Facultad de Ingeniería Eléctrica
Departamento de Automática

Trabajo de Diploma en opción al título de Ingeniero en Automática

Biblioteca para procesamiento de series temporales con fines de monitorización de procesos de unidades de generación eléctrica.

Detección y tratamiento de outliers, valores perdidos y ruido en minas de datos.

Autor: Yainier Albuerne Castro.

Tutores: MSc. Ing. David Díaz Martínez.
P.A. Dr. Ing. Luis Vázquez Seisdedos.

SANTIAGO DE CUBA
2016

Pensamiento

Conseguir algo sin arriesgar nada, lograr experiencia sin ningún peligro u obtener una compensación económica sin trabajar es tan difícil como vivir sin haber nacido.

A. J. Gouthey

Dedicatoria

Dedico este trabajo a la persona a quien debo todo lo que soy: mi madre; pues sin su cariño, amor y entrega que no repara en sacrificios no hubiera sido posible el logro de esta y otras tantas metas.

Agradecimientos

Ha sido un camino largo y repleto de dificultades, sin embargo, no he andado solo. Me han acompañado en esta travesía los mejores integrantes de un equipo, personas que por su constancia, dedicación y amor contribuyeron en la realización de un sueño. Por lo tanto, doy fe de gratitud:

- ✓ *A Idelsa Castro, mi querida madre, por ser excelente ejemplo de sacrificio y esfuerzo, por inspirarme a ser mejor y por dar todo de sí en pos de conseguir mis metas.*
- ✓ *Al resto de la familia por la ayuda brindada de una forma u otra, de manera significativa a mi hermana Yailén Albuerne, a mi hermano Tomasito y a mi primo Ernestico.*
- ✓ *Agradecimiento especial a mi tutor MSc. Ing. David Díaz Martínez por su ayuda incondicional en la elaboración de este trabajo, por enseñarme todo lo que necesitaba aprender al respecto. Gracias por sus conocimientos brindados, su entrega y dedicación.*
- ✓ *A mi tutor P.A. Dr. Ing. Luis Vázquez Seisdedos por su apoyo, colaboración y sus valiosos aportes a este trabajo.*
- ✓ *A Claudia Navarro por la dedicación constante, el amor, los consejos y todo su apoyo.*
- ✓ *A mis amigos por estar siempre a mi lado en las buenas y malas y por ayudarme a crecer como persona, especialmente a Manuel A. Ametller, Carlos D. Vázquez y Javier Cabrera.*
- ✓ *También al excelente grupo humano y profesional que nos guío, corrigió y acompañó durante el desarrollo de nuestros estudios.*

A todos los que caminaron a mi lado a lo largo de esta etapa, quienes me hicieron mucho más fácil la vida y quienes sinceramente me hacen sentir afortunado por haber disfrutado de su compañía y porque me ayudan a seguir creyendo en un futuro mejor.

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 su aprovechamiento en aras de producir un impacto positivo en la operación de plantas térmicas.

El punto de partida es la necesidad de asegurar la calidad de los datos para su posterior procesamiento. Se le da solución a problemas tales como la presencia de outliers o valores atípicos, los datos perdidos y el efecto del ruido en las mediciones. Se hace una revisión de los métodos de filtrado 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 industriales.

La realización de este trabajo se centrará en los problemas de obtención, manipulación y utilización de la información a partir de datos de operación sintetizando todo en un software desarrollado sobre la plataforma Matlab™ que permite utilizar una serie de herramientas que han sido integradas en una interfaz de usuario para facilitar la extracción del conocimiento a partir de las minas de datos del proceso.

Abstract

In this thesis a group of methodologies is presented that try to facilitate the task of exploitation of the information contained in the historical data of process and its use for the sake of producing a positive impact in the operation of thermal plants.

The starting point is the necessity to assure the quality of the data for its later prosecution. Is given solution to such problems as the outliers's presence or atypical values, the lost data and the effect of the noise in the measurements. A revision of the methods is made of having filtered univariable, especially those based on technical wavelets, since these last ones have been shown in the particularly advantageous literature for the filtrate of industrial data.

The realization of this work will be centered in the obtaining problems, manipulation and use of the information starting from operation data synthesizing everything in a software developed on the platform Matlab™ that allows to use a series of tools that have been integrated in user's interface to facilitate the extraction of the knowledge starting from the mines of data of the process.

Nomenclatura

ACP Análisis de Componentes Principales.

BCTI Boundary Corrected Translation Invariant.

Dbn Funciones wavelets pertenecientes a la familia Daubechies.

DCS Distributed Control Systems

DFT Discrete Fourier Transform

EWMA Exponential Weighted Moving Average.

FFT Fast Fourier Transform.

FHM Filtro Híbrido basado en la Mediana.

FIR Finite Impulse Response Filters.

IIR Infinite Impulse Response Filters.

IQP Industria Química y de Proceso.

KDD Knowledge Discovery in Databases.

MAR Missing at Random

MCAR Missing Completely at Random

MD Minería de Datos.

MNI Missing Non- Ignorable

NT Nodal Test

OLMS On Line Multiscale Filtering.

RB Ruido Blanco

RD Reconciliación de Datos.

RNA Redes Neuronales Artificiales.

SNR Signal Noise Ratio

Contenido

Introducción

Las mediciones de las variables de proceso aportan información clave para un amplio número de tareas en ingeniería, motivo por el cual cada día, a medida que crece y se desarrollan los equipos y técnicas para recopilar estos datos, también incrementa la necesidad de contar con herramientas para explotar estas minas de datos. Esto ha provocado la necesidad de técnicas de análisis de datos que ayuden a manejar estas minas para extraer información útil y con ella soportar la aplicación de diversas tareas operativas.

Al disponer de esta información, se ayuda a mejorar la operación de la planta mediante la identificación de aquellas variables que están influyendo sobre la eficiencia y el comportamiento de la misma.

Para la extracción de este conocimiento útil a partir de los datos del proceso es necesario utilizar diversas técnicas de un proceso que se conoce como minería de datos; pero los expertos reconocen que todavía existe en la actualidad una gran insuficiencia de herramientas para estos fines. Los Sistemas de Control Distribuidos (DCS del inglés Distributed Control Systems) no cuentan con la posibilidad de realizar estos procesamientos matemáticos avanzados de las observaciones instrumentales, solo realizan un pre-procesamiento elemental de estas mediciones entre los que podemos citar la detección de fallos en los sensores a partir del cálculo de los gradientes de incremento o decremento de una magnitud en determinado intervalo de tiempo, generación de alarmas a partir de valores muy por encima de las referencia permitidas, etc.

La Minería de Datos (MD) no es más que el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Forma parte de un dominio aún más grande que se conoce como descubrimiento del conocimiento a partir de datos (KDD del inglés Knowledge Discovery in Databases), el cual abarca: la recolección, la selección y la preparación de datos para las etapas siguientes relacionadas con el desarrollo de aplicaciones sobre estas minas.

El error total en una medición, puede representarse convenientemente como la suma de las contribuciones de dos tipos de errores: los errores aleatorios y los errores gruesos. El error aleatorio se puede definir como el efecto acumulativo de

cada una de las incertidumbres que ocasionan y dan lugar a que los datos de una serie de mediciones repetidas fluctúen al azar alrededor del valor medio de los mismos (media estadística). El error grueso es originado por varios factores, entre los más destacados están: mal calibración del instrumento, fugas en el proceso, corrosión, depósitos de sólidos, entre otros. La correcta identificación y tratamiento de estos es primordial para garantizar un buen control de la planta de proceso.

La presencia de valores perdidos (información ausente o faltante) es un problema común a cualquier investigación y no puede ser ignorado en el análisis de datos. Ignorar los datos ausentes puede tener repercusiones graves que van desde la pérdida de potencia de un estudio hasta la aparición de desviaciones inaceptables (Little y Rubin, 1987). La eliminación de tramas con características especiales limita la representatividad o validez externa de los resultados del estudio.

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.

En base a lo antes expuesto, para la realización de este trabajo se plantea como **problema de la investigación** la mala calidad de los datos industriales lo cual dificulta su procesamiento con fines de diagnóstico. Para ello se define como **objeto de la investigación** las series temporales registradas a partir de mediciones instrumentales de variables correspondientes a procesos industriales. El **objetivo** radica en diseñar e implementar una interfaz gráfica en Matlab™ que permita a partir de las facilidades que brindan un conjunto de funciones, la detección y tratamiento de outliers y de datos perdidos y la eliminación del ruido en las mediciones de un sub-proceso industrial. Es por esto que el **campo de acción** es el diseño de algoritmos de cómputo capaces de procesar estas minas de datos en aras de mejorar su calidad bajo el desarrollo de una interfaz gráfica utilizando el Toolbox GUIDE (del inglés Graphical User Interface Development Environment) de Matlab™. Para ello se plantea como **hipótesis** que si se diseñan un conjunto de algoritmos

que sean capaces de detectar (a partir de las series temporales registradas en un subproceso industrial) outliers, mediciones perdidas, ruido y ofrecer un buen tratamiento a estos problemas y estos logran ser integrados en una interfaz gráfica sobre Matlab™ entonces se dispondrá de mediciones con la calidad suficiente como para que puedan pasar a una segunda etapa de procesamiento.

Para el cumplimiento del objetivo propuesto se han asumido las siguientes tareas de investigación:

1- Diagnosticar y fundamentar los problemas y la insuficiencia de herramientas informáticas que procesen las series temporales de procesos multivariados para obtener información útil acerca de la planta.

2- Caracterizar desde el punto de vista gnoseológico, histórico y en la actualidad los aspectos teóricos referidos al procesamiento de las series temporales de procesos industriales.

3- Diseño de algoritmos computacionales para detectar y tratar errores, ausencia de datos y el ruido presente en las mediciones.

4- Construcción de una interfaz de usuario que sea capaz de integrar todas las prestaciones del sistema, entre ellas: graficar señales, visualizar y almacenar resultados numéricos, manejo de ficheros Excel para exportar datos, etc.

5- Probar la efectividad de los algoritmos de cálculo incorporados al producto de cómputo tanto con variables correspondientes a varios procesos industriales reales como con señales generadas a escala de simulación.

Métodos y técnicas empleados en la investigación:

1. Análisis de fuentes documentales.

2. Técnicas y métodos empíricos: Observación, encuesta y entrevista.

3. Método histórico lógico.

4. Método de análisis y síntesis.

5. Métodos experimentales: Programación

Significación Práctica:

La investigación ofrece al sector industrial una herramienta computacional abierta a nuevos códigos, que permite un procesado de grandes volúmenes de series temporales procedentes de observaciones instrumentales conservadas en costosas computadoras servidoras de un proceso industrial.

Abre las perspectivas de cooperación entre la Universidad de Oriente y el potencial de industrias químico-tecnológicas en el Territorio. En el ámbito del Departamento de Control Automático constituye un mayor desarrollo en el área del procesamiento de datos operacionales de procesos y por tanto, en la MD. El diseño e implementación de nuevos algoritmos para la identificación y tratamiento de errores, valores perdidos y ruido tienen en este software una precedencia valiosa para evaluarle su desempeño.

La presente investigación se encuentra organizada de la siguiente forma: una introducción general en la que se exponen las principales motivaciones que llevaron a la realización de esta investigación y en la cual se encuentra además, la fundamentación del diseño metodológico de la misma.

Dos capítulos que constan de introducciones parciales para la mejor comprensión de los objetivos de los mismos, estos a su vez, se encuentran organizados por epígrafes, de manera que resulte más fácil su revisión por parte del lector. Finalmente cada capítulo presenta sus conclusiones, además de las conclusiones generales, recomendaciones y bibliografía.

En el Capítulo I se presenta el estudio teórico del presente trabajo de diploma. Para ello se exponen las generalidades de minas de datos operacionales en la Industria Química y de Procesos (IQP) y la aplicación de varias técnicas para su procesamiento las que permiten la extracción de valiosa información a partir de las mismas que pueda ser útil en la mejora del funcionamiento de la planta.

En el Capítulo II se exponen las características específicas de la aplicación desarrollada, su interfaz de usuario, principales componentes, facilidades que ofrece y los algoritmos computacionales empleados.

Capítulo 1. Minas de datos operacionales en la Industria Química y de Procesos. Su tratamiento.

Introducción.

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). En la actualidad, estas industrias cuentan con sistemas que registran y almacenan diariamente los valores de cada una de las variables que intervienen en el proceso. Estos datos, conocidos como registros históricos o series temporales conforman enormes bases de datos que a pesar de estar presentes no se aprovecha debidamente el potencial de conocimiento existente en ellas. El acceso a esta información depende en gran medida de la calidad que tengan estas minas de datos.

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 (figura 1.1) 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).

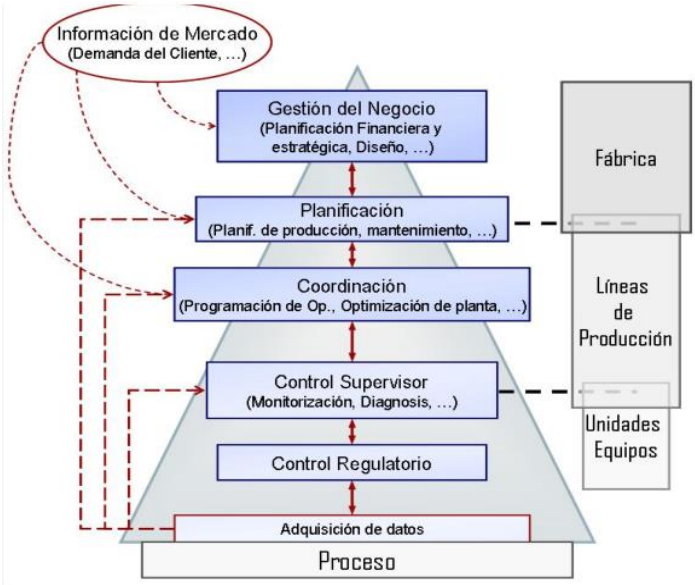

Figura 1.1 Estructura jerárquica - IQP.

En cada nivel, la toma de decisiones se soporta por una variedad de tareas operacionales (control regulatorio, rectificación de datos, monitorización y diagnóstico 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.).

1.2 Minas de Datos de Proceso = Adecuada Información de Soporte.

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 de la industria así como grupos de investigadores de diversos campos 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 Descubrimiento de Conocimiento en Bases de Datos (KDD del inglés Knowledge Discovery in Databases) y la Minería de Datos (MD) (Han y Kamber, 2001; Zabala, 2003).

1.3 Descubrimiento de información en bases de datos.

El KDD es 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.

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 (figura 1.2):

• *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 cómo 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) y el 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 inglés: 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.

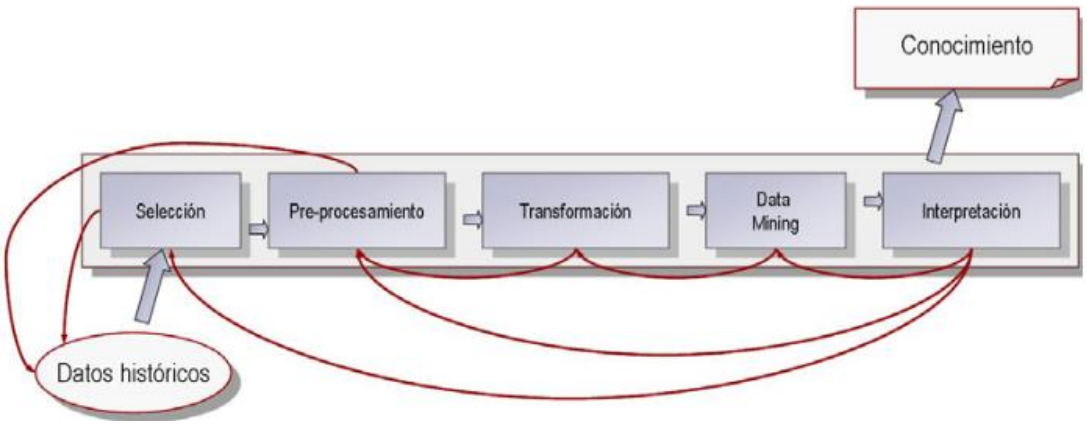

Figura 1.2 Arquitectura genérica del Proceso KDD.

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).

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, 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, 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 y de procesos.

1.4 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 y*, a partir de las mediciones del proceso y tales que:

$$y = y^* + \varepsilon \qquad (1.1)$$

Donde ε 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 1.3), en la IQP 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 dos grandes bloques:

• *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.

• *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.). Éstos 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. 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 1.5.3.2), las funciones wavelets para el filtrado y extracción de tendencias de las variables de proceso (Bakshi y Stephanopoulos, 1994b; Nounou y Bakshi, 1999; Jiang *et al.*, 2003).

• *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 rectificado, como la monitorización o diagnóstico de procesos, es mejor aplicar un filtrado previo al tratamiento de los datos ((Musulin et al ., 2006).

En la literatura también se pueden encontrar versiones multivariable de varios filtros univariable 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.

1.4.1 Tratamiento de outliers y errores aleatorios.

En cualquiera de los dos enfoques anteriores, la rectificación se orienta a la eliminación de errores aleatorios. Un valor atípico (en inglés outlier) es una observación que es numéricamente distante del resto de los datos. Estas observaciones anómalas han recibido diferentes nombres en diferentes áreas como, por ejemplo, observaciones atípicas, anómalas, discordantes, excepcionales, defectuosas, aberrantes, contaminantes, nocivas, entre otras. Las estadísticas derivadas de los conjuntos de datos que incluyen valores atípicos serán frecuentemente engañosas. 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 pérdida de oportunidad de conocer mejor y posiblemente mejorar la operación de un proceso.

