas de selección. En este sentido, se han comprobado las ventajas de las *wavelets daubechies db4*, *db8* y, especialmente, *db1* para el filtrado de señales con presencia de patrones estacionales diversos.
- Proponer un paso de estimación del nivel asociado a la descomposición de una señal usando *wavelets* que ofrece una alternativa simple a la decisión sobre el nivel a fijar cuando se quiere analizar una o varias señales de proceso disponibles. Lo anterior, utilizado como paso previo de las estrategias de filtración con *wavelets* (estrategia *levashrink*) redunda en mayor autonomía de las aplicaciones sobre todo para casos en línea y sin perder calidad en la precisión de los estimados.
- Abrir una nueva vía para conseguir una respuesta más efectiva al problema de la selección de la función *wavelets* a través de métodos de rectificación combinada. Dicho método no se muestra suficientemente valido en términos de precisión de los estimados, pero abre una vía alternativa para solucionar el problema planteado.

En un segundo enfoque, se aborda la rectificación asociada al problema de la Reconciliación de Datos para casos dinámicos y lineales (capítulo 3). A diferencia de la Reconciliación de Datos de sistemas en estado estacionario, los desarrollos de RD para casos dinámicos siguen siendo comparativamente muy poco numerosos, además de que las estrategias propuestas son de difícil aplicación en situaciones reales. La aportación de esta tesis es significativa. Se propone una estrategia que integra el rectificado individual de las variables mediante *wavelets* (estrategia *levashrink* desarrollada en el capítulo anterior) con la reconciliación conjunta de estos rectificados basados en una técnica de representaciones polinomiales. Los aportes más significativos de esta integración son:

- La técnica propuesta explota de manera efectiva la redundancia temporal y funcional de los datos a través del uso de las tendencias filtradas del proceso.
- Gracias al uso de las tendencias, se logra una reducción de la complejidad en la estimación de la varianzas de las variables del proceso, que ha sido un tema de continuo debate en la literatura.
- Como resultado, el método propuesto mejora de forma sustancial la estimación de los métodos de reconciliación dinámica basados en representaciones polinomiales, siendo a la vez competitivo frente a otras alternativas de RD dinámicas.

Asimismo, la aplicabilidad de esta estrategia se asegura en situaciones donde se requiere rectificación en línea a través de un enfoque de horizonte móvil. Finalmente, se propone una primera extensión del método para casos no lineales que generalizan los resultados de la propuesta.

Contribuciones en el área de Supervisión de Procesos

En el área de supervisión de procesos los trabajos se han orientado a revisar las estrategias existentes para la monitorización de situaciones específicas que se presentan en la industria química y que en la literatura no se han estudiado suficientemente.

Una de estas situaciones es la monitorización de procesos que pueden verse frecuentemente afectados por anormalidades de aparición lenta (capítulo 4). Aunque el número de estrategias propuestas en la literatura no es muy alto, se propone un análisis comparativo de estas. Todas las estrategias se basan en combinar una técnica de filtrado con un Análisis de Componentes Principales. La comparación llevada a cabo incluye algunas variantes en las que se adopta el filtrado con *wavelets* incluyendo las variantes que se proponen en el capítulo 2 de la tesis. Las estrategias propuestas en esta tesis obtienen en muchos casos resultados superiores cuando se comparan con los métodos actuales para la detección de este tipo de perturbaciones. Adicionalmente, reducen drásticamente la aparición de problemas paralelos como la generación de alarmas falsas.

También se ha estudiado la supervisión de procesos multioperacionales. La cantidad de estrategias de supervisión propuestas para este tipo de procesos en la IQP es considerable, la mayoría de las cuales se han decantado hacia el uso combinado de Técnicas Estadísticas Multivariables, particularmente el ACP, con técnicas de *Clustering* basadas en Lógica Difusa (TEM-*CLD*).

- Una primera contribución en esta área ha sido el estudio comparativo de las estrategias existentes en cuanto a las diversas técnicas *clustering* que utilizan (capítulo 5). Es la primera vez que se establece un análisis comparativo riguroso sobre este tipo de estrategias y para el caso de aplicaciones en la IQP.
- Como consecuencia de la comparación, se ha logrado establecer cual de ellas brinda una mejor identificación de *clusters* de formas diversas y métodos de interpretación y manejo eficiente de *outliers* cuando estos están presentes en los datos disponibles.
- El estudio realizado añade extensiones a las técnicas existentes que conducen a un mejor manejo de todos los aspectos mencionados anteriormente.
- Adicionalmente, se establecen alternativas para afrontar tanto situaciones en las que se tiene conocimiento previo del número de grupos en que se estructuran los datos como aquellas situaciones en las que se carece de dicho conocimiento previo de los datos.

Finalmente, se proponen estrategias de apoyo al análisis y monitorización de procesos continuos multioperacionales (capítulo 6) basadas en la combinación de los resultados anteriores sobre las estrategias TEM-*CLD* con ACP multigrupo. Las estrategias resultantes, aplicadas sobre un caso práctico muy ilustrativo, muestran como gracias al buen manejo de los problemas estudiados en el capítulo 5 para el *clustering* se pueden considerar aspectos como el manejo de las transiciones de operaciones, tema comúnmente obviado en la literatura existente, junto con una exitosa monitorización de anormalidades cuando estas ocurren.

Finalmente, se muestra una extensión de los métodos anteriores para la supervisión de procesos afectados por ciclos de decaimiento en las operaciones. Se muestra a través de un caso de estudio muy ilustrativo, el potencial de la estrategia propuesta para explotar la información de los datos históricos y asistir no solo en la monitorización sino en el la planificación de las operaciones y el mantenimiento.

Trabajo Futuro

A lo largo de la tesis se ha visto que los resultados obtenidos tras la evaluación de las diversas estrategias propuestas son bastante satisfactorios. No obstante, para ampliar el margen de generalización de cada una, sería deseable hacer nuevas evaluaciones sobre escenarios académicos e industriales. Esta propuesta es valida tanto para las estrategias de filtración y reconciliación dinámica (capítulos 2 y 3) como para las estrategias de supervisión que se estudian a lo largo de los capítulos 4 a 6.

Con la estrategia de rectificación combinada que se propuso en el capítulo 2, se ha abierto una puerta para la solución al planteamiento de la selección de la wavelets más adecuada para filtración. En efecto, la combinación de funciones *wavelets Daubechies* propuesto ha arrojado resultados bastante prometedores. Pese a ello, sería deseable explorar nuevas alternativas de combinación con las que se mejoren los resultados obtenidos para un amplio margen de señales. Las alternativas a probarse podrían incluir la combinación de los coeficientes wavelets

En cuanto a los desarrollos de Reconciliación Dinámica, sería deseable hacer evaluaciones adicionales de la extensión no lineal propuesta que incluyan comparaciones con otras estrategias existentes en la literatura.

En la literatura reciente sobre supervisión de procesos, se ha destacado el uso de *Support Vector Machine* como técnica de *clustering* y clasificación de datos y Análisis de Componentes Independientes como técnica estadística multivariable. Podría ser interesante integrar ambas técnicas dentro del marco de las estrategias de TEM-CLD que se han trabajado en esta tesis y ver si aportan alguna mejora en el análisis y/o la monitorización de procesos.

Por último, se ha visto que la información aportada por las estrategias desarrolladas en el capítulo 6 puede ser muy útil para asistir tareas posteriores como la diagnosis de fallos y la planificación de operaciones. Luego, sería interesante evaluar la integración de las estrategias propuestas con sistemas de diagnosis y planificadores de modo de establecer el uso optimo de la información compartida.

ANEXOS

A. Análisis Multiescala o Multiresolución

El Análisis multiresolución o multiescala describe procedimientos mediante los cuales la información procedente de un sistema se intenta reorganizar en categorías llamadas niveles, resoluciones o escalas, de manera tal que en cada uno de ellos el nivel de detalle de la información es distinto. Para intentar ilustrar lo anterior, supóngase una imagen de un bosque cuya información se reorganiza mediante un análisis multiresolución. En el nivel más alto de la jerarquía resultante la información que se recoge es esta: una gran franja de color verde para el bosque y otra de color azul para el cielo. Bajando al siguiente nivel, se logra tener un mayor detalle y nitidez en la imagen tal que se puede distinguir un árbol de otro. Si se baja otro nivel, es posible diferenciar las ramas, ver los claros entre las ramas, etc. Así, si se sigue bajando en la jerarquía, se logrará un detalle cada vez de los diversos objetos en la imagen. De este modo, un análisis multiresolución proporciona una manera de agrupar cosas para revelar aspectos de la estructura que dependen de la escala de actividad.

De un modo más formal, los métodos multiescala se definen como procesos de aproximación basados en el uso de una secuencia de subespacios anidados, $\mathbf{V} = \{\mathbf{V}_l\}_{l\geq 0}$ $l\geq 0$, de resolución incrementada tal que se cumplen las siguientes propiedades (Mallat, 1989):

$$\mathbf{V}_l \subset \mathbf{V}_{l+1} \quad \forall \quad l \in \mathbb{Z} \tag{A.1}$$

$$\bigcap_{l\in\mathbb{Z}} \mathbf{V}_l = \{0\} \tag{A.2}$$

$$\bigcup_{l\in\mathbb{Z}} \mathbf{V}_l = L^2(\mathbb{R}) \tag{A.3}$$

Donde el índice l indica la resolución o escala, $L^2(\mathbb{R})$ es un espacio apropiadamente definido para describir funciones a más de una escala de resoluciones $\mathbb{Z}$ indica el conjunto de los números enteros y $\mathbb{R}$ indica el conjunto de los números reales. Luego, dada una señal de mediciones **y**(t), se pueden obtener aproximaciones suavizadas de la misma, en los distintos espacios $\mathbf{V}_l$, siempre que se disponga de funciones bases, definidas en cada espacio $\mathbf{V}_l$, para hacer la proyección correspondiente. Adicionalmente, por cada $\mathbf{V}_l$ existe un subespacio $\mathbf{W}_l$ que constituye el complemento ortogonal de $\mathbf{V}_l$ en $\mathbf{V}_{l+1}$ tal que:

$$\mathbf{V}_{l+1} = \mathbf{V}_l \oplus \mathbf{W}_l \tag{A.4}$$

Donde $\oplus$ denota la suma directa ortogonal. Por aplicación recursiva de la ecuación (A.4) a sucesivos valores de l, se puede obtener la siguiente expresión multiescala:

$$\mathbf{V}_L = \mathbf{V}_{lo} \oplus \bigoplus_{i=lo}^{L-1} \mathbf{W}_i \tag{A.5}$$

Luego, se puede obtener la aproximación de una función a una resolución $l = L$ a partir de una aproximación a una resolución más baja y una secuencia de detalles a distintos valores de $l < L$.

El análisis multiescala es muy útil para describir una señal. La idea es simple: Dada una señal cuyo valor de resolución es l=0, a una primera resolución o escala l=1, se separa la información de la misma en 2 partes: una principal $\mathbf{As_{l=1}}$ que atrapa o retiene un mayor porcentaje de información (determinística de baja frecuencia de la señal) y otra residual **Rs** que idealmente solo retiene información de alta frecuencia asociada al ruido en la señal. Posteriormente, al ir aumentando l (i=2,…L), se puede seguir separando más información de alta frecuencia de la señal original que permanece en cada nueva $\mathbf{As_{l=i}}$. Con esta separación de información se puede facilitar la eliminación ruido de la señal y, además, se pueden ir descubriendo distintos patrones, inicialmente ocultos a distintas escalas, en la señal.

B. Reconciliación basada en los Filtros de Kalman, de los flujos de masa.

El Filtro de Kalman (FK) es una técnica de filtración basada en espacios de estados que ha sido probada para reconciliación (Benqlilou et al, 2002; Narasimhan y Jordache, 2002). En esta tesis se ha adoptado como la DDR de referencia para comparar con la propuesta WEPA del capítulo 3. A continuación se describe el FK para reconciliar los flujos de masa en el caso de la sección 3.4.1. El modelo dinámico para este caso queda descrito por la ecuación (3.27). Sin embargo, en la aplicación del FK se requiere de un modelo dinámico en forma de espacios de estados discretos con la forma general:

$$\mathbf{x}(k_C) = \mathbf{A}(k_C)\mathbf{x}(k_{C-1}) + \mathbf{B}(k_C)\mathbf{o}(k_C) + \omega(k_{C-1}) \tag{B.1}$$

$$\mathbf{u}(k_C) = \mathbf{H}(k_C)\cdot\mathbf{x}(k_C) + \varepsilon(k_C) \tag{B.2}$$

donde $\mathbf{x}(k_c)$ es el vector de las variables de estado, $\mathbf{o}(k_c)$ es el vector de entradas/salidas manipuladas, $\omega(k_c)$ es el vector de perturbaciones aleatorias en el modelo, $\mathbf{u}(k_c)$ es el vector de mediciones y $\varepsilon(k_c)$ es el vector de errores aleatorios en las mediciones. $\mathbf{A}(k_c)$, $\mathbf{B}(k_c)$, y $\mathbf{H}(k_c)$ se conocen como las matrices de transición, de ganancia del control y de las observaciones respectivamente. Son matrices del sistema de dimensión apropiada.

La ecuación B.1 describe la evolución dinámica de las variables de estado y la ecuación B.2 es el modelo de las mediciones el cual describe la relación entre las variables $\mathbf{x}(k_c)$ y $\mathbf{u}(k_c)$. En aplicaciones de FK (por ejemplo, control) los valores verdaderos de $\mathbf{o}(k_c)$ se asumen a ser conocidos. Sin embargo, en muchas situaciones solo se dispone de mediciones de $\mathbf{o}(k_c)$ contaminadas con ruido aun cuando se requieren los valores estimados de estas variables. En estos casos, se consideran a estas variables como variables de estado lo que conduce a un nuevo vector de variables de estado $\mathbf{x}^{+}(k_C) = \left[\mathbf{x}_1(k_C), \mathbf{x}_2(k_C), \ldots, \mathbf{o}_1(k_C), \mathbf{o}_2(k_C), \ldots\right]$. En consecuencia, $\mathbf{B}(k_c)$, es despreciada y el modelo de las ecuaciones (B.1) y (B.2) pasa a ser:

$$\mathbf{x}(k_C) = \mathbf{A}(k_C)\cdot\mathbf{x}(k_{C-1}) + \omega(k_{C-1}) \tag{B.3}$$

$$\mathbf{u}(k_C) = \mathbf{H}(k_C)\cdot\mathbf{x}(k_C) + \varepsilon(k_C) \tag{B.4}$$

La formulación anterior se utiliza en el caso de estudio considerado. También, se considera que los coeficientes de las matrices **A** y **H** no cambian con el tiempo. Luego, las matrices **x,** **A** y **H** quedan fijadas como sigue:

$$\mathbf{x} = \begin{bmatrix} q_0 \\ q \\ V \end{bmatrix}; \qquad \mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix}; \qquad \mathbf{H} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{B.5}$$

Asimismo, las matrices covarianza de $\varepsilon(k_c)$ (**R**) y $\omega(k_c)$ (**S**) se asumen a ser:

$$\mathbf{R} = \begin{bmatrix} 0.05 & 0 & 0 \\ 0 & 0.001 & 0 \\ 0 & 0 & 0.01 \end{bmatrix}; \qquad \mathbf{S} = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.01 & 0 \\ 0 & 0 & 10 \end{bmatrix}; \tag{B.6}$$

Tras la definición del modelo en su forma de espacios de estados, se aplica la reconciliación con FK según el siguiente algoritmo (Narasimhan et al, 2002):