El error total en una medición, puede representarse convenientemente como la suma de las contribuciones de dos tipos de errores: los errores aleatorios y los errores gruesos. El error aleatorio se puede definir como el efecto acumulativo de cada una de las incertidumbres que ocasionan y dan lugar a que los datos de una serie de mediciones repetidas fluctúen al azar alrededor del valor medio de los mismos (media estadística). Los errores aleatorios son errores pequeños debido a la fluctuación normal del proceso (variación de la presión atmosférica, fluctuaciones de tensión para instrumentos eléctricos) o a la variación aleatoria inherente en la operación del instrumento. El término error aleatorio implica que ni la magnitud ni la señal del error pueden predecirse con certeza y la única manera posible para que estos errores puedan caracterizarse es mediante el uso de distribuciones de probabilidad. El error grueso es originado por varios factores, entre los más destacados están: mal calibración del instrumento, fugas en el proceso, corrosión, depósitos de sólidos, entre otros. Dado que el error grueso es más preocupante, se han creado una serie de pruebas basadas en la estadística para su detección.

El método más usado para detectar error grueso es la Prueba de Hipótesis Estadística. Se declara un error grueso si la prueba estadística computada excede un valor crítico el cual se selecciona siguiendo una función de distribución normal estándar. Si el valor de la prueba estadística no excede el valor crítico, entonces la hipótesis nula H_0 es aceptada, y esto significa que la medida no contiene error grueso. Si el valor de la prueba estadística excede el valor crítico, entonces la hipótesis alternativa H_1 es aceptada, lo cual significa que la medida contiene un error

grueso. La prueba estadística de hipótesis incluye: Global Test, Nodal Test, Measurement Test y Generalized Likelihood Ratio, entre otras.

1.4.1.1 Reconciliación de datos.

La reconciliación de datos es una herramienta que permite la estimación única de variables cuyas mediciones son conflictivas. Esta técnica se basa en la propiedad de redundancia de las variables y en diferentes estimadores estadísticos que permiten ajustar la mediciones, con el fin de lograr que las leyes de conservación sean obedecidas al realizar los cálculos, por tanto, esta herramienta permite realizar un control estricto de los balances de masa y energía en los procesos industriales. En principio la reconciliación de datos se realiza por medio de la minimización de la función $\sum W_i * \rho(\varepsilon_i)$ donde ρ es la función objetivo, ε_i y W_i son el error y el factor de ponderación (desviación estándar) de la i-esima medición respectivamente; esta función de minimización esta sujeta a restricciones impuestas por los balances de masa y energía, los cuales dependen del sistema en estudio.

La definición de la función objetivo varía de acuerdo a las consideraciones tomadas con respecto al comportamiento de los errores. La reconciliación de datos convencional, implementa como función objetivo la función de mínimos cuadrados; este método cuenta con una fórmula analítica para su desarrollo y considera que el error sigue una distribución normal con media en cero, por lo que no deben existir errores gruesos dentro de las mediciones al realizar las operaciones, de lo contrario, se presenta un fenómeno conocido como *smearing effect*, en el cual los valores reconciliados se alejan de su valor convencionalmente verdadero, afectando a las variables libres de errores. El grado del smearing effect depende del grado de redundancia del sistema, de la ubicación del error, de las desviaciones estándar de las medidas y de tamaño del error grueso presente en estas. Otras funciones objetivo consideran que el comportamiento del error sigue una distribución normal con media diferente de cero o combinan dos distribuciones normales teniendo en cuenta el efecto que generan los errores gruesos, mejorando así los resultados de los reconciliados en presencia de estos errores; sin embargo, por la complejidad de estas funciones objetivos se requiere de métodos de optimización no lineales para ser solucionados, por tanto sus tiempos de cálculo son mayores; entre estas funciones se puede mencionar: Cauchy, Normal Contaminada, Lorentzian, Fair entre otras.

1.4.1.2 Métodos estadísticos para la detección de errores gruesos.

Si las medidas no contienen error aleatorio, se puede decir que el error grueso se detectaría simplemente con la violación de las restricciones (Balances de masa y energía). Sin embargo en la realidad la mayoría de las medidas tiene error aleatorio, este problema se soluciona aplicando al error una distribución probabilística.

Hay diferentes formas para identificar un error grueso:
I) Análisis teórico de todos los efectos que conducen a un error grueso.
II) Con mediciones de una variable por dos métodos con precisión diferente.
III) Comprobando la satisfacción de las ecuaciones de balance.

Esta tercera alternativa es en particular atractiva porque es relativamente simple y se basa en relaciones de validez absoluta, a saber, en la conservación de la masa y la energía. De las varias técnicas disponibles, las más usadas son la Prueba Global, la Prueba de Medición (Mah y Tamhane, (1982)), la Prueba de Medición Iterativa Modificada desarrollada por Serth y Heenan (1986), la Prueba Nodal (NT del inglés Nodal Prove), Cociente de Probabilidad Generalizada presentado por Narasimhan y Mah (1987), la Prueba de Componentes Principales y la Prueba de Potencia Máxima-Mínima. Entre ellos, tres tipos de estrategias se han desarrollado para identificar y corregir múltiples errores gruesos: de eliminación en serie, compensación serie, simultáneas o de compensación colectiva. El método Cociente de Probabilidad Generalizada permite identificar errores gruesos múltiples de cualquier tipo usando una estrategia serie de compensación. Para mejorar la eficiencia de la detección de errores gruesos en procesos industriales, una nueva estrategia son los métodos combinados con varias pruebas estadísticas. Estos métodos sólo detectan correctamente en presencia de un error grueso, con excepción de la Prueba Global. Cuando existen múltiples errores gruesos se requiere, además del uso de las pruebas anteriormente mencionadas, de una estrategia para la correcta identificación y localización de todos los errores gruesos. Estas estrategias se clasifican en dos grupos:

- Estrategias Simultáneas.
- Estrategias en serie:
    - Eliminación en serie
    - Compensación en Serie

La mayoría de las técnicas estadísticas están basadas sobre la prueba Hipotética (contraste de hipótesis):

*Hipótesis nula: H_0 Se presenta cuando el error de la medida se considera dentro de los límites establecidos para el error aleatorio

* Hipótesis alternativa: H_1 Se presenta cuando el error de las medidas está fuera del límite establecido para el error aleatorio.

Esta prueba estadística hipotética es comparada con un umbral o valor crítico. Si el valor de la prueba estadística no excede el valor crítico entonces la hipótesis nula (H_0) se acepta y estas medidas no contienen error grueso, pero si la prueba estadística excede el valor crítico entonces la hipótesis alternativa (H_1) es aceptada y estas medidas contienen error grueso. Sin embargo el resultado de la prueba hipotética no es perfecto. Una prueba puede declarar la presencia de error grueso cuando este no se encuentra, entonces la hipótesis nula es verdadera y este caso se llama Error Tipo I (falsa alarma). Por otro lado la prueba declara la medida libre de error grueso, cuando en realidad existe, este caso se denomina Error Tipo II.

1.4.2 Tratamiento de valores perdidos.

En la mayoría de los estudios muestreales y/o censales, principalmente en la medición de unidades, existen múltiples obstáculos, tales como perder una medición, lo que genera espacios vacíos que producen problemas en el análisis posterior. La presencia de valores perdidos (información ausente o faltante) es un problema común a cualquier investigación y no puede ser ignorado en el análisis de datos. Ignorar los datos ausentes puede tener repercusiones graves que van desde la pérdida de potencia del estudio hasta la aparición de desviaciones inaceptables (Little y Rubin, 1987). La eliminación de tramas con características especiales limita la representatividad o validez externa de los resultados del estudio. El análisis de valores perdidos ayuda a resolver varios problemas ocasionados por los datos incompletos. Además, los datos perdidos pueden reducir la precisión de los estadísticos calculados, porque no se dispone de tanta información como originalmente se pensaba. Otro problema radica en que los supuestos subyacentes a muchos procedimientos estadísticos se basan en casos completos y los valores perdidos pueden complicar la teoría exigida.

Desde hace ya varias décadas, se ha venido estudiando la forma de “llenar” estos espacios vacíos, con el fin de obtener un conjunto de datos completos para analizarse por la vía de los métodos estadísticos tradicionales. Sin embargo, esta situación se complica cuando se presentan en una matriz de datos formada por diversas variables sobre la cual se realizan estudios multivariantes, haciéndose necesario la aplicación de métodos que convenientemente imputen conjuntamente los datos (Geng y Li, 2003). Con la ayuda del avance de la computación, se han desarrollado nuevas formas de estudiar los datos faltantes multivariantes, obteniéndose una variedad de técnicas basadas en diferentes enfoques según las características de la data. Aun así, todavía son muchas las deficiencias que enfrentan las técnicas actuales y que son necesarias resolverlas, como las desviaciones en las estimaciones, alteración de la relación entre las variables, cambios en las varianzas, entre otros.

Se distinguen tres mecanismos de pérdida de datos:

•Datos perdidos completamente al azar (MCAR del inglés Missing Completely at Random). Cuando las características de las tramas con información son las mismas que las de las tramas sin información. Dicho de otra manera, la probabilidad de que una trama presente un valor ausente en una variable no depende ni de otras variables ni de los valores de la propia variable (Gleason y Staelin, 1975). Las observaciones con datos perdidos son una muestra aleatoria del conjunto de observaciones.

•Datos perdidos al azar (MAR del inglés Missing at Random). Cuando las tramas con datos incompletos son diferentes significativamente de los que presentan datos completos en alguna variable, y el patrón de ausencia de datos puede ser predecible a partir de variables con datos observados en la base de datos del estudio que no muestran ausencia de datos. La probabilidad de que se produzca la ausencia de una observación depende de otras variables pero no de los valores de la variable con el valor ausente. Es imposible probar si la condición MAR es satisfecha y la razón es que dado que no se conoce la información faltante no se puede comparar los valores de aquellas tramas que tienen información con los que no la tienen.

• Datos perdidos no ignorables o no debidos al azar (MNI del inglés Missing Non-Ignorable, o MNAR del inglés Missing Not At Random). Cuando la probabilidad de

los datos perdidos sobre una variable Y depende de los valores de dicha variable una vez que se han controlado el resto de las variables.

1.4.2.1 Métodos de imputación de datos.

Cuando no se pueden ignorar los datos faltantes, la manera más adecuada de tratarlos es llenar esos espacios faltantes con valores admisibles; a este procedimiento es lo que denominamos imputación. La importancia de los procedimientos de imputación no radica sólo en reducir la desviación por la ausencia de datos, sino también en producir un conjunto de datos rectangulares y “limpios” sin datos faltantes (Geng y Li, 2003).

Los métodos de imputación consisten en estimar los valores ausentes en base a los valores válidos y/o casos de la muestra. La estimación se puede hacer a partir de la información del conjunto completo de variables o viene de algunas variables especialmente seleccionadas. Usualmente los métodos de imputación se utilizan con variables métricas (de intervalo o de razón) y deben aplicarse con precaución porque pueden introducir relaciones inexistentes entre los datos.

Son muchas las técnicas de imputación que han surgido, sobre todo desde la década de los setenta, que emplean enfoques univariantes y multivariantes. Se han empleado enfoques basados en modelos como: funciones de verosimilitud, regresión y descomposición de matrices en valores singulares, entre otros. A pesar de estos avances, no se ha encontrado una metodología capaz de reproducir la data o que pueda resolver en forma totalmente satisfactoria el tratamiento de los datos faltantes, razón por la cual, aún se sigue investigando en busca de mejorar las técnicas existentes.

Los métodos de imputación se pueden resumir en los siguientes casos:

• *Sustitución por la media de las observaciones:* Consiste en sustituir los valores ausentes por la media de los valores válidos.

• *Sustitución por constante:* Consiste en sustituir los valores ausentes por constantes cuyo valor viene por razones teóricas o relacionadas con investigación previa.

• *Imputación por regresión:* Consiste en estimar los valores ausentes en base a su relación con otras variables mediante el Análisis de Regresión.

1.4.2.2 Selección de la técnica adecuada de imputación.

Seleccionar un método de imputación adecuado es una decisión de gran importancia, ya que para un conjunto de datos determinado, algunas técnicas de imputación podrían dar mejores aproximaciones a los valores verdaderos que otras. Para la selección de la técnica de imputación adecuada, no hay reglas específicas, dependerá entonces del tipo del conjunto de datos, tamaños del archivo, tipo de no respuesta, patrón de pérdida de respuesta, de los objetivos de la investigación, características específicas de la población, características generales de la organización del estudio, software disponible (Entilge 1996), importancia de los valores agregados o de los valores puntuales (microdato), distribuciones de frecuencias de cada variable, marginal o conjunta, etc.

Fellegi y Holt (1971), plantean que: “La técnica de imputación seleccionada debe superar las reglas de validación, cambiando lo menos posible los registros, manteniendo la frecuencia de la estructura de la data”. Goicoechea (2002), resume los criterios a tomar en consideración para seleccionar la técnica adecuada para imputar:

1. Tipo de variable a imputar
2. Parámetros que se desean estimar.
3. Tasas de no respuesta y exactitud necesaria.
4. Información auxiliar disponible.

Según Goicoechea (2002) los pasos que se llevan a cabo para realizar imputación son:

Paso 1: una vez que se dispone de un archivo con datos faltantes, se recopila y valida toda la información auxiliar disponible que pueda ser de ayuda para la imputación.

Paso 2: se estudia el patrón de pérdida de respuesta. Posteriormente se observa si hay un gran número de registros que simultáneamente tienen no respuesta en un conjunto de variables.

Paso 3: se seleccionan varios métodos de imputación posibles y se contrastan los resultados.

Paso 4: se calculan las varianzas para los distintos métodos seleccionados con el objetivo de obtener estimaciones con el mínimo sesgo y la mejor precisión.

Paso 5: se concluye a partir de los resultados obtenidos.

1.5 Ruido. Métodos de filtrado.

En general, se considera ruido a todas las perturbaciones que interfieren sobre las señales transmitidas o procesadas. Puede ser definido como una señal no deseada que interfiere con la medición de otra señal. Cuando la señal principal es analógica, el ruido será perjudicial en la medida que lo sea su amplitud respecto a la señal principal. Cuando las señales son digitales, si el ruido no es capaz de producir un cambio de estado, dicho ruido será irrelevante, sin descartar que el ruido nunca se puede eliminar en su totalidad. La principal fuente de ruido es la red que suministra la energía eléctrica, y lo es porque alrededor de los conductores se produce un campo magnético a la frecuencia de 50 ó 60 Hz. Además, por estos conductores se propagan los parásitos o el ruido producido por otros dispositivos eléctricos o electrónicos. Los datos del proceso suelen ser ruidosos, y el nivel del ruido cambia con el tiempo y las condiciones del proceso. En consecuencia, en el procesamiento de las mediciones adquiridas y transmitidas con diferentes técnicas es necesario considerar el ruido que los contamina y la taza de muestreo con que se leen. Por último, ya que la magnitud del ruido cambia con las condiciones de funcionamiento, el procedimiento debe ser independiente de la varianza del ruido.
Las señales industriales son afectadas por una amplia gama de ruidos. Típicamente, estos tienen naturaleza aditiva o multiplicativa. La figura 1.3.b muestra el efecto del ruido sobre una señal, se puede apreciar que la señal con ruido está distorsionada y es evidente que contiene otras componentes de frecuencias además de la original.

El ruido aditivo sigue la siguiente regla (Couch II, 2002):

$$x(t) = y(t) + \upsilon(t) \qquad (1.2)$$

Donde: x(t) señal contaminada con ruido

y(t) señal original sin ruido

υ(t) ruido presente.

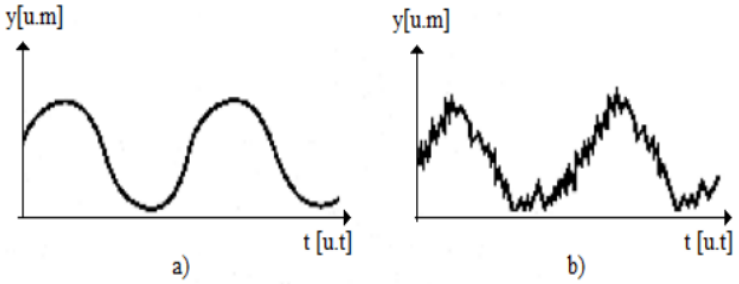

Figura 1.3 Representación de señales: a) Original b) Efecto del ruido.

Para medir la influencia del ruido sobre una señal se utilizan dos funciones propuestas por D. Donoho (Donoho *et al.*, 1993) la relación señal ruido (SNR del inglés Signal Noise Ratio) y el Error Medio Cuadrático (MSE); donde N es la cantidad de muestras de la señal, X es la señal original y $\bar{X}$ la señal tratada. Mientras mayor es la relación, menos efecto tiene la señal no deseada sobre la señal útil. Se define como la proporción entre la potencia de la señal con la potencia del ruido:

$$SNR = -20\log(sqrt(\frac{\sum_{i=1}^{N}(X_i-\bar{X})^2}{\sum_{i=1}^{N}{X_i}^2}) \quad (1.3)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2 \quad (1.4)$$

El ruido gaussiano es una de las formas más comunes en las que éste aparece en las señales industriales, generalmente es causado por procesos aleatorios de la corriente eléctrica o por las agitaciones térmicas de los elementos conductivos. Debe su nombre a que sus amplitudes siguen una distribución de Gauss (Coursey, 2003), independientemente de que exista una correlación del ruido en el tiempo o no. Dependiendo de su frecuencia o sus características de tiempo, el ruido de proceso puede clasificarse en una de varias categorías como sigue:

1- Ruido de Banda Estrecha: Un proceso de ruido con una banda estrecha tal como el zumbido de 50/60 Hz del suministro de electricidad.

2- Ruido blanco: Ruido puramente aleatorio que tiene un espectro de frecuencia plano. El ruido blanco teóricamente contiene todas las frecuencias con igual intensidad.

3- Ruido blanco limitado en banda: Un ruido con un espectro plano y un ancho de banda limitado que usualmente cubre el espectro del dispositivo o la señal de interés.

4- Ruido coloreado: Ruido de banda ancha cuyo espectro tiene una forma no plana; los ejemplos son ruido de rosado, el marrón y el ruido autorregresivo.

5- Ruido Impulsivo: Consiste en pulsos de duración breve de amplitud y duración aleatorios.

6- Ruido de Pulsos Transitorio: Consiste en pulsos de relativamente larga duración

A su vez, la autocorrelación es otra de las propiedades en la caracterización de las señales de ruido. Esta propiedad está relacionada con la forma de distribución del espectro de potencia.

1.5.1 Ruido blanco.

El ruido blanco (RB) es una señal aleatoria (proceso estocástico) que se caracteriza por el hecho de que sus valores de señal en dos tiempos diferentes no guardan correlación estadística (figura 1.4). Como consecuencia de ello, su densidad espectral de potencia (PSD, siglas en inglés de Power Spectral Density) es una constante, es decir, su gráfica es plana. Esto significa que la señal contiene todas las frecuencias y todas ellas muestran la misma potencia. Igual fenómeno ocurre con la luz blanca, de allí la denominación.

Figura 1.4 Ilustración de a) Ruido blanco b) Su espectro de potencia.

La densidad espectral de potencia del ruido blanco está dado por la ecuación 1.5:

$$\mathrm{PSD(f)} = \int_{-\infty}^{\infty} r_{NN}\,(\mathrm{t}) e^{-j2\pi ft} dt = \Sigma^2 \qquad (1.5)$$

donde t y f son el tiempo y la frecuencia respectivamente y r_{NN} es la función de auto-correlación de un ruido blanco de tiempo continuo.

Esta ecuación muestra que el ruido tiene un espectro de potencia constante.

El ruido blanco es una señal no correlativa, es decir, en el eje del tiempo la señal toma valores sin ninguna relación unos con otros. Si la PSD no es plana, entonces se dice que el ruido está "coloreado" (correlacionado).

1.5.2 Ruido coloreado.

Aunque el ruido es una señal aleatoria, puede tener características y propiedades estadísticas. La densidad espectral (potencia y distribución en el espectro de frecuencia), es una de esas propiedades, que pueden ser utilizadas para distinguir los diferentes tipos de ruido. Esta clasificación, por densidad espectral, da la terminología con el nombre de diferentes tipos de colores. Dependiendo de la forma concreta que tenga su PSD, se definen varios "colores" para el ruido.

Dos variedades clásicas de ruido coloreado (RC) son llamadas ruido rosado (figura 1.5) y ruido marrón.

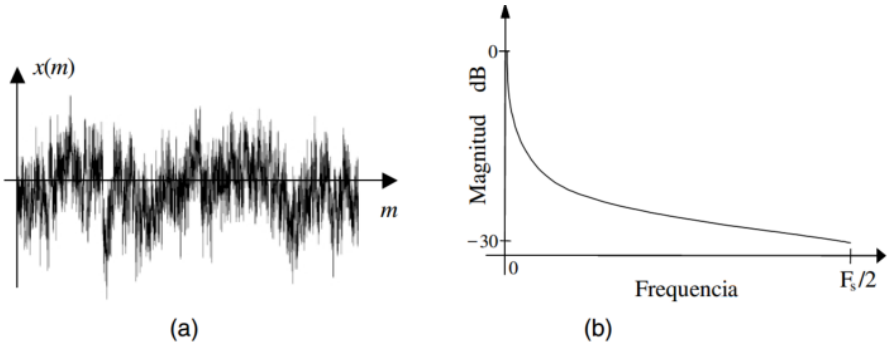

Figura 1.5 a) Señal con ruido rosado b) Su espectro de amplitud.

El ruido rosa es un ruido cuyo nivel sonoro está caracterizado por una densidad espectral inversamente proporcional a la frecuencia. El ruido marrón es un tipo de ruido conocido también como ruido rojo, o ruido browniano en honor a Robert Brown, el descubridor del movimiento browniano. No es un ruido muy común, pero existente en la naturaleza. El ruido marrón está compuesto principalmente por frecuencias graves y medias.

En el decursar del tiempo, para reducir el ruido, se han utilizado diferentes técnicas de filtrado lineal, principalmente esquemas basados en filtros que trabajan directamente sobre la señal en cuestión. Otro conjunto de técnicas para remover el ruido de las señales son las basadas en el uso de transformadas, donde la más conocida y utilizada es la Transformada Discreta de Fourier (DFT del inglés Discrete Fourier Transform), ideal para señales estacionarias cuyos valores varían muy poco

en el tiempo, pero no igual de efectiva para el análisis de señales con variaciones significativas. Tanto las técnicas de filtrado lineal como las basadas en FT presentan importantes dificultades para el tratamiento de señales no estacionarias (Dolabdjian *et al.*, 2002). En este contexto aparece la transformada Wavelet.

1.5.3 Filtrado.

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.

1.5.3.1 Rectificación basada en filtros univariable.

Por lo general 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 y(k) compuesta por una realización de datos sucesivos y afectados por ruido gaussiano o aleatorio tal que y(k) = (y(1), y(2),...). Se acepta ampliamente que y(k) puede expresarse como sigue (Bakhtazad *et al.*, 1999; Craigmile y Percival, 2002):

$$y(k) = y^*(k) + \varepsilon(\mathrm{k}) \tag{1.6}$$

Donde y*(k) es la tendencia verdadera (no contaminada con ruidos) de la señal original y ε(k) es el error aleatorio que se añade a la medición (ruido). Esta descripción supone la no presencia de errores gruesos afectando a y(k). La meta de la filtración es estimar un valor apropiado de y*(k) (ŷ*(k)) mediante la eliminación de la mayor parte del ruido.

Los filtros lineales 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):

$$\hat{\mathrm{Y}}^*(k) = \sum_{j=0}^{nc-1} b_i \, \mathrm{Y}(\mathrm{k} - \mathrm{i}) \tag{1.7}$$

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:

$$\sum_j b_{l=1} \tag{1.8}$$

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 inglés 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 donde 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 inglés 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.

FIR se trata de un tipo de filtros digitales cuya respuesta a una señal impulso como entrada tendrá un número finito de términos no nulos. Para obtener la salida sólo se basan en entradas actuales y anteriores. Su expresión general está dada por la ecuación:

$$y_n = \sum_{k=0}^{N} b_k \, x(nT_s - kT_s) \tag{1.9}$$

En la expresión anterior N es el orden del filtro, que también coincide con el número de términos no nulos y su número de coeficientes del filtro es N+1 y T_s Periodo de muestreo. Los coeficientes son b_k. Aplicando la transformada Z a la expresión anterior obtenemos:

$$H(Z) = \sum_{k=0}^{N} b_k Z^{-k} = h_0 + h_1 Z^{-1} + h_2 Z^{-2} + \cdots + h_N Z^{-N} \tag{1.10}$$

De esta manera se obtiene la estructura básica de este tipo de filtros como se muestra en la figura 1.6.

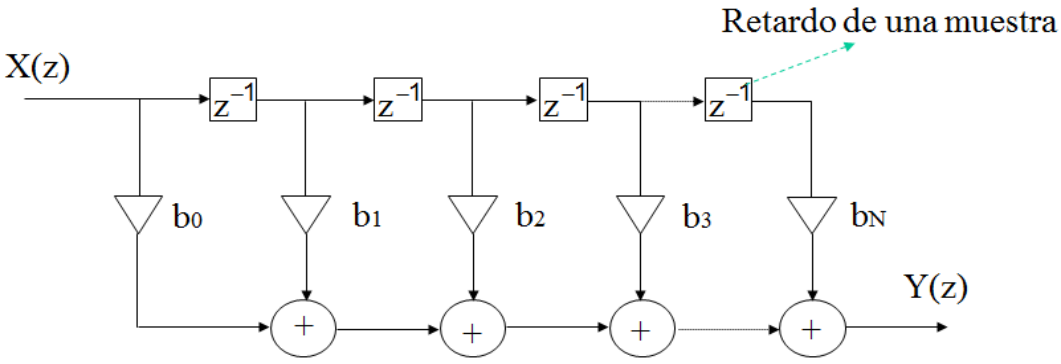

Figura 1.6 Estructura básica de un filtro FIR.

Los filtros FIR tienen la desventaja de necesitar un orden mayor respecto a los IIR para cumplir las mismas características. Esto se traduce en un mayor gasto computacional.

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:

$$\hat{y}^*(k) = \alpha y(k) + (1 - \alpha)\,\hat{y}^*(k-1) \qquad (1.11)$$

Donde α es una constante de suavización que recoge el efecto de los pesos asignados a cada observación.

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 inglés FIR-Median Hybrid Filtres o FMH). La forma general de un FHM es como sigue (Kosanovich et al, 1997):

$$\hat{Y}^*(k) = \text{median}(\text{FIR}(y(k-1), y(k-2), \ldots, y(k), \text{FIR}(y(k+1), y(k+2), \ldots, y(k+nt))) \qquad (1.12)$$

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).

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 1.7 (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 sinusoidal, un fallo puntual de un equipo, una degradación en la operación de otro equipo y el ruido.

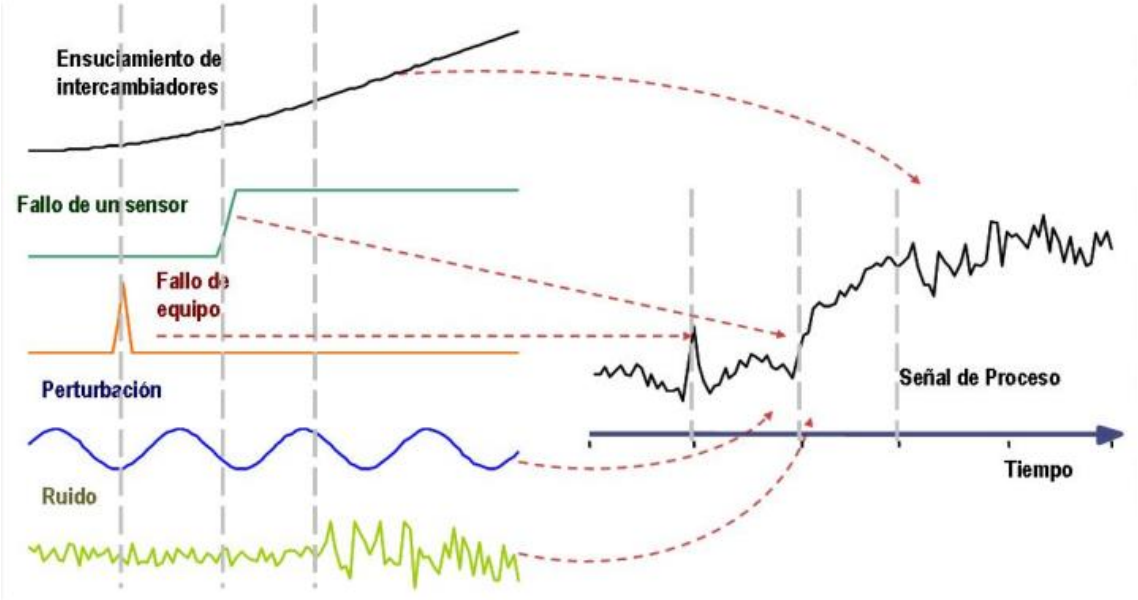

Figura 1.7 Ilustración de una señal de proceso.

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.

1.5.3.2 Las técnicas wavelets.

El término wavelet significa onda pequeña. La pequeñez se refiere al hecho que esta función (ventana) es de longitud finita (compactamente soportada) y el término onda se refiere a la condición que esta función es de naturaleza oscilatoria.

La transformada de ondícula o transformada wavelet es un tipo especial de transformada matemática que representa una señal en términos de versiones trasladadas y dilatadas de una onda finita (denominada óndula madre). La teoría de

ondículas está relacionada con campos muy variados. Todas las transformaciones de ondículas pueden ser consideradas formas de representación en tiempo-frecuencia.

Las Transformadas de Wavelets (WT) están dadas por la Transformada Wavelet Continua (CWT del inglés Continuous Wavelet Transformed) y la Transformada Wavelet Discreta (DWT del inglés Discrete Wavelet Transformed). Son dos herramientas que permiten el análisis de señales de manera similar a la Transformada de Fourier con la diferencia que la WT puede entregar información temporal y frecuencial en forma cuasi-simultánea, mientras que la TF sólo da una representación frecuencial. De acuerdo al principio de incertidumbre de Heisenberg, existen limitaciones con la resolución en el tiempo y frecuencia, pero es posible realizar un análisis usando la Transformada Wavelet, que permite examinar la señal a distintas frecuencias y con diferentes resoluciones. La WT da una buena resolución temporal y baja resolución en frecuencia para eventos de altas frecuencias y da una buena resolución frecuencial pero poca resolución temporal en eventos de bajas frecuencias.

Las wavelets representan un conjunto de funciones matemáticas especialmente diseñadas para análisis multirresolución o multiescala. 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 [$\varphi_{l,\upsilon}$ (k), $\upsilon \in Z$ in V_j] y para el espacio de los detalles con las funciones wavelets [$\psi_{l,\upsilon}$ (k) l=1,…,L, $\upsilon \in Z$ en 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.

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:

$$\phi_{\mathrm{L},\upsilon} = 2^{-L}\phi(2^{-L}k - \upsilon), \upsilon \epsilon z \qquad (1.13)$$

$$\Psi_{L,\upsilon=}2^{-l}\Psi(2^{-l}k-\upsilon), \mathrm{l}=1,\ldots,\mathrm{L};\ \upsilon\epsilon z \quad (1.14)$$

Utilizando las funciones anteriores, una señal y(k) se puede descomponer como sigue (representación multiescala):

$$y(k)=\sum_{\Upsilon\epsilon z} \mathrm{A}_{\mathrm{L},\upsilon}\phi_{\mathrm{L},\upsilon}(k)+\sum_{l=1}^{L}\sum_{\Upsilon\epsilon z} d_{\mathrm{L},\upsilon}\Psi_{\mathrm{L},\upsilon}(k) \quad (1.15)$$

Donde $\alpha_{l,\upsilon}$ 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, A_L (k), y los componentes de detalles, D_l (k), con (l=1,…, L):

$$A_L(k)=\sum_{\Upsilon\epsilon z} \mathrm{A}_{\mathrm{L},\upsilon}\phi_{\mathrm{L},\upsilon}(k) \quad (1.16)$$

$$D_l(k)=\sum_{\Upsilon\epsilon z} \mathrm{D}_{\mathrm{L},\upsilon}\Psi_{l,\upsilon}(k) \quad (1.17)$$

A_L(k) y D_l(k) se construyen a partir de los coeficientes $\alpha_{l,\upsilon}$ y $d_{l,\upsilon}$ respectivamente representando información a distintas escalas. Luego, la ecuación 1.17 se rescribe como sigue:

$$Y(k)=A_L(k)+\sum_{l=1}^{L} D_l(k) \quad (1.18)$$

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 ŷ*(k). Similarmente, 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 ŷ*(k) (Addison, 2002).

1.5.3.3 Rectificación utilizando wavelets.

Las wavelets se han utilizado con éxito en aplicaciones de filtrado de datos explotando para ello la capacidad de descomponer una señal. La idea principal del filtrado con wavelets se basa en transformar los datos a la base wavelet. En esta nueva base los coeficientes estimados que tienen valores más altos (principalmente los coeficientes de A_L) representan la información útil de la señal mientras que los coeficientes con valores más pequeños (principalmente los coeficientes de los D_l) se asociarán al ruido. Mediante una modificación apropiada de los coeficientes de 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 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 tres pasos (figura 1.8):

• 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 tomando el inverso de la wavelet usada, sobre los coeficientes que quedan tras el thresholding (Bloque W -1).

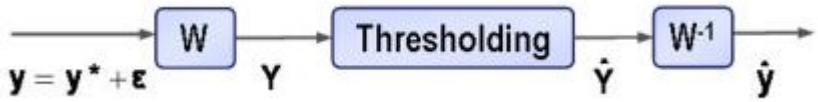

Figura 1.8 Método waveshrink de filtración usando wavelets.

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

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

$$d_j^H = \begin{cases} d_j & if |d_j| > \beta \\ 0 & if |d_j| \leq \beta \end{cases} \tag{1.19}$$

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.

$$d_j^S = \begin{cases} sign d_j \left(|d_j - \beta| \right) & if |d_j| > \beta \\ 0 & if |d_j| \leq \beta \end{cases} \tag{1.20}$$

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).

Conclusiones Parciales.

1. Se describen los problemas 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.

2. Se muestra la importancia de considerar y eliminar los errores gruesos y los valores perdidos y se explican los principales métodos que existen para este fin.

3. Se analiza el efecto del ruido en las mediciones a la hora de procesar los datos y las diferentes técnicas existentes para lograr su eliminación llegando a la consideración de que el procesamiento mediante Transformada Wavelet es el más adecuado para su implementación en este trabajo por su buen comportamiento ante señales correlacionadas.

Capítulo 2. ISLab: plataforma para el procesamiento de mediciones industriales.

Introducción.