- **Paso 1**: Se fijan los estimados iniciales de las matrices covarianzas de los errores **T** y de los estados del sistema **x**, como sigue:

$$\mathbf{T}(k_{0/0}) = \mathbf{T}(k_{c-1/c-1}) = \mathbf{S} \tag{B.7}$$

$$\mathbf{x}(k_0) = \mathbf{x}(k_{c-1/c-1}) = \mathbf{y}(k_0) \tag{B.8}$$

- **Paso 2**: Predicción de los estados y su matriz de covarianza, usando los estimados iniciales de $\mathbf{P}_{k|k-1}$ y $\mathbf{x}_k$,

$$\hat{\mathbf{x}}(k_{c/c-1}) = \mathbf{A}(k_c) \cdot \hat{\mathbf{x}}(k_{c-1/c-1}) \tag{B.9}$$

$$\mathbf{T}(k_{c/c-1}) = \mathbf{A}(k_c) \cdot \mathbf{T}(k_{c-1/c-1}) \cdot \mathbf{A}^T(k_c) + \mathbf{R} \tag{B.10}$$

- **Paso 3**: Se computa la matriz de ganancia del FK como sigue:

$$\mathbf{K}(k_c) = \mathbf{T}(k_{c/c-1}) \cdot \mathbf{H}^T \cdot \left(\mathbf{H} \cdot \mathbf{T}(k_{c/c-1}) \cdot \mathbf{H}^T + \mathbf{S} \right)^{-1} \tag{B.11}$$

- **Paso 4**: Con las nuevas observaciones disponibles $\mathbf{y}(k_c)$, se actualizan los estimados de los estados como sigue:

$$\hat{\mathbf{x}}(k_{c/c}) = \hat{\mathbf{x}}(k_{c/c-1}) + \mathbf{K}(k_c) \cdot \left(\mathbf{y}(k_c) - \mathbf{H} \cdot \hat{\mathbf{x}}(k_{c/c-1}) \right) \tag{B.12}$$

- **Paso 5**: Se actualiza la matriz covarianza de los estados como sigue:

$$\mathbf{T}(k_{c/c}) = \left(\mathbf{I} - \mathbf{K}(k_c) \cdot \mathbf{H} \right) \cdot \mathbf{T}(k_{c/c-1}) \tag{B.13}$$

- **Paso 6**: Se actualizan las predicciones de $\hat{\mathbf{x}}(k_{c/c-1})$, $\hat{\mathbf{T}}(k_{c/c-1})$ y $\mathbf{K}(k_c)$, mediante las ecuaciones (B.9), (B.10) y (B.11).

Los pasos 4, 5 y 6 se repiten cada vez que se dispone de nuevas mediciones.

Narasimhan (Narasimhan y Jordache, 2002) mostró que el procedimiento anterior es una aproximación a una RDD en la cual los estimados de los estados representan los valores de los estados y variables reconciliadas. Asimismo, mostró que si en la formulación anterior del modelo de espacios de estados discretos (ecuaciones B.3 y B.4) se considera el modelo libre de perturbaciones ($\omega(k_c)=0$ en cada instante de tiempo), la estrategia resultante se aproxima a una RD en estado estacionario.

C. Ejemplo del reactor continuo no isotérmico.

Los reactores químicos son las unidades más básicas e importantes a las que se enfrenta un ingeniero en la industria química. En la literatura es muy habitual encontrarse con modelos de reactores que se utilizan como casos de estudio para la aplicación y/o prueba de distintas teorías y metodologías de soporte a diversas tareas operacionales. En particular, los reactores de tipo tanque agitado y continuo, también conocido como *CSTR* (siglas inglesas de C*ontinuous Stirred Tank Reactor*), se han utilizado con mucha frecuencia para comparativas en trabajos de control de procesos (Luyben, 1990; Chen y Peng, 1999; Marlin, 2002; Martinsen *et al*., 2004; Gao y Budman, 2005; Wright y Kravaris, 2005), monitorización y diagnostico de fallos (Chen y McAvoy, 1998; Li y Wang, 1999; Juricek *et al*., 2001; Chen y Liao, 2002), análisis de tendencias (Sun *et al*., 2003; Dash *et al*., 2004), reconciliación de datos (Liebman *et al*., 1992; Romagnoli y Sánchez, 2000), etc. Dado el atractivo de los *CSTR* como escenarios de prueba para diversas estrategias, en esta tesis se utilizan modelos académicos de reactores *CSTR* para probar distintas propuestas.

C.1 Modelo del reactor.

Para las pruebas del capítulo 3 se escoge un *CSTR* utilizado en la literatura sobre métodos de reconciliación de datos en estado dinámico (Liebman *et al*., 1992; Romagnoli y Sánchez, 2000). El detalle del modelo asociado es como sigue:

El reactor opera en modo continuo. Dentro del mismo se produce una reacción exotérmica de primer orden que consiste en la descomposición de un reactante A para producir B. El proceso puede describirse por un conjunto de ecuaciones diferenciales y algebraicas (C.1 – C.4) que representan:

- El cambio en la cantidad de volumen *V* dentro del reactor (C.1).
- El cambio en la concentración *Ca* del reactante A (C.2).
- El cambio en la temperatura *T* del reactor (C.3).
- La constante de velocidad del reactor (C.4).

$$\frac{dV}{dt} = q_0 - q \tag{C.1}$$

$$\frac{dCa}{dt} = \frac{q_0}{V} Ca_0 - \left(\frac{q_0}{V} + K \right) Ca \tag{C.2}$$

$$\frac{dT}{dt} = \frac{1}{V}\left(T_0.q_0 - T.q_0\right) + \frac{-\alpha_d.\Delta H.Ca_r.K.Ca}{\rho.Cp.T_r} + \frac{-U.A_R}{\rho.Cp.V}.\left(T - T_C\right) \tag{C.3}$$

$$K = K_0.exp\left(\frac{-E_A}{T.T_r}\right) \tag{C.4}$$

Ca_0 y T_0 representan la concentración y la temperatura del flujo de alimentación al reactor, q_0 y q son los flujos de entrada y salida y K es la constante de velocidad del reactor que se calcula de acuerdo a la expresión de Arrhenius (ecuación C.4). El resto de parámetros queda definido en la tabla C1. Previo a las simulaciones con este modelo, la concentración y la temperatura se han de escalar utilizando una concentración nominal de referencia Ca_r y una temperatura nominal de referencia T_r (Romagnoli y Sánchez, 2000).

Tabla C1. Parámetros y constantes físicas del reactor *CSTR*.

Parámetro	Valor	Unidades	
q	10	$cm^3 \cdot s^{-1}$	Flujo de salida
V	1000	Cm^3	Volumen en el reactor
ΔH	-27000	$cal \cdot gmol^{-1}$	Calor de Reacción
ρ	0.001	Gcm^3	Densidad
C_p	1.0	$cal \cdot (gk)^{-1}$	Capacidad calorífica
U	$5.0x10^{-4}$	$cal \cdot (cm^2 sK)^{-1}$	Coeficiente de Transf. de Calor
A_R	10.0	Cm^2	Área de Transf. de Calor
T_C	340	K	Temperatura del refrigerante
T_0		K	Temperatura de la alimentación
T		K	Temperatura del tanque (reactor).
K_0	$7.86x10^{12}$	s^{-1}	Constante de Arrhenius
K		s^{-1}	Constante de velocidad del reactor
E_A	14090	K	Energía de Activación
α_d	1		Parámetro de desactivación del catalizador
Ca		$Gmol \cdot cm^{-3}$	Concentración en la alimentación
Ca_0		$Gmol \cdot cm^{-3}$	Concentración en el Reactor

C.2 Modelo ampliado del reactor

El modelo anterior se selecciona como primer caso de estudio para la comparación de las estrategias que se presentan en el capítulo 4. No obstante, para acercarse a las condiciones de los experimentos de los trabajos de la literatura donde aparecen algunas de las metodologías a probar, se añaden ciertas variaciones al modelo como:

- Considerar el balance de energía en la camisa del reactor.

$$\frac{dT_C}{dt} = q_{C0} \times (T_{C0} - T_C) + \frac{(U \cdot A_H) \cdot (T - T_C)}{\rho_C \cdot Cp_C \cdot V_C} \tag{C.5}$$

- Un controlador proporcional de nivel mediante el flujo de salida q.
- Otro controlador para la temperatura del reactor T mediante el flujo de enfriamiento en la chaqueta q_C.