En condiciones ideales se obtendrían los valores reales de las mediciones industriales lo cual supondría una gran ventaja y facilidad para el procesamiento de datos. En condiciones reales cada serie temporal que se registra viene contaminada y esto conlleva a la necesidad de una herramienta que le dé un tratamiento en pos de reducir o eliminar esta contaminación. En este capítulo se fundamentan los aspectos necesarios para el diseño de un sistema capaz de detectar outliers, datos ausentes así como ruido a partir de mediciones industriales y a su vez ofrecer un tratamiento adecuado a dichos problemas con el objetivo de poder entregar mediciones listas para pasar a otra etapa de procesamiento, como: detección de estados estacionarios o detección de fallos. Con este propósito nace el Laboratorio de Señales Industriales (**ISLab** del inglés **I**ndustrial **S**ignals **Lab**oratory). A continuación se exponen sus generalidades, interfaz de usuario, principales componentes, facilidades que ofrece y los algoritmos computacionales empleados.

2.1 Generalidades.

El ISLab es un software construido usando las facilidades del Matlab™ versión 7.10 del 2010. Es un laboratorio virtual diseñado para el procesamiento de las mediciones industriales con fines de monitorización. El software utiliza algunas herramientas que han sido programadas a partir de la necesidad de resolver determinados problemas que afectan las series temporales, particularmente aquellas que proceden de la industria de procesos.

El programa está montado sobre una interface de usuario sencilla y práctica que facilita la interacción con las funciones del sistema. Para ello se usó el Toolbox GUIDE (del inglés Graphical User Interface Development Environment) del Matlab™, un ambiente de desarrollo integrado para la confección de interfaces gráficas de usuario que permite crear de modo interactivo esta interface y ejecutar programas que necesiten ingreso continuo de datos.

Este software (figura 2.1) enfoca su desarrollo en la solución de tres problemas fundamentales (outliers, datos perdidos y ruido) y de esa forma se encuentra dividido en tres etapas que dan solución a cada problema. El dato de salida de la primera etapa es entrada para la segunda y la salida de ésta pasa a la última. Esta concepción flexibiliza la introducción de mejoras.

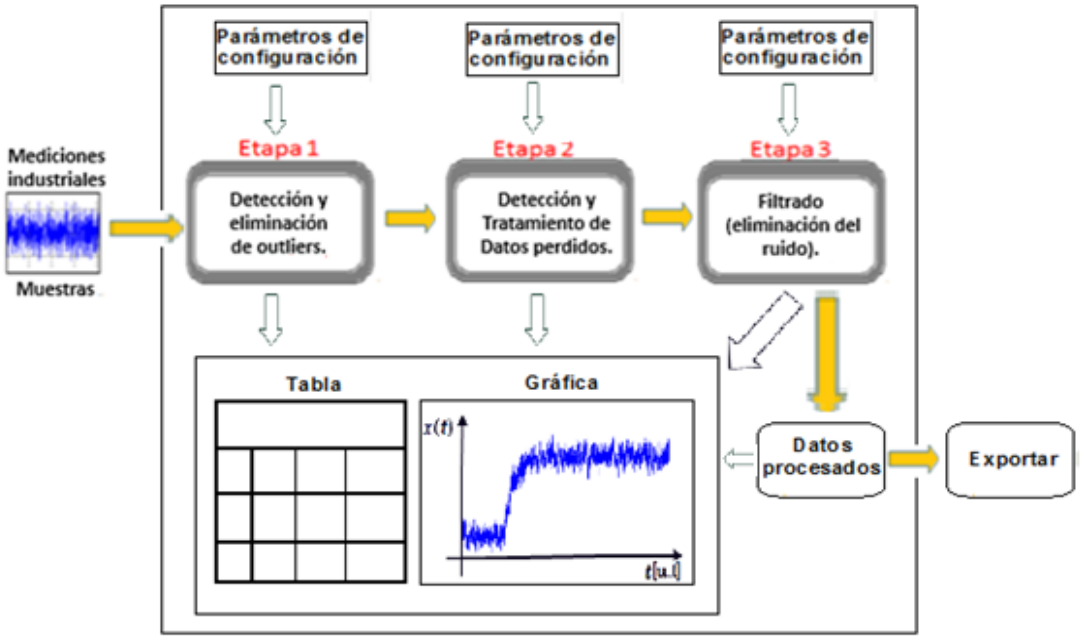

Figura 2.1 Estructura general del ISLab.

Es importante señalar que los métodos y algoritmos explicados en este trabajo son solo los seleccionados para su implementación en cada una de las etapas del procesamiento pero no son los únicos que pueden ser incorporados. El ISLab es una herramienta abierta a la incorporación de nuevos métodos de acuerdo a las ventajas o desventajas de cada uno de ellos en cuanto a cantidad de parámetros en entrada requeridos, carga computacional o exactitud en la estimación. El software está concebido de manera que se mantiene la estructura de entrada-salida de cada bloque funcional independientemente del método utilizado por lo que se garantiza la escalabilidad y operatividad del sistema.

2.2 Análisis de la primera etapa. Principales características.

Las técnicas estadísticas más habituales para el análisis de datos están basadas en modelos que suponen un cierto número de outliers. Sin embargo, en la práctica, puede ocurrir que el modelo supuesto es capaz de describir acertadamente una gran parte de los datos, pero un pequeño grupo de observaciones parecen seguir un comportamiento diferente.

Se podría ignorar la presencia de estos datos atípicos, pero entonces se estaría perjudicando el análisis de múltiples formas dependiendo del tipo de datos que se procese. En series temporales, los datos atípicos son más complejos que en otro tipo de modelos debido a la estructura temporal. Por esto la primera tarea que realiza el ISLab es la detección y eliminación de estas mediciones erróneas, usando un algoritmo capaz de reconocer cuando la medición es equivocada de acuerdo a las características de la señal bajo tratamiento.

2.2.1 Detección de valores atípicos.

El procedimiento para la detección de outliers realiza dos funciones principales (figura 2.2):

- Estimar las medias y desviaciones típicas.
- Describir el patrón de los errores gruesos, es decir, las posiciones dentro de la trama de datos donde se encuentran.

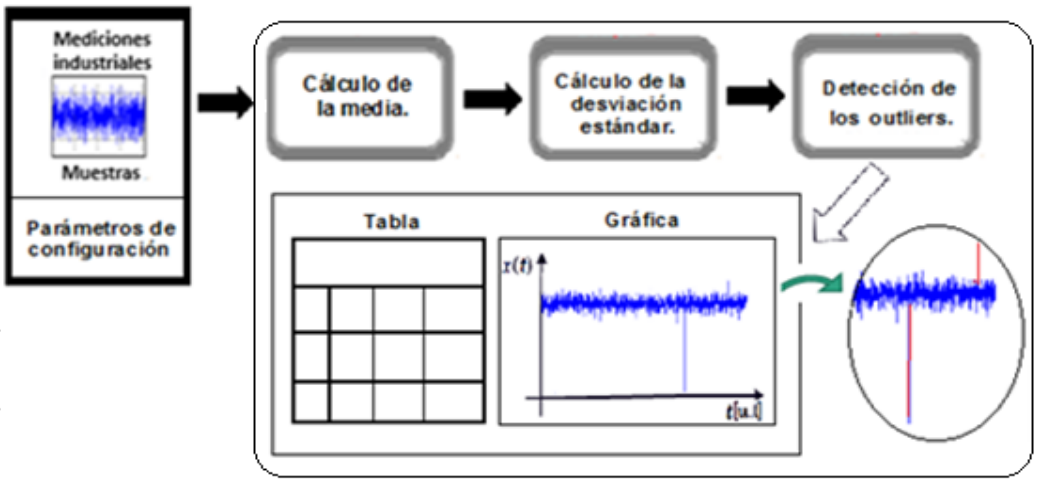

Figura 2.2 Diagrama funcional de la detección de outliers en el ISLab.

El primer paso para lograr la detección de los outliers es determinar el tamaño de la data, luego se establece una ventana con un ancho determinado (en cantidad de muestras) sobre la cual se realizan las operaciones en cada uno de los pasos (figura 2.10). El primer estadígrafo que se obtiene en esa ventana es el valor característico de la serie de datos más conocido como media aritmética (ecuación 2.1). Luego se halla una medida del grado de dispersión de los datos con respecto al valor promedio, dicho de otra manera, la desviación estándar o simplemente el promedio o variación esperada con respecto a la media aritmética (ecuación 2.2). Si el valor absoluto de la diferencia entre la muestra y la media es mayor o igual que n (factor

de multiplicación) veces la desviación estándar entonces se considera fuera de los límites del valor crítico lo que convierte a esta muestra en un outlier (figura 2.3).

$$\bar{x} = \frac{1}{j}\sum_{i=1}^{j} \frac{x_1+x_2+\cdots+x_j}{j} \quad (2.1)$$

$$s^2 = \frac{1}{j}\sum_{i=1}^{j}(x_i - \bar{x})^2 \quad (2.2)$$

Donde: x_i Muestras, $\bar{x}$ Media aritmética, j cantidad de muestras y s desviación estándar.

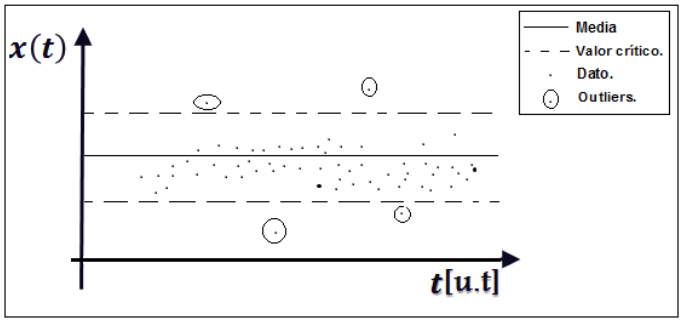

Figura 2.3 Representación gráfica de la detección de outliers.

La ventana tiene un ancho (por defecto 10) que puede ser cambiado por el usuario al igual que el valor de n. Matemáticamente este procedimiento está representado por la ecuación 2.3.

$$|X - \bar{X}| >= n * s \quad (2.3)$$

Donde: X muestras, $\bar{X}$ Media aritmética, n factor de multiplicación y s desviación estándar.

Para la implementación de este procedimiento se utilizó la función Matlab *bsxfun*. Este comando realiza una operación binaria (fun) entre dos matrices (A y B) elemento por elemento. La sintaxis del comando se representa del siguiente modo: *C=bsxfun(fun,A,B).* La operación que realiza está dada por un conjunto de funciones Matlab; las que emplea el procedimiento para la detección de los outliers son *@minus* para la *sustracción* y *@gt* para *mayor que* quedando una línea de código que es el equivalente de la ecuación 2.3:

*I = bsxfun(@gt,abs(bsxfun(@minus,tramo,media)),n*desv); % tramo es la ventana móvil, la función abs se emplea para tomar el valor absoluto de la diferencia.*

2.2.2 Eliminación de los valores atípicos.

Una vez que se han detectado las anomalías y se tienen sus posiciones se procede a su eliminación. Este proceso de eliminación deja espacios vacíos sobre los que se deben insertar valores permisibles y lógicos en lugar de los defectuosos (figura 2.4).

Para rellenar los espacios donde estaban los errores gruesos eliminados se eligió un método de imputación que consiste en una regresión multilineal robusta que se lleva a cabo con el uso del comando *robustfit* que devuelve un vector con los valores estimados para la imputación.

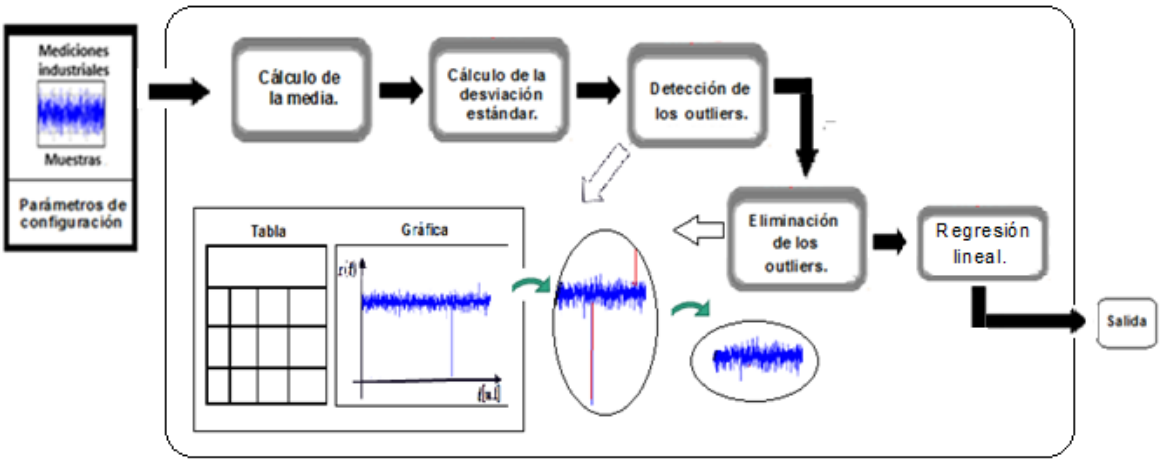

Figura 2.4 Diagrama funcional de la eliminación de outliers en el ISLab.

Este tipo de regresión es una forma de análisis diseñada para eludir algunas limitaciones tradicionales de los métodos paramétricos y no paramétricos que emplea el método de los mínimos cuadrados. Es una regresión robusta diseñada para no ser excesivamente afectada por violaciones de los supuestos por el proceso de generación de datos subyacentes, algo que no sucederá debido a que éstos ya han sido detectados y eliminados.

2.3 Etapa 2. Tratamiento de los datos perdidos.

Los casos con valores perdidos representan un reto importante, ya que los procedimientos de modelado tradicionales simplemente descartan estos casos para el análisis. En ocasiones, cuando hay pocos valores perdidos (aproximadamente, menos del 5 % del número total de casos) y dichos valores pueden considerarse perdidos de forma aleatoria (es decir, que la pérdida de un valor no depende de otros valores), entonces, se considera que la lista de datos es relativamente “segura” y no se le da tratamiento a este problema.

El ISLab no sigue este concepto, a cada valor ausente detectado se le realiza un tratamiento, llenando este espacio vacío en correspondencia con la forma de onda de la señal. De esta forma se garantiza descartar los problemas ya mencionados asociados a trabajar con series afectadas con ausencia de algunos de sus valores.

El procedimiento de análisis de valores perdidos realiza dos funciones principales:

- Describe el patrón de los datos perdidos, es decir, las posiciones donde se encuentran.
- Rellena (imputa) los valores perdidos con valores estimados utilizando un método de regresión multilineal robusta.

2.3.1 Detección de datos ausentes.

Una vez que la señal es procesada en la primera etapa resulta una señal libre de errores gruesos. La detección de los valores perdidos es el segundo bloque de procesamiento, el algoritmo implementado queda representado en la figura 2.5.

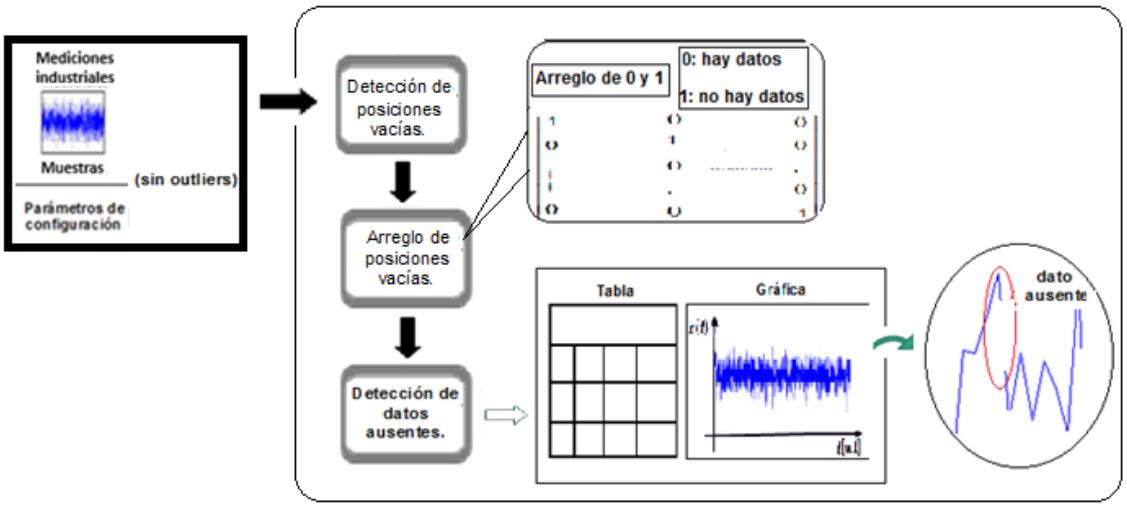

Figura 2.5 Diagrama funcional de la detección de datos perdidos en el ISLab.

Básicamente consiste en una especie de censo poblacional en el que se define a la población como el número de muestras por las que está compuesta la señal y se realiza una encuesta, dicho de algún modo, a la posición ocupada por cada muestra creando un arreglo multidimensional conformado por ceros y unos donde cada cero representa aquella posición en la que se encuentra un valor lógico (dato numérico) y los unos indican las posiciones que no contienen ningún valor o contengan un valor ilógico o no numérico. De esta manera se obtienen todas las posiciones en las que existe ausencia de valores y a la vez la cantidad de estos espacios vacíos.

El comando usado para tales propósitos es:

MD=find(isnan(X)); % devuelve en un arreglo las posiciones de los valores no numéricos (posiciones vacías).

2.3.2 Eliminación de datos perdidos.

La importancia de los procedimientos de imputación no radica sólo en reducir la desviación por la ausencia de datos, sino también en producir un conjunto de datos rectangulares y "limpios" sin datos faltantes (Geng y Li, 2003).

Para llenar aquellos lugares vacíos que fueron encontrados en la serie temporal se eligió el método de basado en regresión multilineal robusta que se lleva a cabo con el uso del comando del Matlab *robustfit* (epígrafe 2.2.2). Cuando se imputa, se logra obtener una base de datos completa, la cual permitirá llevar a cabo metodologías de análisis de datos comunes y el uso de software tradicionales para su manejo.