Al redefinir el modelo mediante la adición de la ecuación anterior más los controladores, los valores de todos los parámetros se cambian a los valores de las diferentes constantes, de los valores iniciales y del estado estacionario que aparecen reportados en el libro de Luyben para el caso de un modelo *CSTR* (Luyben, 1990). En la figura C.1 se muestra el diagrama del reactor.

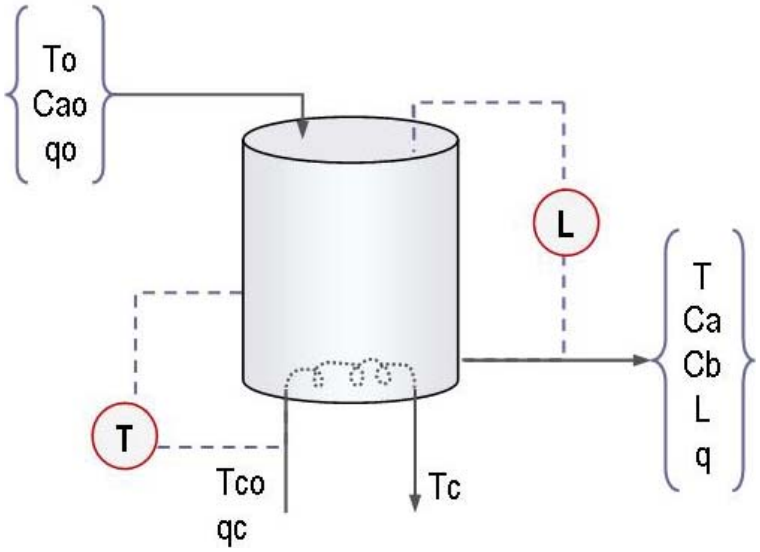

Figura C.1. Esquema del reactor continuo (*CSTR*).

C.3 Modelo afectado por ensuciamiento.

En el capítulo 6 se propone analizar la posible variación en las condiciones de operación de un proceso debidas a ensuciamiento. Como caso de estudio se toma de nuevo el *CSTR* de la sección anterior con la variante principal de la presencia de un efecto de ensuciamiento sobre el sistema de intercambio de calor dentro del tanque. Se prueban varios valores de estado estacionario tal que las dinámicas asociadas al efecto de ensuciamiento simulado sean lo suficientemente ilustrativas para el análisis a realizar. En la tabla C2 se resumen los diferentes valores del estado estacionario que se han probado.

Tabla C2. Distintas condiciones de trabajo para el *CSTR* afectado por ensuciamiento.

Variable	Condiciones de Operación (CO_i)				
	CO_1	CO_2	CO_3	CO_4	CO_5
T_0 (k)	323	323	323	323	323
T(k)	340.6	353.1	363.5	369.3	374.4
Ca_0(mol/lt)	1.9	1.9	1.8	1.9	1.8
Ca (mol/ lt)	1.53	1.21	0.85	0.6934	0.568
Tc(k)	335	342	348	351	354

D. Tratamiento de Outliers con ACP

Esta estrategia aparece en varios trabajos de la literatura relacionada con supervisión (Fernández Pierna *et al*., 2002; Chiang *et al*., 2003). Consiste en el siguiente algoritmo:

1. Se toma la matriz de datos del proceso **Y** y se obtiene el correspondiente ACP según se explica en la sección 4.1.

2. Se calculan los *SPE* y T^2 y sus correspondientes límites $SPE_{\lim}$ y $T^2_{\lim}$.

3. Se comparan los *SPE* y T^2 con los límites calculados. Todas las observaciones que superen los límites calculados se consideran como *outliers*. Y se etiquetan como n_{out}, mientras que las observaciones que no son *outliers* se etiquetan como n_{ACP}.

4. Se recalcula el ACP pero solo utilizando los datos etiquetados como n_{ACP}.

5. Se repiten 2 y 3 tal que se ajustan los valores de n_{out} y n_{ACP}.

6. El procedimiento anterior se repite continuamente entre los pasos 2 y 5. Se detendrá si se cumple alguno de los siguientes criterios:
   $n_{out} = 0$, tras una iteración.
   $n_{out} < n/2$, donde n representa el número de muestras.

En la propuesta de la estrategia anterior se parte de que en un conjunto de datos el número de *outliers* puede ser muy alto (casi la mitad de los datos normales). En tales casos, si se aplica el procedimiento anterior solo hasta el paso 3, únicamente se detectaran los *outliers* más desviados. Por ello proponen la iteración continua. No obstante, en muchos casos reales el número de *outliers* es significativamente menor al 5 % del total de los datos (Pearson, 2002), situaciones para las cuales algunos autores (Chiang et al., 2003) reconocen que el procedimiento aplicado entre los pasos 1, 2 y 3, ya es suficiente. Luego, en el trabajo del capítulo 4, la estrategia anterior se aplica solo hasta el paso 3. Adicionalmente, el cálculo de los límites se basa en la distribución empírica de los correspondientes estadísticos (*SPE* y T^2) tal y como se propone en la sección 4.3.1.1.

BIBLIOGRAFÍA

Abu-el-zeet, Z. H., P. D. Roberts y V. M. Becerra. Enhancing Model Predictive Control Using Dynamic Data Reconciliation. *AIChE Journal,* 48(2), pp.324-333. (2002).

Addison, P. S. *The Illustrated Wavelet Transform Handbook: Applications in Science, Engineering, Medicine and Finance.* Institute of Physics Publishing, Bristol, UK.(2002).

Albuquerque, J. S. y L. T. Biegler. Data Reconciliation and Gross-Error Detection for Dynamic Systems. *AIChE Journal,* 42(10), pp.2841. (1996).

Amand, T., G. Heyen y B. Kalitventzeff. Synergy between data reconciliation and principal component analysis: Plant monitoring and fault detection. *Computers & Chemical Engineering,* 25, pp.501-507. (2001).

Amiri, M. Fuzzy C-Means *Clustering*. Comp. Eng. Dept., Sharif University of Technology, Tehran, Iran.(2003).
http://ce.sharif.edu/~m_amiri/download.html#Y_FCMC.

Apte, C., B. Liu, E. P. D. Pednault y P. Smyth. Business Applications of Data Mining. *Communications of the ACM,* 45(8), pp.49-53. (2002).

Armstrong, J. S. Combining Forecasts: The End of the Beginning or the Beginning of the End. *International Journal of Forecasting,* 5, pp.585-588. (1989).

Babuska, R., P. J. van der Veen y U. Kaymak. Improved Covariance Estimation for Gustafson-Kessel Clustering. *IEEE International Conference on Fuzzy Systems,* pp.1081-1085. (2002).

Bagajewicz, M. J. Data Reconciliation and Instrumentation Upgrade. Overview and Challenges, in Fourth International Conference on Foundations of Computer-Aided Process Operations FOCAPO'03 Coral Springs, Florida. (2003).

Bagajewicz, M. J. y Q. Jiang. Integral Comparison of steady state and integral dynamic data reconciliation. *Computers & Chemical Engineering,* 24, pp.2367-2383. (2000).

Bagajewicz, M. J. y Q. Jiang. Integral Approach to Plant Linear Dynamic Reconciliation. *AIChE Journal,* 43(10), pp.2546. (1997).

Bakhtazad, A., A. Palazoglu y J. A. Romagnoli. Process Data De-noising Using Wavelet Transform. *Intelligent Data Analysis,*(267), pp.285. (1999).

Bakshi, B. R. Multiscale analysis and modeling using wavelets. *Journal of Chemometrics,* 13(3-4), pp.415-434. (1999).

Bakshi, B. R. Multiscale PCA with application to multivariate statistical process monitoring. *AIChE Journal,* 44(7), pp.1596-1610. (1998).

Bakshi, B. R., P. Bansal y M. N. Nounou. Multiscale rectification of random errors without fundamental process models. *Computers & Chemical Engineering,* 21, pp.S1167-S1172. (1997).

Bakshi, B. R. y G. Stephanopoulos. Representation of Process Trends-IV. Induction of real-time patterns from operating data for diagnosis and supervisory control. *Computers & Chemical Engineering,* 18(4), pp.303-336. (1994a).

Bakshi, B. R. y G. Stephanopoulos. Representation of Process Trends-III. Multiscale Extraction of Trends from Process Data. *Computers & Chemical Engineering,* 18(4), pp.267-302. (1994b).

Balasko, B., Abonyi, J., y Feil, B. Fuzzy Clustering and Data Analysis Toolbox For Use with Matlab. Dept. Process Eng., University of Veszprem. Veszprem, Hungary.(2004). www.fmt.vein.hu/softcomp

Barbosa, V. P., M. R. M. Wolf y R. Maciel Fo. Development of data reconciliation for dynamic nonlinear system: application the polymerization reactor. *Computers & Chemical Engineering,* 24, pp.501-506. (2000).

Barni, M., A. Mecocci y A. Mecocci. Comments on a Possibilistic Approach to Clustering. *IEEE Trans. Fuzzy Systems,* 4(3), pp.393-396. (1996).

Bates, J. M. y C. W. J. Granger. The Combination of Forecasts. *Operational Research Quarterly,* 20, pp.319-325. (1969).

Belanger, P. W. y W. L. Luyben. Inventory control in processes with recycle. *Industrial & Engineering Chemistry Research,* 36(1), pp.706-716. (1997).

Benqlilou, C., Bagajewicz, M. J., Espuña, A., and Puigjaner, L. A Comparative Study of Linear Dynamic Data Reconciliation Techniques., in 9th Mediterranean Congress of Chemical Engineering Barcelona, España. pp.P-8-31. (2002).

Benqlilou, C., Tona, R. V., Espuña, A., and Puigjaner, L. On-Line Application of Dynamic Data Reconciliation, in 4th Conference on Process Integration, Modelling and Optimisation for Energy Saving and Pollution Reduction (Klemes, J. ed.). Milano, Italy. pp.403-406. (2001).

Bezdek, J. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Publishing Corporation, New York, N.Y.(1981).

Chatfield, C. *Time Series Forecasting.* Chapman & Hall., New York.(2002).

Chen, C. y S. T. Peng. Intelligent process control using neural fuzzy techniques. *Journal of Process Control,* 9(6), pp.493-503. (1999).

Chen, G. y T. McAvoy. Predictive On-Line Monitoring of Continuous Processes. *Journal of Process Control,* 8(5-6), pp.409-420. (1998).

Chen, J. H., D. S. Hill y J. L. Liu. Pocess monitoring usin distance-based adaptive reasonance theory. *Industrial & Engineering Chemistry Research,* 41, pp.2465-2479. (2002).

Chen, J. H. y C. M. Liao. Dynamic process fault monitoring based on neural network and PCA. *Journal of Process Control,* 12(2), pp.277-289. (2002).

Chen, J. H., C. M. Liao, F. R. J. Lin y M. J. Lu. Principle component analysis based control charts with memory effect for process monitoring. *Industrial & Engineering Chemistry Research,* 40(6), pp.1516-1527. (2001).

Chen, J. H. y J. L. Liu. Mixture principal component analysis models for process monitoring. *Industrial & Engineering Chemistry Research,* 38(4), pp.1478-1488. (1999).

Cheung, J. y G. Stephanopoulos. Representation of process trends-part II. The Problem Of Scale And Qualitative Scaling. *Computers & Chemical Engineering,* 14(4/5), pp.511-526. (1992a).

Cheung, J. y G. Stephanopoulos. Representation of process trends-part I. A formal representation framework. *Computers & Chemical Engineering,* 14(4/5), pp.495-510. (1992b).

Chiang, L. H., R. J. Pell y M. B. Seasholtz. Exploring process data with the use of robust outlier detection algorithms. *Journal of Process Control,* 13(5), pp.437-449. (2003).

Chintalapudi, K. K. y M. Kam. The credibilistic Fuzzy C-Means Algorithm. *IEEE Int. Conf. Syst. , Man, Cybernetics,* 2, pp.2034-2039. (1998).

Chiu, S. L. Fuzzy Model Identification Based on Cluster Estimation. *J. Intell. Fuzzy Systems,* 2, pp.267-278. (1994).

Cho, S. B. y J. H. Kim. Combining multiple neural networks by fuzzy integral for robust classification. *EEE Trans. on Syst. Man and Cybernetics,* 25(2), pp.380-384. (1995).

Choi, S. W., C. K. Yoo y I. B. Lee. Overall statistical monitoring of static and dynamic patterns. *Industrial & Engineering Chemistry Research,* 42(1), pp.108-117. (2003).

Cios, K. J., W. Pedrycz y R. W. Swiniarski. *Data mining methods for knowledge discovery.* Kluwer Academic., Boston, MA.(1998).

Craigmile, P. F. and B. D. Percival. Wavelet-Based Trend Detection and Estimation., *In Encyclopedia of Environmetrics* (*El-Shaarawi, A. H. y W. W. Piegorsch* eds.). John Wiley & Sons, Chichester, England. pp. 2334-2338. (2002).

Dash, S., M. R. Maurya, V. Venkatasubramanian y R. Rengaswamy. A novel interval-halving framework for automated identification of process trends. *AIChE Journal,* 50(1), pp.149-162. (2004).

Dash, S., R. Rengaswamy y V. Venkatasubramanian. Fuzzy-Logic based trend classification for fault diagnosis. *Computers & Chemical Engineering,* 27(3), pp.347-362. (2003).

Davis, J. F., M. Piovoso, M. Piovoso y B. R. Bakshi. Process Data Analysis and Interpretation. *Advances in Chemical Engineering,* (2000).

DeCicco, J. Simulation o fan Industrial Polyvinyl Acetate CSTR and Semi-Batch Reactor utilizing MATLAB and SIMULINK: Version 1.0. Department of Chemical and

Environmental Engineering, Illinois Institute of Technology.(1998). http://www.chee.iit.edu/~cinar.

Delurgio, S. A. *Forecasting: Principles and Applications.* Irwin/McGraw-Hill, New York.(1998).

Demirli, K., S. X. Cheng y P. Muthukumaran. Subtractive Clustering based modeling of job sequencing with parametric search. *Fuzzy Sets and Systems.,* 137(2), pp.235-270. (2003).

Donoho, D. L. Wavelet Shrinkage and W.V.D. - A Ten-Minute Tour., in International Conference on Wavelets and Applications Toulouse, France. (1992).

Donoho, D. L. y I. M. Johnstone. Adapting to Unknown Smoothness via Wavelet Shrinkage. *J. American Statistical Association,* 90(432), pp.1200-1224. (1995).

Donoho, D. L. y I. M. Johnstone. Ideal spatial adaption via wavelet shrinkage. *Biometrika,* 81, pp.425-455. (1994).

Downs, J. y E. Vogel. A Plant-Wide Industrial-Process Control Problem. *Computers & Chemical Engineering,* 17(3), pp.245-255. (1993).

Doymaz, F., A. Bakhtazad, J. A. Romagnoli y A. Palazoglu. Wavelet-based robust filtering of process data. *Computers & Chemical Engineering,* 25(11-12), pp.1549-1559. (2001).

Duda, R. O., P. E. Hart y D. G. Stork. *Pattern Classification.*, 2 edn. John Wiley & Sons, Inc., N.Y.(2001).

Edgar, T., D. Himmelblau y L. Lasdon. *Optimization of Chemical Processes.*, 2 edn. Chemical Engineering Series, McGraw-Hill International, Singapore.(2001).

Epstein, N. Optimum evaporator cycle with scale formation. *Canadian Journal of Chemical Engineering,* 57, pp.659. (1979).

Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth y R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press/MIT, Menlo Park, California.(1996).

Fernández Pierna, J. A., F. Wahl, O. E. de Noord y D. L. Massart. Methods for outlier detection in prediction. *Computers & Chemical Engineering,* 63(1), pp.27-39. (2002).

Ferrer, A. Aplicación del Control Estadístico Multivariable. *Automática e Instrumentación,*(326), pp.62-72. (2002).