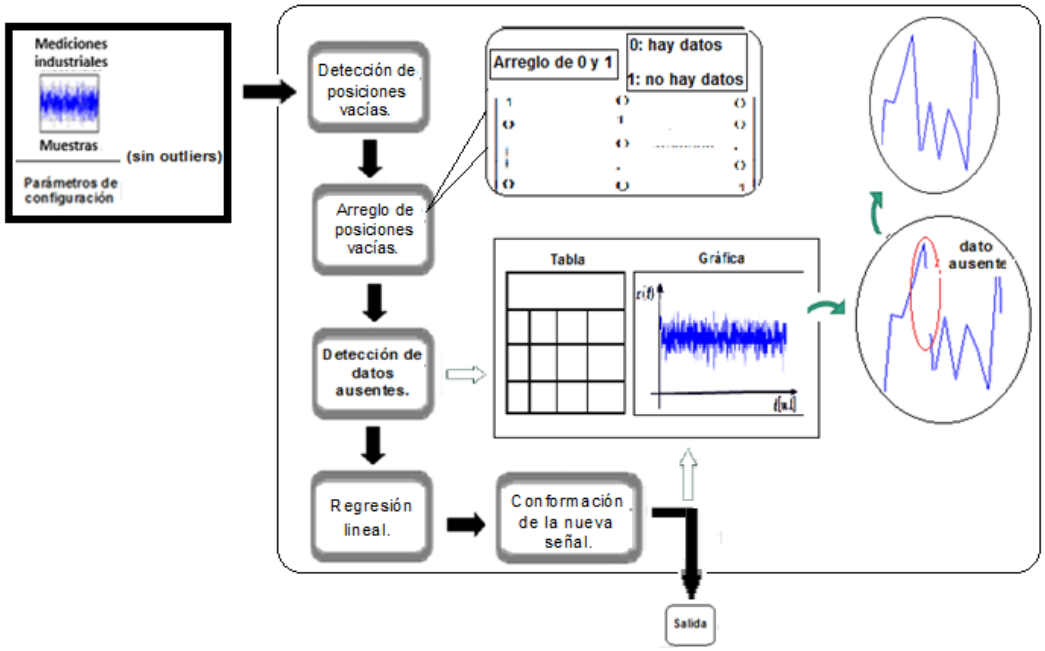

Figura 2.6 Diagrama funcional del tratamiento a datos perdidos en el ISLab.

De este modo se obtiene a la salida de esta etapa un archivo de datos consistentes y libre no solo de errores gruesos sino además de datos perdidos; elementos que repercutirían de manera negativa en un futuro procesamiento de la señal.

Luego de haber pasado por las dos etapas antes descritas, los datos resultantes llegan a una última etapa de gran importancia que será la encargada de extraer o reducir el ruido. A continuación se describen los elementos fundamentales que componen la etapa de filtrado, sus características esenciales y algoritmos de trabajo.

2.4 Filtrado. Caracterización de la etapa final.

Un filtro digital es un tipo de filtro que opera sobre señales discretas y cuantizadas, implementado con tecnología digital, bien como un circuito digital o como un programa informático. Es un sistema que, dependiendo de las variaciones de las señales de entrada en el tiempo y amplitud, realiza un procesamiento matemático sobre dicha señal, obteniéndose en la salida el resultado del procesamiento matemático o la señal de salida.

Hay varios tipos de filtros así como distintas clasificaciones para estos filtros. Una de las principales características del ISLab es que ofrece una variada selección de filtros (figura 2.7), todos con un marco de diseño flexible que brindan la posibilidad de ser configurados por el propio usuario variando sus parámetros fundamentales (recomendado solo a conocedores).

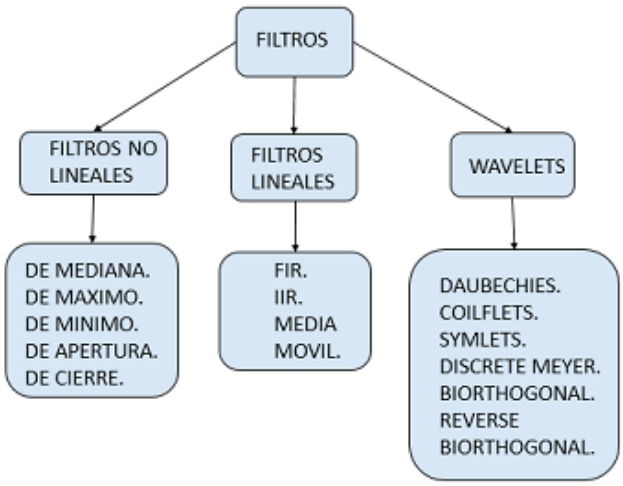

Figura 2.7 Filtros que posee el ISLab.

2.4.1 Tipos de filtros.

Con el objetivo de brindar una herramienta flexible que ofrezca a los especialistas la posibilidad de actuar de acuerdo a sus criterios particulares en cuanto a las principales técnicas de reducción del ruido en las mediciones, se ofrece un conjunto de filtros (figura 2.8) con características y modos de trabajo diferentes pero eficaces y precisos.

El ISLab contiene cinco filtros no lineales entre los que se pueden citar el de mediana, de máximo y de mínimo. Estos filtros se especializan en la reducción del ruido impulsivo o ruido no gaussiano. El filtro de media móvil así como los clásicos FIR e IIR son algunos de los filtros lineales más usados desde hace ya un buen

tiempo y aun se siguen empleando en el procesamiento de señales debido a su buen desempeño.

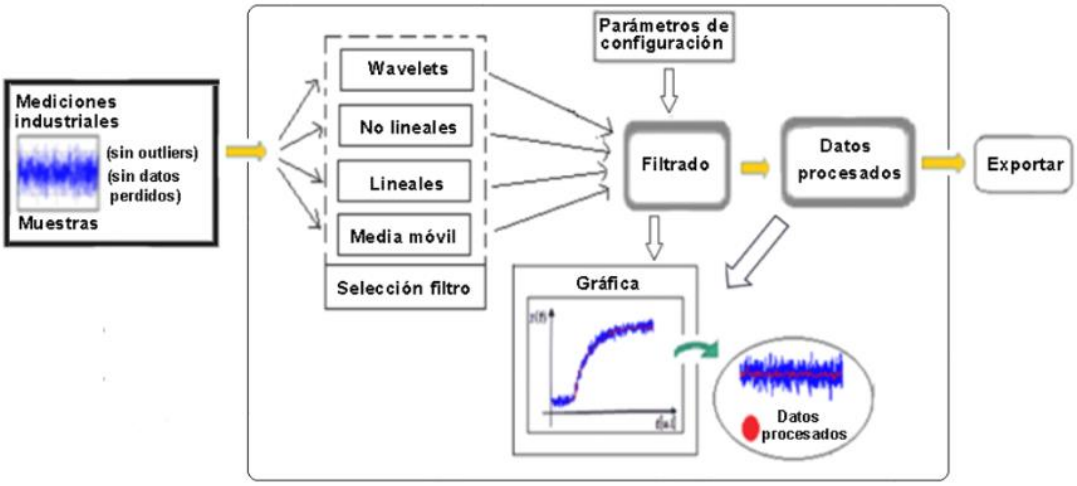

Figura 2.8 Diagrama funcional del filtrado en el ISLab.

Varias figuras destacadas en el campo del procesamiento de señales han demostrado la naturaleza multiescala de las series temporales provenientes de mediciones industriales. En consecuencia con esto y como prestación principal, el ISLab tiene implementado seis familias de wavelets, funciones de gran aceptación en la comunidad científica internacional que realizan un análisis multirresolución, a varias escalas, de las señales y que en los años recientes han probado su excelente desempeño en el campo del filtrado de señales.

2.4.2 Filtros no lineales.

Los filtros no lineales se caracterizan fundamentalmente porque no aplican un operador lineal a una señal variable en el tiempo, lo que quiere decir que la señal de salida no se calcula como la media ponderada (pesos) de la señal de entrada y anteriores muestras de salida. Generalmente se aplica en casos de señales con presencia de ruido no gaussiano (ruido impulsivo sobre todo). En la figura 2.9 Se muestra el diagrama general de funcionamiento de estos filtros, donde x(N) representa la señal de entrada y N la cantidad de muestras con i=1,2,…,N.

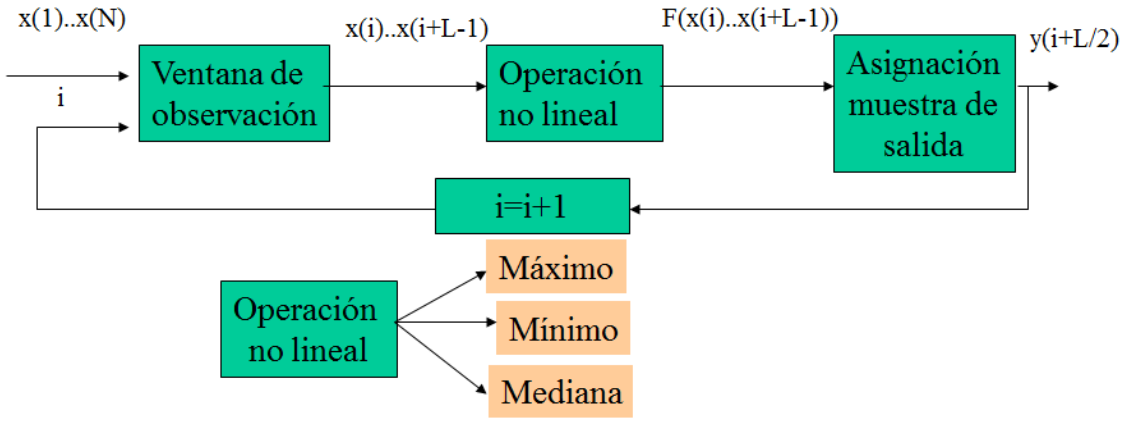

Figura 2.9 Diagrama general de funcionamiento de los filtros no lineales.

Estos filtros son más difíciles de usar y de diseñar que los lineales debido a que las herramientas matemáticas más potentes de análisis de señales (por ejemplo, la respuesta al impulso y la respuesta en frecuencia) no se pueden utilizar en ellos. Así, los filtros lineales se utilizan a menudo para eliminar el ruido y la distorsión creada por los procesos no lineales, simplemente porque el filtro no lineal adecuado es difícil de diseñar y construir.

De acuerdo al criterio de varios expertos el filtrado no lineal es mucho mejor que el filtrado lineal. El problema se podría englobar desde los siguientes puntos:

- El propósito es eliminar el ruido.
- El ruido es impulsivo.
- La señal y el ruido ocupan la misma banda de frecuencia.

El filtrado no lineal ha resultado ser satisfactorio en solventar esos problemas. A continuación, se hará un conciso resumen de los principales filtros digitales no lineales con los que cuenta el ISLab.

2.4.2.1 Filtro de mediana.

El ISLab pone en manos del operador un conjunto de filtros no lineales que dan la posibilidad de obtener no solo valores próximos a los reales sino también aquellos que bordean sus límites inferiores y superiores, lo cual podría resultar muy útil en dependencia del estudio que se vaya a realizar con los datos y que representa un aumento de las prestaciones que ofrece este software.

En el ámbito de la estadística, la mediana representa el valor de la variable de posición central en un conjunto de datos ordenados. El filtrado de mediana viene dado por la expresión:

$$y[m] = MED\{x[m-k], x[m-k+1], \dots, f[m], f[m+k-1], f[m+k]\} \quad (2.4)$$

donde m representa la posición en que se encuentra el valor central y k es la cantidad de valores que existen antes y después de éste de forma tal que MED toma el valor central después de la ordenación.

Existen dos métodos para el cálculo de la mediana:

1. Considerando los datos en forma individual, sin agruparlos.
2. Utilizando los datos agrupados en intervalos de clase.

En el caso del filtro de mediana implementado en el ISLab, se utilizó el segundo método para el cálculo de la mediana. Una vez conocida la longitud de la serie temporal se fracciona usando una ventana de ancho *n* (en cantidad de muestras) con la cual recorre la señal como se muestra en la figura 2.10 mientras se va realizando el cálculo en cada sección que van quedando almacenados en un arreglo. El ancho de la ventana móvil es una magnitud más intuitiva y comprensible para un usuario no experto en cuestiones de filtrad. El comando usado para la operación de cálculo de la mediana es:

```
Y(k)=median(a(k-n:k+n));      %Cálculo de la mediana en la ventana
```

Los restantes filtros no lineales implementados en la plataforma tienen una manera de proceder similar. Los filtros de máximo y de mínimo usan los comandos:

```
Y(k)=max(a(k-n:k+n));      %Cálculo de los máximos en la ventana

Y(k)=min(a(k-n:k+n));   %Cálculo de los mínimos en la ventana
```

Donde:

k-n y k+n: límites de la ventana
k: número de muestras
n: ancho de la ventana.

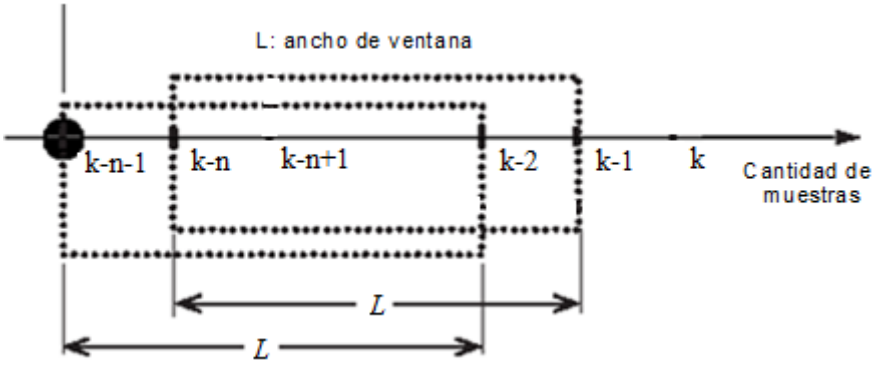

Figura 2.10 Metodología de trabajo del algoritmo de fraccionamiento (ventana móvil).

Los filtros de apertura y de cierre están conformados por la unión de un filtro de mínimo con uno de máximo en el caso del primero y la unión contraria da como resultado un filtro de cierre.

2.4.3 Filtros lineales.

Se dice que un filtro es lineal si se puede aplicar el principio de superposición. Es decir, supongamos que se tienen las entradas *f1(t)* y *f2(t)* que al pasar por el filtro por separado se convierten en *g1(t)* y *g2(t)* respectivamente. El filtro es lineal si se cumple que *f = a1f1(t) + a2f2(t)* se convierte en *g = a1g1(t) + a2g2(t)* después de pasar por el filtro. Una de las principales características de los filtros es conservar ciertas características de la señal eliminando otras. Dentro de las prestaciones que posee el ISLab están dos filtros clásicos que se han venido usando desde hace mucho tiempo y que no pasan de moda: FIR e IIR. Para los sistemas lineales la función de transferencia de estos filtros se define como:

$$H(Z) = \frac{\sum_{i=0}^{n} a_i z^{-i}}{1-\sum_{i=1}^{n} b_i z^{-i}} \quad \text{(sólo sistemas lineales)} \qquad (2.5)$$

Donde se tiene que si b_i=0 para todo i, el filtro es no recursivo (FIR) y si algún b_i Es no nulo, el filtro es recursivo (IIR).

2.4.3.1 Filtro FIR.

Dentro de los tres grandes bloques de métodos de diseño de filtros FIR con fase lineal (el enventanado, el muestreo en frecuencia y el rizado constante) en este trabajo se emplea el método de enventanado el cual se basa en acotar la respuesta ante impulso infinita de un filtro ideal. Como los términos de la respuesta ideal y las

ventanas son simétricos respecto al punto central, el filtro presenta fase lineal. El procedimiento es el siguiente:

- Obtener la respuesta ante impulso del filtro ideal deseado
- Enventanar (truncar) dicha respuesta.
- Desplazar la respuesta ante impulso enventanada un número adecuado de muestras para hacerla causal.

La ventana estará definida como:

$$w(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & n \ge N \end{cases} \quad (2.6)$$

y su expresión en el dominio Z es:

$$W(z) = 1 + z^{-1} + \ldots + z^{-N+2} + z^{-N+1} = \frac{1-z^{-N}}{1-z^{-1}} \quad (2.7)$$

con lo que su respuesta en frecuencia sería:

$$W(\omega) = e^{-j\omega\left(\frac{N-1}{2}\right)} \frac{\mathrm{Sin}\left(\frac{N\omega}{2}\right)}{sin\left(\frac{\omega}{2}\right)} \quad (2.8)$$

Estabilidad, fácil diseño y fase lineal son las principales propiedades del filtro FIR. Para la programación del filtrado se utilizó la función del Matlab *filter* que usa un filtro digital que trabaja para las entradas reales y complejas. La descripción de su relación entrada salida en el dominio de la transformada Z está dada por la ecuación 2.9 donde $Y(Z)$ Salida y $X(Z)$entrada:

$$Y(Z) = \frac{b(1)+b(2)+\cdots+b(nb+1)z^{-nb}}{1+a(2)z^{-1}+\cdots+a(na+1)z^{-na}} X(Z) \quad (2.9)$$

Para la obtención de los coeficientes, una vez que el operador introduce los parámetros de configuración, se utiliza la función *fir1* que recoge esta configuración y si será pasa banda, pasa alto o pasa bajo (como es el caso).