Flehmig, F., R. Von Watzdorf y W. Marquardt. Identification of trends in process measurements using the wavelet transform. *Computers & Chemical Engineering,* 22(Suppl.), pp.S491-S496. (1998).

Gao, J. y H. M. Budman. Design of robust gain-scheduled PI controllers for nonlinear processes. *Journal of Process Control,* 15(7), pp.807-817. (2005).

Gertler, J. *Fault Detection and Diagnosis in Engineering Systems.* Marcel Dekker, New York.(1998).

Ghael, S., A. M. Sayeed y R. G. Baraniuk. Improved wavelet de-noising via filtering in additive noise. *IEEE Transactions of Automated Control,* AC-13, pp.646-655. (1997).

Goebel, M. y L. Gruenwald. A Survey of Data Mining and Knowledge Discovery Software Tools. *SIGKDD Explorations,* 2(1), pp.20-33. (1999).

Grossman, I. Challenges in the new Millennium: Product Discovery and Design, Enterprise and Supply Chain Optimization, Global Life Cycle Assessment, in Process System Engineering - PSE 2003 (Chen, B. y Westerberg, A. W. eds.). Beijing, China. pp.28-47. (2003).

Gustafson, D. E. and Kessel, W. C. Fuzzy Clustering with a Fuzzy Covariance Matrix., in proc. IEEE CDC San Diego, CA, USA. pp.761-766. (1979).

Halkidi, M., Y. Batistakis y M. Vazirgiannis. On Clustering Validation Techniques. *J. Intell. Infor. Syst.,* 17(2/3), pp.107-145. (2001).

Han, J. y M. Kamber. *Data mining: concepts and techniques.* Morgan Kaufmann, San Francisco, CA, USA.(2001).

Hand, D., H. Mannila y P. Smyth. *Principles of Data Mining.* The MIT Press, Cambridge, Massachusetts.(2001).

Harmon, L. y S. Schlosser. CPI Plants Go Data Mining. *Chemical Engineering,* 106(4), pp.96. (1999).

Hwang, D. H. y C. Han. Real-time monitoring for a process with multiple operating modes. *Control Engineering Practice.,* 7, pp.891-902. (1999).

Jackson, J. E. *A User's Guide to Principal Components*. John Wiley&Sons, New York.(1991).

Jain, A. K., M. N. Murty y P. J. Flynn. Data Clustering: A Review. *ACM Computing Surveys,* 31(3), pp.264-323. (1999).

Jain, V. y I. Grossman. Cyclic scheduling of continuous parallel process units with decaying performance. *AIChE Journal,* 44, pp.1623. (1998).

Jang, S. S., B. Josepth y H. Muhai. Comparison of two approaches to on-line parameter and state estimation problem of non-linear systems. *Industrial & Engineering Chemistry Research,* 25, pp.809. (1996).

Jiang, T. W., B. Z. Chen, X. R. He y P. Stuart. Application of steady-state detection method based on wavelet transform. *Computers & Chemical Engineering,* 27(4), pp.569-578. (2003).

Johannesmeyer, M. C., A. Singhal y D. E. Seborg. Pattern matching in historical data. *AIChE Journal,* 48(9), pp.2022-2038. (2002).

Juricek, B., D. E. Seborg y W. Larrimore. Predictive Monitoring for Abnormal Situation Management. *Journal of Process Control,* 11(2), pp.111-128. (2001).

Kano, M., K. Nagao, S. Hasebe, I. Hashimoto, H. Ohno, R. Strauss y B. R. Bakshi. Comparison of multivariate statistical process monitoring methods with applications to the Eastman challenge problem. *Computers & Chemical Engineering,* 26(2), pp.161-174. (2002).

Kivikunnas, S. Overview of Process Trend Analysis Methods and Applications, in Proceedings of Workshop on Applications in Chemical and Biochemical Industry Aachen, Germany. (1999).

Köhler, T. y Lorenz, D. A comparison of denoising methods for one dimensional time series. Technical Report DFG SPP 1114, Preprint series of the DFG priority program 1114.(2004). http://www.math.uni-bremen.de/zetem/DFG-Schwerpunkt/preprints/prep074.pdf

Kosanovich, K. A. y M. J. Piovoso. PCA of Wavelet Transformed Process Data for Monitoring. *Intelligent Data Analysis,* 1(1-4), pp.85-99. (1997).

Kourti, T. Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *Journal of Chemometrics,* 17(1), pp.93-109. (2003).

Kramer, M. A. Autoassociative neural networks. *Computers & Chemical Engineering,* 16, pp.313-328. (1994).

Kresta, J., J. F. MacGregor y T. Marlin. Multivariate Statistical Monitoring of Process Operating Performance. *Canadian Journal of Chemical Engineering,* 69, pp.35-47. (1991).

Krishnapuran, R. y J. M. Keller. A Possibilistic Approach to Clustering. *IEEE Trans. Fuzzy Systems,* 1(2), pp.98-110. (1993).

Krzanowski, W. J. Between-groups comparison of principal components. *J. American Statistical Association,* 74(367), pp.703-707. (1979).

Ku, W., R. H. Storer y C. Georgakis. Disturbance detection and isolation by dynamic principal components analysis. *Chemometrics and Intelligence Laboratory System,* 30, pp.179. (1995).

Li, R. F. y X. Z. Wang. Qualitative/quantitative simulation of process temporal behavior using clustered fuzzy digraphs. *AIChE Journal,* 47(4), pp.906-919. (2001).

Li, R. F. y X. Z. Wang. Combining Conceptual Clustering and Principal Component Analysis for State Space Based Process Monitoring. *Industrial & Engineering Chemistry Research,* 38, pp.4345-4358. (1999).

Li, T., L. Qi, S. Zhu y M. Ogihara. A Survey on Wavelet Applications in Data Mining. *SIGKDD Explorations,* 4(2), pp.49-97. (2002).

Li, W. H., H. H. Yue, S. Valle-Cervantes y S. J. Qin. Recursive PCA for adaptive process monitoring. *Journal of Process Control,* 10(5), pp.471-486. (2000).

Liebman, M. J., T. F. Edgar y L. S. Ladson. Efficient data reconciliation and estimation for dynamic processes using nonlinear programming techniques. *Computers & Chemical Engineering,* 16(10/11), pp.963. (1992).

Lu, N. Y., F. L. Wang y F. R. Gao. Combination method of principal component and wavelet analysis for multivariate process monitoring and fault diagnosis. *Industrial & Engineering Chemistry Research,* 42(18), pp.4198-4207. (2003).

Luyben, W. *Process Modeling, Simulation and Control for Chemical Engineers.* McGraw-Hill, Inc.,(1990).

MacGregor, J. F. Data-Based Latent Variable Methods for Process Analysis, Monitoring and Control., in European Symposium on Computer Aided Process Engineering - 14 (Barbosa Póvoa, A. y Matos, H. eds.). Elsevier, Lisbon, Portugal. (2004).

Makridakis, S., S. Wheelwrigth y R. Hyndman. *Forecasting: Methods and Applications*, 3 edn. Wiley, New York.(1998).

Mallat, S. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence,* 11, pp.674-693. (1989).

Marlin, T. *Process Control, Designing Processes and Control Systems for Dynamic Performance.*, 2 edn. McGraw Hill, New York.(2002).

Martin, E. B., A. J. Morris y S. Lane. Monitoring process manufacturing performance. *IEEE Control Systems Magazine,*(October), pp.26-39. (2002).

Martinsen, F., L. T. Biegler y B. A. Foss. A new optimization algorithm with application to nonlinear MPC. *Journal of Process Control,* 14(8), pp.853-865. (2004).

Mingfang, K., C. Bingzhen y L. Bo. An integral approach to dynamic data rectification. *Computers & Chemical Engineering,* 20, pp.-749. (2000).

Misiti, M., Y. Misiti, G. Oppenheim y J. M. Poggi. *Wavelet Toolbox. User's Guide.*, 3 edn. Natick, MA (USA).(2004).

Musulin, E., M. J. Bagajewicz, J. M. Nougues y L. Puigjaner. Instrumentation Design and Upgrade for Principal Components Analysis Monitoring. *Industrial & Engineering Chemistry Research,* 43, pp.2150-2159. (2004).