2.4.3.2 Filtro IIR.

Uno de los filtros IIR más conocidos y ampliamente usados es el filtro de Butterworth (figura 2.11) el cual se obtiene imponiendo que la respuesta en magnitud del filtro sea máximamente plana en la banda pasante y en la banda no pasante. Esto quiere decir que las (2N-1) primeras derivadas de $|H(f)|^2$ Son cero para $f = 0$ y para $f = \infty$. Solo contienen polos y tienen la función de transferencia:

$$|H(f)|^2 = \frac{H_{max}^2}{1+\left(\frac{f}{f_c}\right)^{2n}} \quad (2.10)$$

donde n es el orden del filtro y f_c La frecuencia de corte (caída de 3 db respecto de la banda pasante).

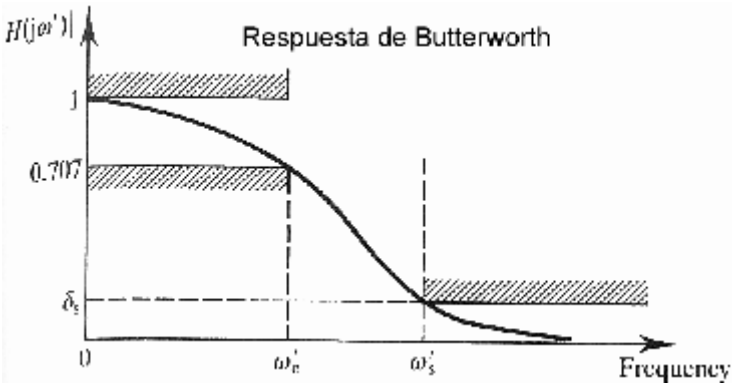

Figura 2.11 Respuesta en frecuencia del filtro de Butterworth.

Los filtros Butterworth son fáciles de construir porque los valores resultantes de los componentes son más prácticos que la mayoría de los otros tipos, y las variaciones de los componentes son menos críticas. Este filtro es el único que mantiene su forma para órdenes mayores (sólo con una caída de más pendiente a partir de la frecuencia de corte). Para su implementación se utilizó la función *butter*.

2.4.3.3 Filtro de media móvil.

Un filtrado de media, queda encuadrado por defecto como un filtro lineal. Para un filtro de media de una dimensión de longitud N = 2 k + 1, su salida para un instante j, queda definido mediante:

$$y[m] = \frac{1}{2k+1}\sum_{j=-k}^{k} x(m+j) \quad (2.11)$$

El filtro de media móvil diseñado básicamente consiste en fijar un ancho de ventana e ir calculando el valor medio de la variable dentro de esa ventana hasta llegar al final de la muestra y guardar los valores calculados dentro de un arreglo. Aunque presenta su equivalencia, este método tiene la ventaja de que no se necesitan configurar los parámetros de los coeficientes del numerador y denominador del polinomio de un filtro tradicional (como cuando se usa el comando *filter* del Matlab™). El comando usado es:

Y=mean(X) %Retorna el valor medio de los elementos del arreglo X.

2.4.4 Wavelets.

La transformada wavelet discreta (DWT) es capaz de entregar suficiente información tanto para el análisis como para la reconstrucción de una señal con una significativa reducción del tiempo de procesamiento, además, es mucho más fácil de implementar que la CWT. La transformada wavelet estima la potencia del ruido que presenta la señal y de esta forma realiza de manera más precisa la reducción del nivel del ruido contenido en la señal.

Para ser útil, la teoría de wavelets debe disponer de algoritmos rápidos para su uso en computadores, es decir, un método para encontrar los coeficientes wavelet y para reconstruir la función que representan. Existe una familia rápida de algoritmos basados en el análisis multirresolución que junto a los algoritmos piramidales, se desarrollaron para descomponer señales de tiempo discreto. La idea es obtener una representación tiempo-escala de una señal discreta. En este caso, filtros con distintas frecuencias de corte son usados para analizar la señal en diferentes escalas. Estas operaciones cambian la resolución de la señal, y la escala se cambia mediante operaciones de interpolación y submuestreo. Por tanto, la representación tiempo-escala de una señal digital se obtiene usando técnicas de filtrado digital.

A continuación se explica el funcionamiento de estas funciones para el filtrado de señales y las principales características de las familias wavelets que brinda al usuario el ISLab, fundamentalmente las wavelets Daubechies.

2.4.4.1 Proceso de descomposición de la señal.

La DWT analiza la señal descomponiéndola en una aproximación y en un detalle (nivel), considerando diferentes bandas de frecuencias con distintas resoluciones para cada nivel. La descomposición de la señal en diferentes bandas de frecuencia se obtiene mediante un sucesivo filtrado de pasa bajo y pasa alto, por lo tanto la señal original x[n] se pasa a través de un filtro pasa alto g[n] y de un filtro pasa bajo h[n], ambos de media banda con respuesta al impulso. El filtrado de la señal corresponde a la operación matemática de convolución de ésta con g[n] y h[n] lo cual se define como:

$$x[n] * g[n] = \sum_{-\infty}^{\infty} x[k]g[n-k] \quad (2.12)$$

$$x[n] * h[n] = \sum_{-\infty}^{\infty} x[k]h[n-k] \quad (2.13)$$

Estos filtros eliminan las componentes frecuenciales situadas por debajo y por encima de la mitad del ancho de banda de la señal respectivamente.

Una propiedad importante de la DWT es la relación entre las respuestas impulso de los filtros paso alto y paso bajo. Estos filtros no son independientes entre sí y están relacionados a través de la siguiente ecuación:

$$g[L-1-n] = (-1)^n h[n] \quad (2.14)$$

Donde g[n] es el filtro paso alto, h[n] es el filtro paso bajo y L es la longitud del filtro expresada en número de puntos. La conversión de paso bajo a paso alto se hace a través del factor $(-1)^n$, los filtros que satisfacen esta característica se conocen como filtros espejos en cuadratura (QMF).

Después de este filtrado pueden eliminarse la mitad de las muestras de acuerdo a la regla de Nyquist, ya que la señal ahora tiene una frecuencia superior de $\pi/2$ radianes en vez de π, para ello se eliminan una de cada dos muestras (submuestreo por dos). De esta manera se ha constituido el primer nivel de descomposición, lo que matemáticamente puede expresarse como:

$$y_{high}[k] = \sum_n x[n] * g[2k-n] \quad (2.15)$$

$$y_{low}[k] = \sum_n x[n] * h[2k-n] \quad (2.16)$$

Donde $y_{high}[k]$ Y $y_{low}[k]$ Son las salidas de los filtros pasa alto y pasa bajo, respectivamente, después del submuestreo por dos.

La figura 2.12 muestra un ejemplo de este procedimiento, donde x[n] es la señal original que se va a descomponer y h[n] y g[n] son los filtros pasa bajo y pasa alto, respectivamente. En cada nivel de descomposición el ancho de banda de la señal aparece señalado en la figura como “f”.

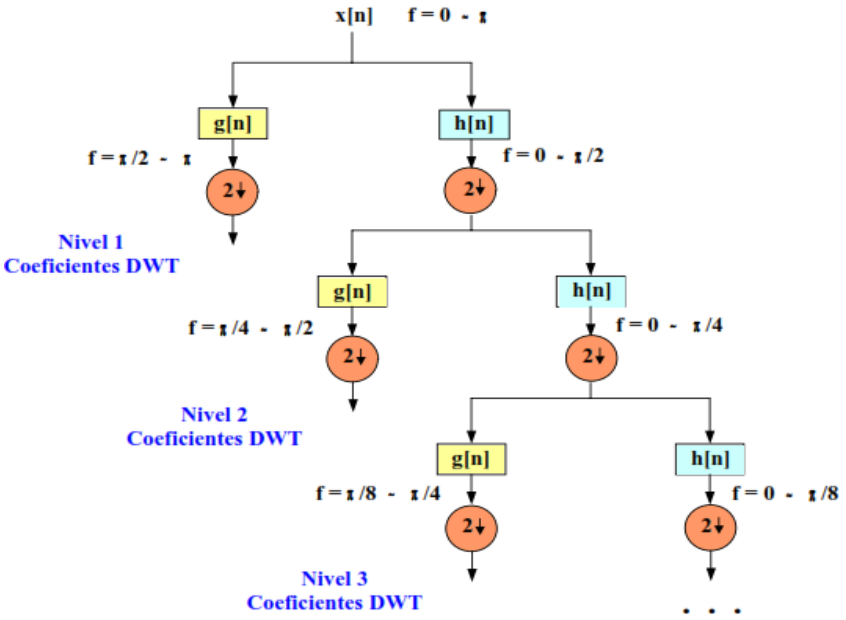

Figura 2.12 Diagrama de codificación de sub-bandas.

En resumen, el procedimiento descrito ofrece una buena resolución en el tiempo para las altas frecuencias y una buena resolución en frecuencia para las bajas frecuencias.

2.4.4.2 Reconstrucción de la señal.

La reconstrucción en este caso es muy simple, dado que los filtros de banda media forman una base ortonormal, para estos efectos el procedimiento anteriormente descrito se sigue en sentido inverso, de este modo la señal es interpolada por dos y pasada a través de los filtros de síntesis g'[n] y h'[n], pasa alto y pasa bajo respectivamente, para posteriormente sumarse ambas salidas.

Las familias más famosas son las desarrolladas por Ingrid Daubechies y que se conocen como las wavelets de Daubechies. El proceso de reconstrucción es similar al de descomposición. La señal en cada nivel es interpolada por dos, pasada por dos filtros de síntesis representados por los operadores G, H (pasa alto y pasa bajo, respectivamente), y entonces las respectivas salidas son sumadas.

Un ejemplo de descomposición se muestra en la figura 2.13, con la aproximación, detalles, y la señal original. Los primeros dos detalles contienen principalmente ruido, mientras que los de mayor nivel aproximan la señal. En este ejemplo se ha utilizado la Wavelet Daubechies 8, con nivel de descomposición 5.

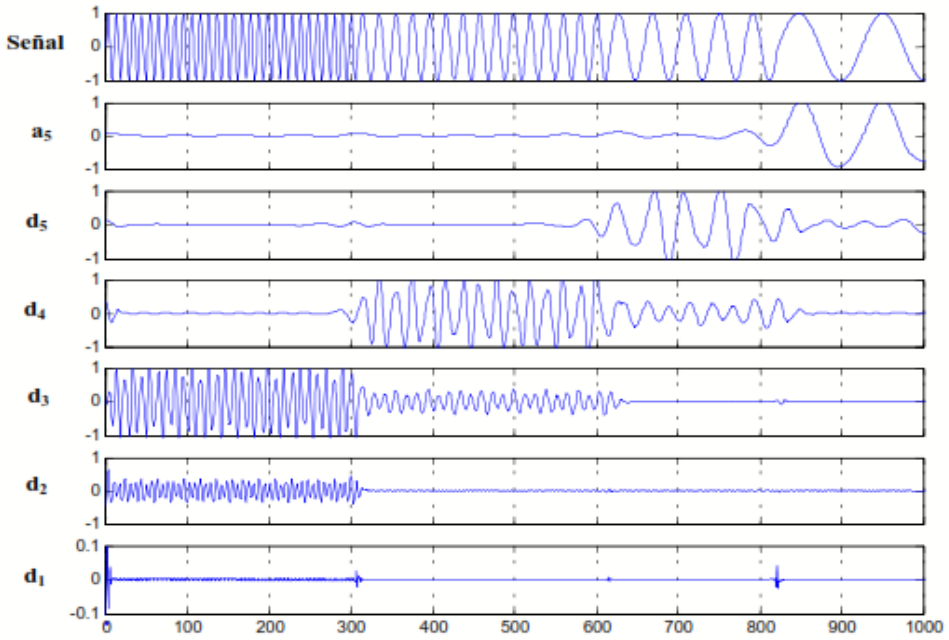

Figura 2.13 Ejemplo de descomposición DWT.

2.4.4.3 Wavelets madre. Daubechies.

El ISLab tiene implementado una variada selección de variantes Wavelets entre las que se destacan las familias Symlets, las Coilflets y las Daubechies. Su eficiencia en representar señales es una clave para su utilidad con problemas como compresión de datos y eliminación de ruido y además han probado ser potentes en facilitar ciertos tipos de problemas computacionales. El término madre da a entender que las funciones con diferentes regiones de actuación que se usan en el proceso de transformación provienen de una función principal o wavelet madre. Es decir, la wavelet madre es un prototipo para generar las otras funciones ventanas.

La wavelet madre se elige entre un conjunto de funciones y cumple el papel de prototipo para todas las ventanas que participan en el proceso, puesto que todas estas ventanas son las versiones dilatadas (o comprimidas) y desplazadas de la wavelet madre. En la figura 2.14 se presentan algunas wavelets madres clásicas.

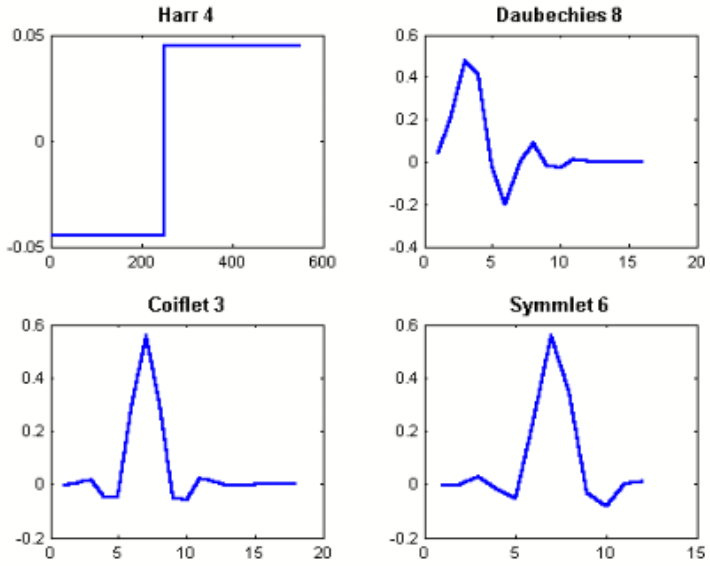

Figura 2.14 Algunas wavelets madre definidas según un eje de tiempo continuo. El número indica la cantidad de momentos nulos.

Dentro de las variedades de transformada wavelet utilizadas para el análisis de señales existe una en particular que desde su descubrimiento por la matemática Ingrid Daubechies ha sido una herramienta de ensueño para los procesadores de señales. Estas wavelets no sólo son ortogonales sino que también se pueden implementar mediante sencillas ideas de filtrado digital, de hecho, mediante cortos filtros digitales. Puede tener un orden N, donde N es un número entero positivo y determina el número de coeficientes de filtro que tiene la wavelet.

Con wavelets, el ruido puede ser eliminado de un gran número de tipos de señales, incluyendo aquellas con saltos, picos y otras ocurrencias no demasiado suaves. La eliminación de ruido por wavelets es superior a las técnicas tradicionales, que eliminan el ruido por medio de filtrados paso bajo, emborronando las zonas abruptas de la señal.

Para la implementación de las familias wavelets con que cuenta el ISLab se emplearon los siguientes comandos:

[c,l] = wavedec(senal,nivel,tipo); %ejecuta una descomposición multinivel.
Sigma = wnoisest(c,l,1); % estima std del ruido a partir de coeficientes nivel 1.
[thr,sorh,keepapp] = ddencmp('den','wv',senal); %valores default para denoising.
[xrec,CXD,LXD] = wden(senal,'sqtwolog',sorh,'one',nivel,tipo); % denoising automático nivel 1, en xrec queda la señal filtrada.

2.5 Generación de señales para experimentación en el Simulink.

Para la comprobación de la efectividad de todos los algoritmos antes mencionados se hace necesario disponer de señales generadas por simulación de las cuales se conozcan de antemano sus características y los resultados que se debía esperar al procesarlas. Para esto se generaron una serie de señales usando el Toolbox Simulink de Matlab con características muy similares a las procedentes de mediciones de variables en procesos térmicos. Se construyen los modelos utilizando los bloques que provee Simulink o los creados por el usuario. Se puede crear virtualmente cualquier sistema dinámico interconectando los bloques de la forma correcta.

En este trabajo se modelarán ondas bases que se puedan emplear para determinar la efectividad del ISLab en el procesamiento de datos. Entre los bloques fundamentales que se toman para generar los modelos se encuentran:

- ✓ Constant: Genera un valor constante.
- ✓ Signal Generator: Genera diferentes formas de onda.
- ✓ Step Input: Genera una función escalón.
- ✓ Ramp Input: Genera una función rampa.
- ✓ Scope: Visualiza señales en ventanas de figura MATLAB (con el escalado que se le indique).
- ✓ Sum: comparador (sumador con signo).

- ✓ Transfer Fcn: Implementa una función de transferencia lineal.

Una vez que se tienen las ondas bases, con el fin de representar señales procedentes de un determinado proceso, estas son contaminadas con ruido blanco o coloreado. Para contaminar la onda base con ruido blanco basta con sumar la señal generada con el bloque Band-Limited White Noise. La generación de RC se realiza inyectando RB a un filtro ARMA (1,1) representado por la siguiente función de transferencia:

$$F(z) = \frac{1+\Theta_1 z^{-1}}{1+\Phi_1 z^{-1}} \qquad (2.17)$$

Donde Θ_1 es el parámetro de la componente de media móvil del filtro y Φ_1 es el parámetro de su componente autorregresiva.

Las distintas combinaciones de bloques dan a lugar a diferentes salidas y formas de onda. Además estos bloques tienen determinados parámetros los cuales una vez se varían influyen en sus salidas dando a lugar a una gran gama de posibilidades en la generación de series temporales.