Musulin, E., Tona, R. V., Ruiz, D., Espuña, A., and Puigjaner, L. Improving Principal Components Analysis Approaches for Monitoring Systems, in 52 Canadian Chemical Enginneering Conference Vancouver, Canada. (2002).

Musulin, E., I. Yélamos y L. Puigjaner. Integration of Principal Component Analysis and Fuzzy Logic Systems for Comprehensive Process Fault Detection and Diagnosis. *Industrial & Engineering Chemistry Research,* 45(5), pp.1739-1750. (2006).

Næs, T. y B. H. Mevik. The flexibility of fuzzy clustering illustrated by examples. *Journal of Chemometrics,* 13(3-4), pp.435-444. (1999).

Narasimhan, S. y C. Jordache. *Data Reconciliation and Gross Error Detection, an intelligent use of process data.* Gulf Publishing Company, Houston, TX (USA).(2000).

Nomikos, P. y J. F. MacGregor. Multivariate SPC charts for Monitoring Batch Processes. *Technometrics,* 37(1), pp.41-59. (1995).

Nomikos, P. y J. F. MacGregor. Monitoring Batch Processes Using Multiway Principal Component Analysis. *AIChE Journal,* 40(8), pp.1361-1375. (1994).

Nounou, M. N. y B. R. Bakshi. On-line multiscale filtering of random and gross errors without process models. *AIChE Journal,* 45(5), pp.1041-1058. (1999).

Optimization-Toolbox. *Optimization Toolbox for Matlab. User's Guide.*, 3 edn. Natick, MA (USA).(2003).

Paiva, R. P., Dourado, A., and Duarte, B. Applying Subtractive Clustering for Neuro-Fuzzy Modelling of a Bleaching Plant, in Proceeding of the 5th European Control Conference - ECC'99 Karlsruhe, Germany. (1999).

Pal, N. R., Pal, K., and Bezdek, J. A mixed c-means clustering model, in Proceedigs of the IEEE Int. Conf. on Fuzzy Systems Spain. pp.11-21. (1997).

Pearson, R. K. Outliers in Process Modelling and Identification. *IEEE Trans. Cont. Syst. Tech.,* 10(1), pp.55-63. (2002).

Piovoso, M. J., K. A. Kosanovich y J. P. Yuk. Process data chemometrics. *IEEE Trans. Instrumentation &. Measurements,* 41, pp.262-268. (1992).

Qin, S. J. Statistical process monitoring: basics and beyond. *Journal of Chemometrics,* 17(8-9), pp.480-502. (2003).

Quantrille, T. y Y. lin. *Artificial Intelligent in Chemical Engineering.* Academic Press, San Diego, CA.(1991).

Rinta-Runsala, E. Off-Line Analysis and Prototyping of Paper Machine Drive Monitoring System. Research report TTE1-2000-27, MODUS-Project., VTT Information Technology, (2001).

Robertson, D. J., J. H. Lee y J. B. Rawlings. A moving Horizon-based Approach for Least-Squares Estimation. *AIChE Journal,* 42(8), pp.2209. (1996).

Rollins, D. K. y S. Devanathan. Unbiased Estimation in Dynamic Data Reconciliation. *AIChE Journal,* 39(8), pp.1330. (1993).

Romagnoli, J. A. y M. C. Sánchez. *Data processing and reconciliation for chemical process operations. Process systems engineering series (Vol. 2).* Academic Press., San Diego, CA(USA).(2000).

Rosen, C. A. *Chemometric Approach to Process Monitoring and Control with Applications to Wastewater Treatment Operation.* PhD Dissertation. Dept. Industrial Electrical Eng. and Automation, Lund University., Lund, Sweden.(2001).

Rowe, A. C. H. y P. Abbott. Daubechies Wavelets and Mathematica. *Computer in Physics,* 9(6), pp.635-648. (1995).

Roy, M., V. R. Kumar, B. D. Kulkarni, J. Sanderson, M. Rhodes y M. vander Stappen. Denoising algorithm using wavelet transform. *AIChE Journal,* 45(11), pp.2461-2466. (1999).

Ruiz, D., J. M. Nougues, Z. Calderon, A. Espuna y L. Puigjaner. Neural network based framework for fault diagnosis in batch chemical plants. *Computers & Chemical Engineering,* 24(2-7), pp.777-784. (2000).

Sanmartí, E., A. Espuña y L. Puigjaner. Batch production and preventive scheduling under equipment failure uncertainty. *Computers & Chemical Engineering,* 21, pp.1157. (1997).

Sebzalli, Y. M. y X. Z. Wang. Knowledge discovery from process operational data using PCA and fuzzy clustering. *Eng. App. Artificial Intelligence,* 14, pp.607-616. (2001).

Sequeira, S. E. *An Evolutive Strategy for On-Line Optimization of Continuous Chemical Processes.* PhD dissertation, Chem. Eng. Dept., Universitat Politècnica de Catalunya, Barcelona, Spain.(2003).

Sequeira, S. E., M. Graells y L. Puigjaner. Off-line and on-line approach for optimal maintenance management of continuous parallel processes with decreasing performance. *Industrial & Engineering Chemistry Research,* 42, pp.176. (2003).

Singhal, A. y D. E. Seborg. Pattern Matching in Historical Batch Data Using PCA. *IEEE Control Systems Magazine,*(October), pp.53-63. (2002a).

Singhal, A. and Seborg, D. E. Clustering of Multivariate Time-Series Data., in Proceedings of the American Control Conference. Anchorage, AK. pp.3931-3936. (2002b).

SP95. *ANSI/ISA-S95.01-1999. Enterprise Control System Integration, part I: Models and Terminology*, 13 edn. International Society for Measurement and Control,(2000).

Srinivasan, R. y M. Qian. Offline Temporal Signal Comparison Using Singular Points Augmented Time Warping. *Industrial & Engineering Chemistry Research,* 44(13), pp.4697-4716. (2005).

Srinivasan, R., P. Viswanathan, H. Vedam y A. Nochur. A Framework for Managing Transitions in Chemical Plants. *Computers & Chemical Engineering,* 29(2), pp.305-322. (2005).

Srinivasan, R., C. Wang, W. K. Ho y K. W. Lim. Dynamic Principal Components Analysis Based Methodology for Clustering Process States in Agile Chemical Plants. *Industrial & Engineering Chemistry Research,* 43, pp.2123-2139. (2004).

Stephanopoulos, G. y C. Han. Intelligent systems in process engineering: A review. *Computers & Chemical Engineering,* 20(6-7), pp.743-791. (1996).

Stockill, D. Decision Confidence - Handling Uncertainty through the Plant Life Cycle using Statistics and Data Mining., in European Symposium on Computer Aided Process Engineering - ESCAPE 12 (Grievink, J. y Schjindel, J. eds.). Elsevier, Amsterdam, Netherlands. pp.70-77. (2002).

Sun, W., A. Palazoglu y J. A. Romagnoli. Detecting abnormal process trends by wavelet-domain hidden Markov models. *AIChE Journal,* 49(1), pp.140-150. (2003).

Taswell, C. The what, how, and why of wavelet shrinkage denoising. *IEEE Computing in Science & Engineering,* 2(3), pp.12-19. (2000).

Teppola, P. y P. Minkkinen. Possibilistic and fuzzy C-means clustering for process monitoring in an activated sludge waste-water treatment plant. *Journal of Chemometrics,* 13(3-4), pp.445-459. (1999).

Teppola, P., S. Mujunen y P. Minkkinen. Adaptive Fuzzy C-means Clustering in Process Monitoring. *Chemometrics and Intelligence Laboratory System,* 45, pp.23-38. (1999).

Terui, N. y H. K. Van Dijk. Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting,* 18, pp.421-438. (2002).

Teymour, F. y W. H. Ray. The Dynamic Behavior of Continuous Polymerization Reactors - VI. Complex Dynamics in Full-Scale Reactors. *Chemical Engineering Science,* 47, pp.4133-4140. (1992a).