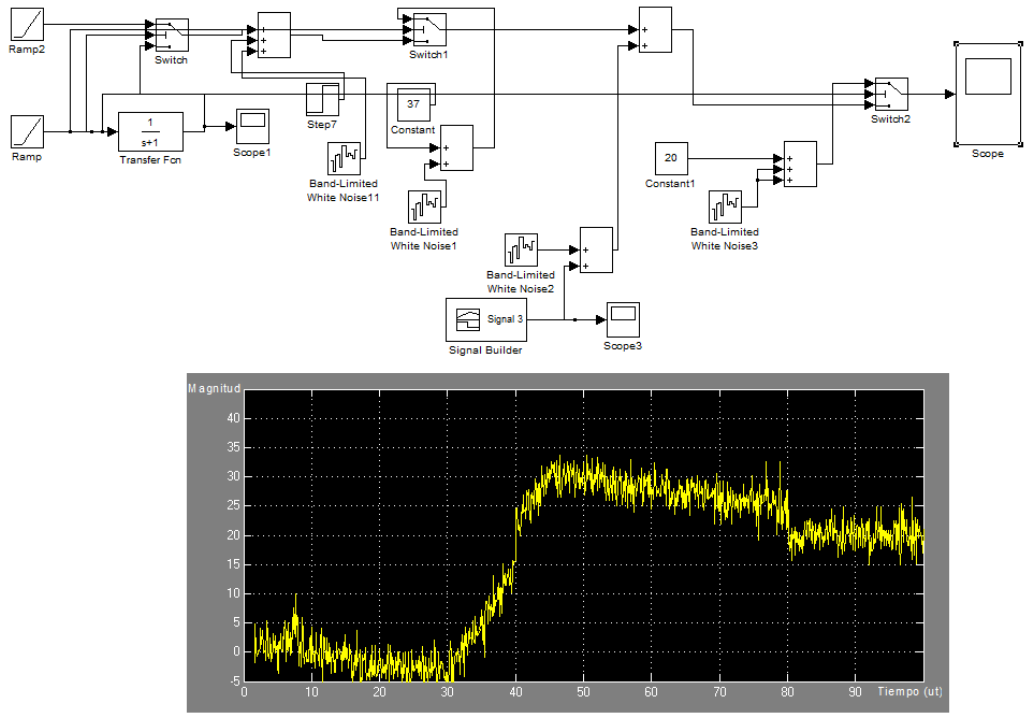

Figura 2.15 Ejemplo de generación de señales empleando el Simulink. La señal está contaminada con ruido blanco.

2.6 Análisis del desempeño de los algoritmos de detección y tratamiento de outliers y valores perdidos.

Para comprobar la efectividad de los algoritmos empleados por el ISLab para la detección de los outliers y de los datos ausentes, así como su comportamiento en función de brindar un tratamiento apropiado y preciso a las señales con estos problemas se tomaron cuatro series temporales contaminadas con ruido blanco con distintas características y desarrollo en el tiempo las cuales fueron generadas en el Simulink (ver epígrafe 2.5). A continuación se verán las principales características y desempeño de los métodos en el procesamiento de estas señales.

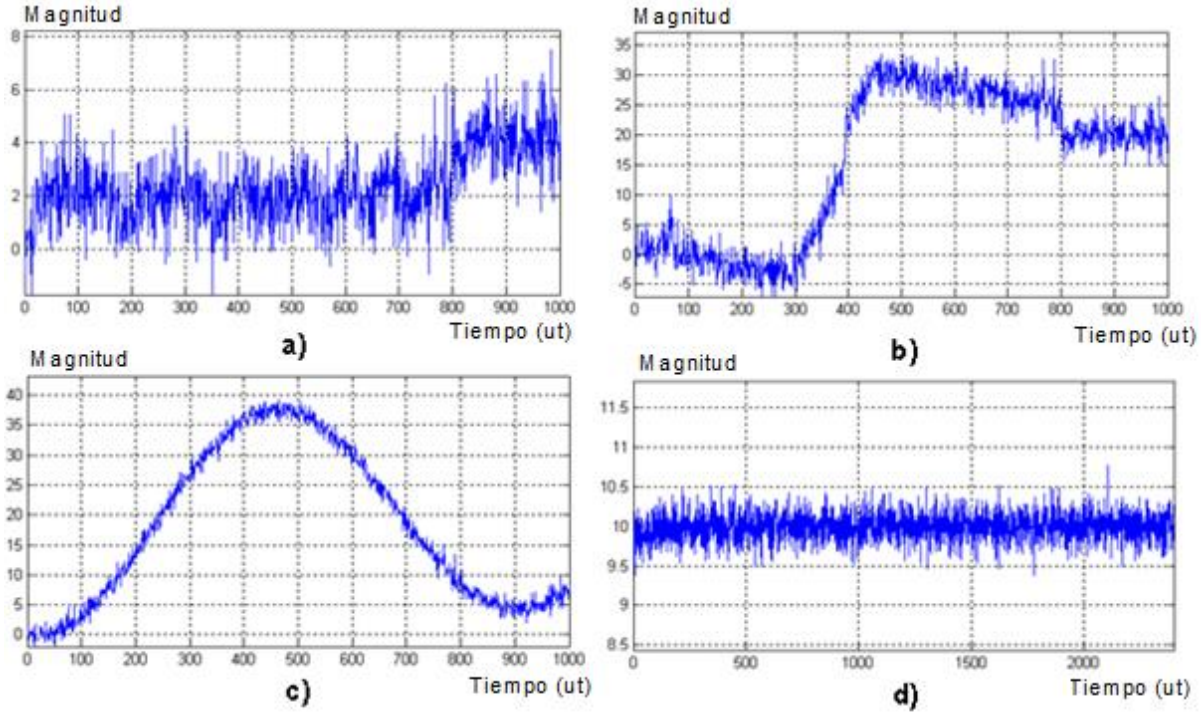

Figura 2.16 Señales empleadas: a) señal V, b) señal X, c) señal Y, d) señal Z.

2.6.1 Algoritmo de detección y tratamiento de outliers.

Para aumentar la fiabilidad de los resultados se tomaron cuatro configuraciones distintas de los parámetros de ajuste del método implementado para la detección y eliminación de los outliers. El ancho de la ventana móvil se tomó igual a diez (10) para lograr una mayor precisión y se tomaron cuatro órdenes o valores de *n*, de dos (2) a cinco (5), los cuales determinan la diferencia que deben tener las muestras respecto a la media de los datos para que éstas sean consideradas valores atípicos.

En cada una de las señales de prueba se cambiaron intencionalmente los valores de algunas muestras por diferentes magnitudes que representan outliers para verificar

luego que éstos sean detectados por el método. Para comprobar la efectividad del algoritmo de imputación de datos, una vez eliminadas las anomalías y sustituidas por valores razonables mediante interpolación, se compara la varianza de la data original (sin cambios en las muestras) con la que se obtenga una vez se hayan rellenado los espacios que ocupaban los outliers. De acuerdo al grado de similitud entre los valores de varianza será el rendimiento del método implementado.

A la señal V se le incorporaron outliers en las posiciones 3, 120 (g), 309 (g), 310 (g), 438, 564, 578, 885 (g); a la señal X se le adicionaron 8 outliers en los siguientes lugares: 103 (g), 177, 220, 292 (g), 518 (g), 843 (g), 867 (g), 909, a la señal Y en las posiciones 14, 59, 100 (g), 215 (g), 278, 317 (g), 435 (g), 590 (g), 639, 735 y 836 (g) mientras que la señal Z sufrió modificaciones en las posiciones: 1621 (g), 1780 (g), 1831 (g), 2107 (g) y 2173 (g). Las posiciones clasificadas con una (g) son aquellas que representan datos atípicos grandes.

2.6.1.1 Evaluación del método.

La figura 2.17 muestra la detección de los outliers para *n*=3 en las cuatro señales modificadas y la tabla 2.1 contiene la cantidad de datos anómalos detectados para cada orden.

TABLA 2.1 Resultados del método de detección y tratamiento de outliers.

Factor de multiplicación (n)	Señal	Outliers	Detectados
2	V	8	26
	X	8	31
	Y	11	15
	Z	5	84
3	V	8	9
	X	8	9
	Y	11	12
	Z	5	5
4	V	8	6
	X	8	6
	Y	11	8
	Z	5	5
5	V	8	4
	X	8	5
	Y	11	6
	Z	5	5

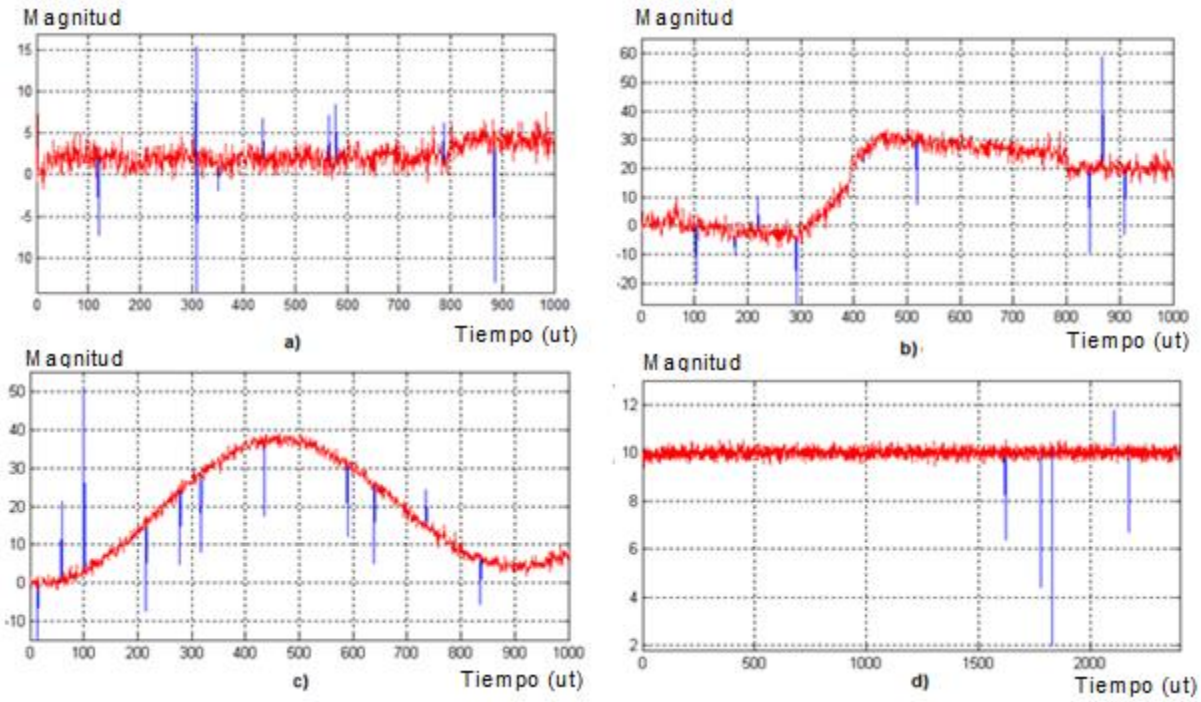

Figura 2.17 Outliers detectados para *n*=3 en: a) señal V, b) señal X, c) señal Y y d) señal Z.

Como se puede observar en la tabla 2.1 todos los outliers que tenían las señales fueron detectados para *n*=3 mientras que para *n*=2 se tomaron como anomalías una gran cantidad de muestras, las cuales, en su mayoría no constituyen errores gruesos. Este efecto viene dado por el hecho de que el valor crítico es menor por lo que el algoritmo además de detectar los valores anómalos también estaría viendo aquellas mediciones que son consideradas errores aleatorios. Por otra parte cuando se evalúa el método para *n*=4 o un orden mayor apenas son detectados algunos outliers. Esto se debe a que aumentó el valor crítico por lo que solo fueron detectados aquellos que eran lo suficientemente grandes y distantes de la media aritmética del conjunto de datos, tal y como se esperaba aquellos que antes fueron clasificados como (g). Como se puede apreciar, de acuerdo al factor de multiplicación que se emplee será el resultado del método por lo que es aconsejable que el valor por defecto (3) solo sea modificado por un especialista o en caso de que se desee realizar un determinado estudio que exija el cambio de este parámetro.

Por otro lado, al comparar los valores de las varianzas se pudo comprobar que la diferencia es mínima por lo que el método demostró ser eficiente en la solución de la detección y tratamiento de los valores atípicos.

2.6.2 Algoritmo de detección y tratamiento de datos perdidos.

El procedimiento para analizar la efectividad del algoritmo de detección y tratamiento de los valores perdidos fue similar al anterior con la diferencia de que en lugar de plantar outliers se eliminaron algunas muestras o se cambiaron por caracteres no numéricos. La comprobación del método fue del mismo modo. El ancho de ventana escogido fue diez (10).

A la señal V se le generaron valores ausentes en las posiciones 5, 42, 43, 44, 142, 508, 593, 678, 740, 850, 851, 925, 968; a la señal X se le adicionaron en los siguientes posiciones: 27, 82, 142, 337, 409, 529, 530, 531, 607, 694, 837, a la señal Y en las posiciones 12, 139, 244, 245, 544, 625, 766 y 928 y la señal Z sufrió cambios en los lugares: 10, 85, 167, 244, 340, 546, 980, 1657, 2000, 2367. De esta manera se obtuvieron cuatro señales con distintos valores perdidos cada una.

2.6.2.1 Evaluación del método.

La figura 2.18 muestra la detección de los datos ausentes en las cuatro señales modificadas y la tabla 2.2 contiene la cantidad de datos perdidos detectados y eliminados.

Todos los lugares vacíos que fueron generados en las señales fueron correctamente detectados y se comprobó que los datos imputados prácticamente no alteraron el valor de la varianza que poseía la data original por lo que se considera que este método cumple eficazmente con su objetivo.

TABLA 2.2 Resultados del método de detección y tratamiento de valores perdidos.

Señal	Datos perdidos	Detectados	Eliminados
V	13	13	13
X	11	11	11
Y	8	8	8
Z	10	10	10

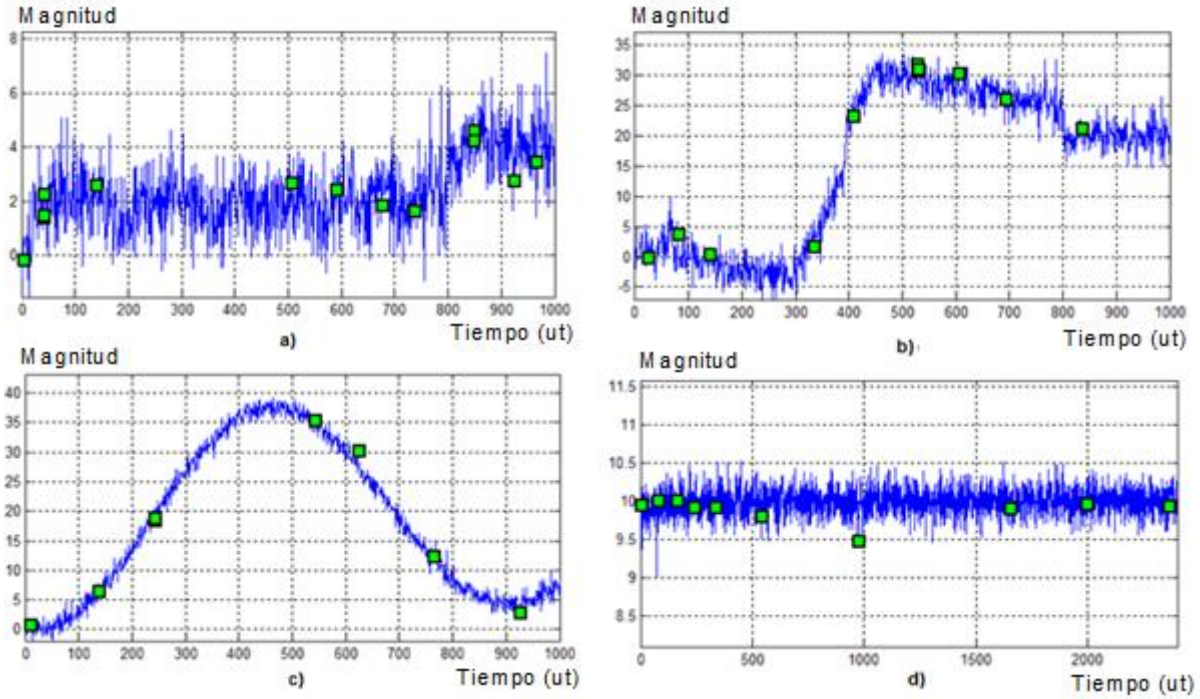

Figura 2.18 Datos perdidos detectados en: a) señal V, b) señal X, c) señal Y y d) señal Z.

2.7 Análisis del desempeño de las técnicas de reducción de ruido implementadas en el ISLab.

En el presente epígrafe se realiza un análisis y comparación de los diferentes métodos de supresión de ruido que brinda el ISLab comprobando su rendimiento ante señales industriales reales y simuladas. En particular se analiza el comportamiento de los métodos de un filtro IIR (específicamente un filtro de Butterworth), un filtro FIR mediante enventanado, los filtros de mediana y de media móvil y una técnica Wavelet para la reducción de ruido gaussiano en estas señales. A tales efectos se utilizó un conjunto de señales patrón con características similares a las de un generador de vapor.

2.7.1 Selección de parámetros.

En la gran mayoría de los procesos industriales en la actualidad se almacenan en grandes servidores los valores de las mediciones instrumentales de cada una de las variables que intervienen en los diferentes procesos. Ésta sería la fuente de procedencia real de los datos con ruido a filtrar. Por lo general, estas muestras se toman cada un segundo, incluso hay variables de lento desarrollo como la temperatura que se puede muestrear en periodos de tiempo más amplios pero generalmente los sistemas se configuran para realizar la toma de muestras de todas las variables cada un segundo que da un margen lo suficientemente aceptable como

para no perder información sobre la dinámica incluso de las variables más veloces. Para probar la efectividad de los métodos, en este trabajo se generan señales mediante Matlab Simulink de las cuales se conocen su forma de onda base (antes de agregarle ruido) y se considera esta misma frecuencia de muestreo $f_s = 1\,Hz$.