Teymour, F. y W. H. Ray. The Dynamic Behavior of Continuous Polymerization Reactors - V. Experimental Investigation of Limit-Cycle Behavior for Vinyl Acetate Polymerization. *Chemical Engineering Science,* 47, pp.4121-4132. (1992b).

Tjoa, I. B. y L. T. Biegler. Reduced Successive Quadratic Programming Strategy for error-in-variables Estimation. *Computers & Chemical Engineering,* 16(6), pp.523. (1992).

Tona, R. V., Espuña, A., and Puigjaner, L. A Historical Data Based Methodology to identify longer variations and optimal process operating conditions, in 51st Canadian Chemical Engineering Conference Halifax, Canada. pp.375- (2001).

Trygg, J. y S. Wold. PLS regression on wavelet compressed NIR spectra. *Chemometrics and Intelligence Laboratory System,* 42, pp.209. (1998).

Vedam, H. y V. Venkatasubramanian. PCA-SDG based process monitoring and fault diagnosis. *Control Engineering Practice.,* 7, pp.903-917. (1999).

Venkatasubramanian, V., R. Rengaswamy y S. N. Kavuri. A review of process fault detection and diagnosis Part II: Quantitative model and search strategies. *Computers & Chemical Engineering,* 27(3), pp.313-326. (2003a).

Venkatasubramanian, V., R. Rengaswamy, S. N. Kavuri y K. Yin. A review of process fault detection and diagnosis Part III: Process history based methods. *Computers & Chemical Engineering,* 27(3), pp.327-346. (2003b).

Venkatasubramanian, V., R. Rengaswamy, K. Yin y S. N. Kavuri. A review of process fault detection and diagnosis Part I: Quantitative model-based methods. *Computers & Chemical Engineering,* 27(3), pp.293-311. (2003c).

Wachs, A. y D. R. Lewin. Improved PCA methods for process disturbance and failure identification. *AIChE Journal,* 45(8), pp.1688-1700. (1999).

Wang, X. Z. Knowledge Discovery through Mining Process Operational Data., *In Application of Neural Networks and other Learning Technologies, in Process Engineering* (*Mujtaba, I. M. y M. A. Hussain* eds.). Imperial College Press, London. pp. 287-327. (2001).

White, D. C. Creating the smart plant. *Hydrocarbon processing,* 82(10), pp.41-50. (2003).

Whiteley, J. R., J. F. Davis, A. Mehrota y S. C. Ahalt. Observations and Problems Applying ART2 for Dynamic Sensor Pattern Interpretation. *IEEE Trans. Systems, Man, and Cybernetics. Part A: Systems and Humans.,* 26(4), pp.423-437. (1996).

Wise, B. M. y N. B. Gallagher. The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control,* 6(6), pp.329-348. (1996).

Wold, S. Exponentially Weigthed Moving Principal Component Analysis and Projection to Latent Structures. *Chemometrics and Intelligence Laboratory System,* 23, pp.149. (1994).

Wold, S., A. Berglund y N. Kettaneh. New and old trends in chemometrics. How to deal with the increasing data volumes in R&D&P (research, development and production) - with examples from pharmaceutical research and process modeling. *Journal of Chemometrics,* 16(8-10), pp.377-386. (2002).

Wright, R. A. y C. Kravaris. Two-degree-of-freedom output feedback controllers for nonlinear processes. *Chemical Engineering Science,* 60(15), pp.4323-4336. (2005).

Wu, X., P. S. Yu, G. Piatetsky-Shapiro, N. Cercone, T. Y. Lin, R. Kotagiri y B. W. Wah. Data Mining: How Research Meets Practical Development? *Knowledge and Information Systems,* 5, pp.248-261. (2003).

Yager, R. R. y D. P. Filev. Approximate Clustering via the Mountain Method. *IEEE Transactions on Systems, Man, And Cybernetics,* 24(8), pp.1279-1284. (1994).

Yamashita, Y. Supervised learning for the analysis of process operational data. *Computers & Chemical Engineering,* 24(2-7), pp.471-474. (2000).

Yang, Y. Combining forecasting procedures: some theoretical results. *Econometric Theory,* 20, pp.176-222. (2004).

Yoo, C. K., P. A. Vanrolleghem y I. B. Lee. Nonlinear modelling and adaptive monitoring with fuzzy and multivariate statistical methods in biological wastewater treatment plants. *Journal ofBiotechnology,* 105, pp.135-163. (2003).

Yoon, S. y J. F. MacGregor. Principal-Component Analysis of Multiscale Data for Process Monitoring and Fault Diagnosis. *AIChE Journal,* 50(11), pp.2891-2903. (2004).

Yoon, S. y J. F. MacGregor. Statistical and causal model-based approaches to fault detection and isolation. *AIChE Journal,* 46(9), pp.1813-1824. (2000).

Zabala, A. Data Mining: Convirtiendo datos en información (parte 1). *Solo Programadores,* pp.26-36. (2003).

Zhong, W. y J. Yu. Improve nonlinear Soft sensing modeling by combining multiple models. *Hydrocarbon processing,* 79(4), pp.108-112. (2000).

¿Cuantas hojas impresas ha costado esta tesis?

1746 hojas impresas al final, otros varios cientos el mes del deposito y varios miles desde el día inicial.

¿Cuantas horas ha costado este trabajo?

5-7 horas diarias más de pensar, corregir y de escribir que de estudiar, experimentar y analizar durante todos los lunes a viernes del año 2006 contados desde marzo hasta septiembre… más 7-10 horas diarias de estudiar, pensar, dudar, experimentar, analizar, volver a dudar, calcular y escribir, desanimarse y animarse, de caer (equivocarse) y levantarse durante todos los lunes a sábados de 24 estaciones continuas a los pies del tibidabo. Centenares de horas dispersas de discusiones, correcciones, sugerencias y de esperas de parte de Antonio y Puigjaner o de Pugjaner y Antonio.

Las horas también necesarias y dispersas de los amigos y conocidos de la universidad y de los amigos del paseo San Juan; de esta tierra y de otras tierras; de pocos veranos o muchas primaveras.

¿Cuánto material se ha usado en esta tesis?

Dejando a un lado las hojas impresas con mis pocas ideas propias y las impresas con muchas y diversas ideas ajenas…

Dejando al lado las comidas y la ropa, los pasajes, medicinas, champúes y menudencias, la pensión y las llamadas a mis padres en otras tierras, y todo, todo, todo, todo, todo,… Todo aquello que ahora no cuenta…

En la tesis se han gastado las fundas de varias sillas, parte del tiempo de vida de argent, ferro, lantani, tantal, clor y algunos ordenadores sin nombre de elemento. Varias horas de calefacción y también de ventilación artificial. Los ahorros de un año de trabajo y el sueldo de mis padres de 2 meses enteros, el dinero de 4 años de beca que concedió la Generalitat y varios meses de dinero que este último año y de la UPC recibí; más de un centenar de libros que me encontré por allí: por mates, por industriales, y por toda la UPC, por la UB, por la UAB, y por la Pompeu también, el Internet en el que reconozco que muchas horas perdí, el teléfono que use en bastantes ocasiones y los softwares que me descargue de la intranet.

¿Y Cual es el aporte de esta tesis?

De entrada lo que se ha escrito, con sus virtudes y deficiencias. 5 trabajos que se muestran aquí. Solo 5 granitos de arena para lanzarlos a una playa bastante, bastante extensa.

Otros trabajos escritos que no se pudieron incluir por que faltaba corregir cosas sobre lo que se había hecho y sobre lo que se quiso decir.

Lo que se hizo con honradez y con intenciones sinceras de aportar conocimientos, soluciones ó simples ideas nuevas ó con el deseo transparente de convertir en realidad los sueños de aportar un poco de hormigón a las bases del edificio de una mejor Venezuela.

Lo que no se puede escribir ni explicar con palabras. Las historias de muchos años por Barcelona y algunos lugares muy concretos en el espacio y el tiempo de la península ibérica. Y… lo que se escribirá en el futuro como fruto de esta cosecha, muy rico o muy pobre... No se, no se, no se.