El orden del filtro de Butterworth se toma $n = 4$, en el caso del FIR que requiere coeficientes de órdenes mucho mayor se toma $n = 500$ y de manera común para ambos se toma como frecuencia de corte $f_c = 0.07$. La magnitud del ancho de ventana escogido para los filtros de mediana y de media móvil fue de diez, un valor pequeño que garantiza una mayor precisión en la estimación de los valores reales de la señal. En el caso del método Wavelet, que sigue una filosofía diferente, se utiliza la función madre Daubechies 4 (db4) con nivel de descomposición igual a cinco.

2.7.2 Generación de las señales.

En la definición de las formas de ondas base para las señales sintéticas, se tuvo en cuenta los regímenes en la operación en los generadores de vapor que transcurren entre transiciones que pueden ser modeladas por exponenciales y rampas. En este caso se usa una exponencial generada como la respuesta de un sistema lineal de primer orden con constante de tiempo de 200 segundos a una excitación de cambio en escalón. La señal básica es generada usando el modelo Simulink. Este corre a un paso fijo de una unidad de tiempo produciendo 4467 muestras. Su constante de tiempo es de T=200 ut.

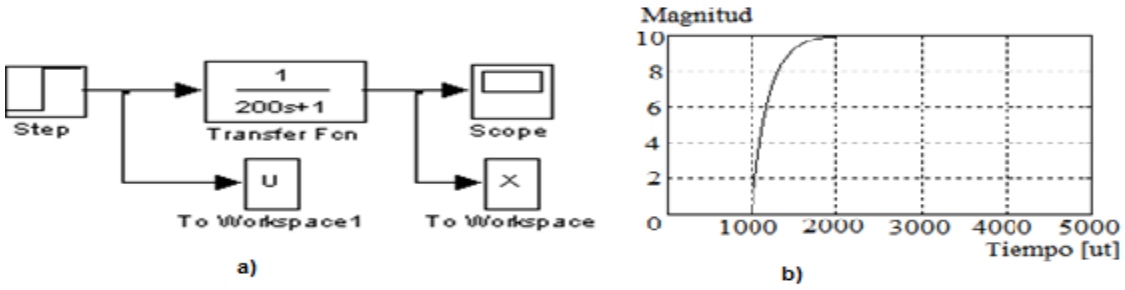

Figura 2.19 a) Modelo Simulink para generar la señal base. Sistema de primer orden con T=200 s, b) respuesta a un escalón de diez unidades.

Las señales de ensayo finales se obtienen al contaminar la onda base con ruido de diferentes características. Para ello, se genera una base de datos con vectores de

ruidos que contiene RB y RC de distribución normal, con los siguientes valores de varianza: $\Sigma^2 = 0.01$, $\Sigma^2 = 0.03$, $\Sigma^2 = 0.06$, $\Sigma^2 = 0.09$, $\Sigma^2 = 0.12$, $\Sigma^2 = 0.25$ $y \Sigma^2 = 0.50$.

La base de datos final está compuesta por 14 vectores, de ellos siete contaminados con RB y siete con RC.

2.7.3 Resultados y discusión.

Partiendo de que se conoce la señal base (sin ruido), las métricas que se utilizaron para la evaluación de los algoritmos fueron el MSE y la SNR, descritos en el epígrafe 1.5.

Los resultados del desempeño de cada método para cada uno de los vectores de RB y RC generados se muestran en la tabla 2.3. En negritas están resaltados los mejores resultados. Como se puede notar a medida que aumenta la varianza del ruido, la reducción del mismo se hace más ineficiente, tendencia esta que siguen los cinco algoritmos pero siempre sobresale como más efectivo el de la transformada Wavelet y la diferencia con los demás es más significativa en el caso del RC.

El filtro de mediana es el segundo con un comportamiento más acertado, lo que demuestra la eficiencia de los filtros no lineales, aunque solo se diferencia ligeramente con el de media móvil. Seguido de éstos se encuentra el filtro FIR el cual es levemente superior al filtro de Butterworth. Queda demostrada la fortaleza de la Transformada Wavelet Discreta en el procesamiento de señales de origen industrial que en su mayoría son señales que están autocorrelacionadas (RC).

TABLA 2.3 Resultados de los métodos para RB Y RC

Algoritmo	Ruido	MSE(RB)	SNR(RB)	MSE(RC)	SNR(RC)
Wavelet	0.01	**0.00047**	**51.717**	**0.00113**	**47.971**
Butterworth		0.00175	46.053	0.00360	42.940
FIR		0.00165	46.322	0.00352	43.030
Mediana		0.00089	48.979	0.00180	45.953
Media Móvil		0.00099	48.536	0.00225	44.966
Wavelet	0.03	**0.00102**	**48.380**	**0.00257**	**44.401**
Butterworth		0.00480	41.687	0.01026	38.392
FIR		0.00463	41.849	0.01014	38.445
Mediana		0.00213	45.208	0.00449	41.979
Media Móvil		0.00298	43.755	0.00672	40.232
Wavelet	0.06	**0.00239**	**44.722**	**0.00628**	**40.523**
Butterworth		0.00994	38.531	0.02152	32.177
FIR		0.00950	38.725	0.02111	35.260
Mediana		0.00503	41.483	0.00975	38.615
Media Móvil		0.00627	40.531	0.01438	36.928
Wavelet	0.09	**0.00284**	**43.967**	**0.00742**	**39.798**

Butterworth		0.01391	37.069	0.02984	33.757
FIR		0.01332	37.260	0.02928	33.839
Mediana		0.00630	40.508	0.01271	37.462
Media Móvil		0.00866	39.128	0.01948	35.609
Wavelet	0.12	**0.00376**	**42.746**	**0.00987**	**38.559**
Butterworth		0.01850	35.833	0.03974	32.513
FIR		0.01774	36.015	0.03903	32.591
Mediana		0.00837	39.278	0.01691	36.223
Media Móvil		0.01155	37.879	0.02597	34.359
Wavelet	0.25	**0.00896**	**38.982**	**0.02433**	**34.643**
Butterworth		0.03733	32.784	0.08094	29.424
FIR		0.03592	32.951	0.07969	29.491
Mediana		0.01856	35.819	0.03967	32.521
Media Móvil		0.02374	34.750	0.05417	31.168
Wavelet	0.50	**0.01574**	**36.535**	**0.04253**	**32.218**
Butterworth		0.07401	29.813	0.16016	26.460
FIR		0.07120	29.980	0.15750	26.533
Mediana		0.03893	32.602	0.07522	29.742
Media Móvil		0.04600	31.877	0.10498	28.295

2.8 Desempeño del ISLab ante señales reales tomadas del generador de vapor # 2 de la CTE “Lidio Ramos” FELTON.

Para la comprobación de la efectividad de los métodos programados en el ISLab ante señales reales se utilizarán mediciones tomadas del generador de vapor de la unidad # 2 de la de la Central Termoeléctrica “Lidio Ramos” FELTON. Esta fábrica cuenta con dos unidades de generación eléctrica de 250 MW cada una. Fue construida por la empresa checa SESTLMACE (Slovenske Energeticke Strojarne) y puesta en operación para el sistema eléctrico cubano en noviembre del 2001.

Se eligen un conjunto de variables de las más representativas de la dinámica del sistema como son la potencia activa de generación y el flujo de combustible. Cada señal está compuesta por un total de 40000 muestras (figura 2.20).

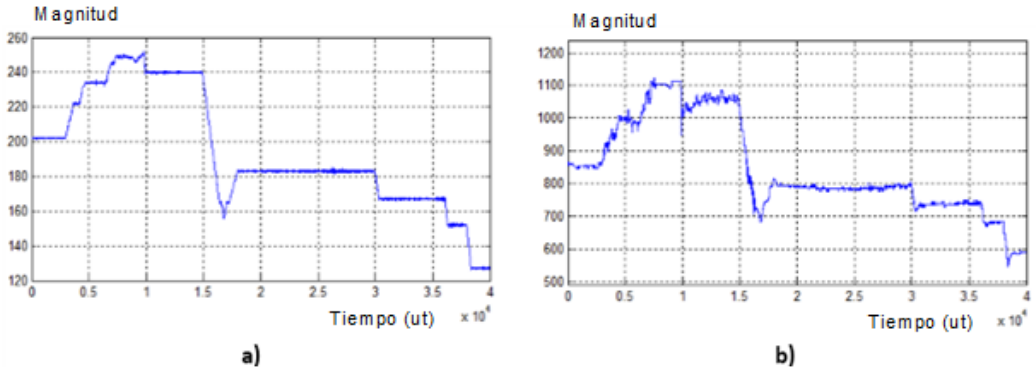

Figura 2.20 Mediciones sin procesar: a) potencia activa, b) flujo de petróleo.

Para la detección y eliminación de outliers se escogió un factor de multiplicación equivalente a tres y un ancho de ventana igual a diez con lo cual se detectaron y fueron suprimidos un total de 120 muestras que fueron consideradas anomalías en el caso de la potencia activa y 104 en la señal del flujo de combustible. Una vez que las señales pasaron por la primera etapa de procesamiento que realiza el ISLab se pasó a la próxima en la que no fueron detectados valores perdidos por lo que en esta etapa no se hizo necesario imputar valores admisibles en las datas. La longitud de la ventana móvil empleada en este algoritmo para el tratamiento de los datos ausentes fue igual al utilizado anteriormente. En la última etapa para la reducción del ruido se tomó la técnica que mejor resultados dio en las pruebas que se le hicieron a los distintos filtros: Wavelet Daubechies 4 (db4) con un nivel de descomposición igual a cinco.

Una vez procesados estos datos en el **Lab**oratorio de **S**eñales **I**ndustriales se obtuvo un juego de señales suavizadas, sin presencia de muestras erróneas, consistente, series temporales más próximas a las mediciones ideales que se esperan obtener en todo proceso, aquellas que no presentan problemas de calidad (figura 2.21).

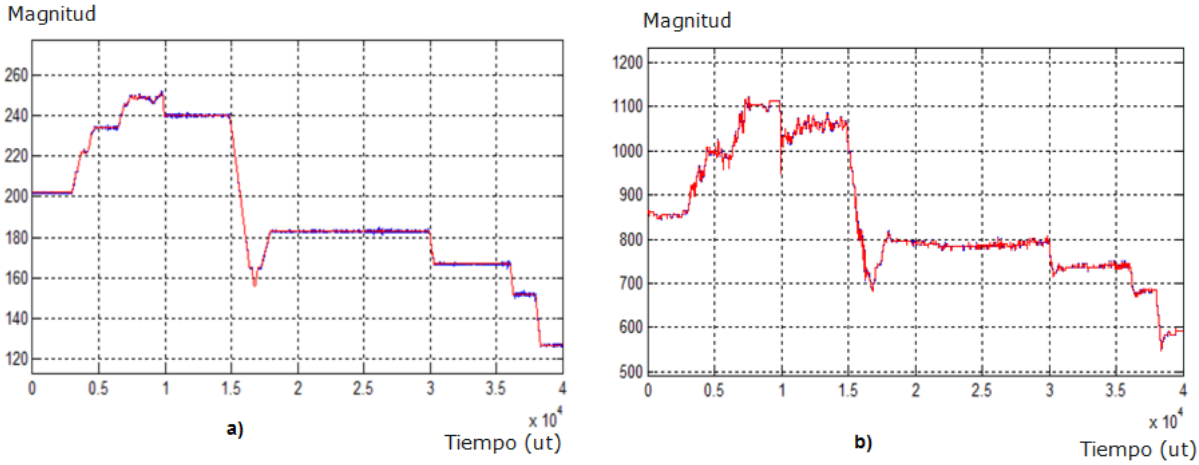

Figura 2.21 a) Potencia activa y b) flujo de petróleo (en azul señal original y en rojo la filtrada).

En la figura 2.21 b) se hace difícil distinguir la señal original debido a que las señales están conformadas por una elevada cantidad de muestras, la figura 2.22 muestra un acercamiento a la gráfica donde ambas señales se distinguen un poco mejor.

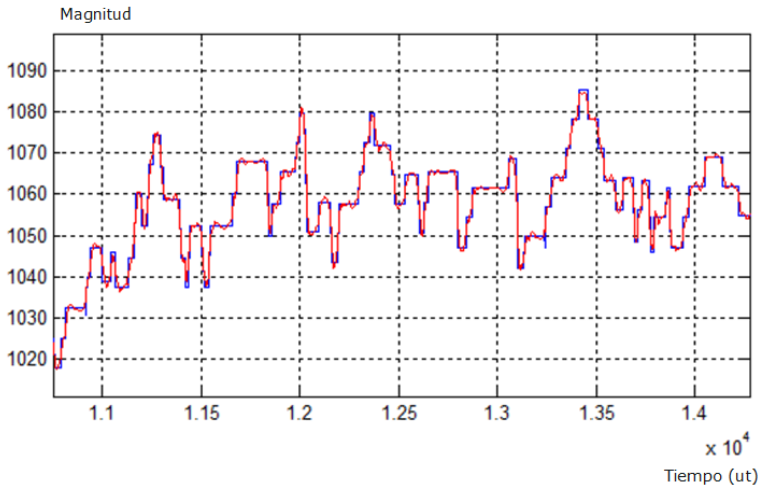

Figura 2.22 Fragmento de la señal de flujo de petróleo (en azul señal original y en rojo la filtrada)

Esto ejemplifica como el ISLab con sus diferentes métodos y algoritmos (abiertos a cambios y modificaciones) puede ser usado de forma efectiva para monitorear el comportamiento de las variables en las unidades de generación eléctrica (extensible a otras) constituyendo la base para el análisis en retrospectiva de las mediciones y para el diseño de métodos y procedimientos para alcanzar metas orientadas al desarrollo de sistemas de cómputo que actúen como asistente automatizado de la operación, entre otras posibilidades.

Conclusiones Parciales.

Se ha logrado establecer una aplicación de software en Matlab™ llamada ISLab que bajo una interfaz flexible le permite a especialistas y a investigadores, dado un registro de series temporales correspondientes a registros históricos de procesos industriales, detectar y tratar los outliers y datos perdidos que contiene, así como reducir el ruido de tal modo que se obtiene a la salida una señal lista para pasar a otra etapa de procesamiento o de extracción de información. Además, se probó la efectividad de los algoritmos implementados en el ISLab para dar solución a estos problemas demostrando que las técnicas basadas en funciones wavelets para la reducción de ruido en señales procedentes de mediciones industriales son las que mejores resultados presentan, fundamentalmente las funciones madre Daubechies, seguidas de los filtros no lineales.

Conclusiones Generales

Las conclusiones del presente trabajo son las siguientes:

- Se exponen los procedimientos y algoritmos empleados para realizar el preprocesamiento a las series temporales.
- Se implementó un algoritmo efectivo para la detección y tratamiento de outliers.
- Fue desarrollado un método de cómputo capaz de detectar y tratar los datos ausentes presentes en las series de tiempo.
- Fueron implementadas diversas técnicas para la reducción del ruido presente en las mediciones industriales dentro de las cuales se destacó la técnica Wavelet.
- Se diseñó y programó el ISLab, una herramienta que une todas las funciones implementadas para el procesamiento de las señales industriales en una interfaz gráfica de fácil acceso para el usuario.

Recomendaciones

Al culminar la tesis se ha visto que los resultados obtenidos tras la evaluación de las diversas estrategias propuestas son bastante satisfactorios. Para ampliar el margen de generalización de cada una sugerimos:

- Hacer nuevas evaluaciones sobre escenarios académicos e industriales.

- Implementar nuevos métodos para la detección de outliers y valores perdidos así como nuevos algoritmos para realizar la imputación de datos y la regresión lineal (o no lineal).

- Implementar técnicas novedosas de reducción de ruido que gozan de popularidad entre la comunidad científica como los filtros de partículas, de Wiener, etc. Para evaluar su efectividad o posible mejora ante la técnica Wavelet que dio buenos resultados en este trabajo.

- Llevar a cabo la comunicación en línea del proceso con el (ISLab), lo cual permitiría la obtención en tiempo real de mediciones preprocesadas.

- Dotar de nuevas prestaciones al software no solo en el ámbito del preprocesamiento sino también en la parte más importante de la Minería de Datos: la extracción de información de estas señales.

- Continuar perfeccionando gráficamente la interfaz creada para facilitar la interacción con el usuario.

Bibliografía

Addison, P. S. (2002). The Illustrated Wavelet Transform Handbook: Applications in Science,Engineering, Medicine and Finance. Institute of Physics Publishing, Bristol, UK.

Apte, C., B. Liu, E. P. D. Pednault y P. Smyth (2002). Business Applications of Data Mining. Communications of the ACM, 45(8), pp.49-53.

Bakhtazad, A., A. Palazoglu y J. A. Romagnoli (1999). Process Data De-noising Using Wavelet Transform. Intelligent Data Analysis, (267), pp.285.

Bakshi, B. R. (1999). Multiscale analysis and modeling using wavelets. Journal of Chemometrics, 13(3-4), pp.415-434.

Bagajewicz, M. J. (2003). Data Reconciliation and Instrumentation Upgrade. Overview and Challenges, in Fourth International Conference on Foundations of Computer-Aided Process Operations FOCAPO'03 Coral Springs, Florida.

Cios, K. J., W. Pedrycz y R. W. Swiniarski (1998).Data mining methods for knowledge discovery. Kluwer Academic., Boston, MA.

Chiang, L. H., R. J. Pell y M. B. Seasholtz (2003). Exploring process data with the use of robust outlier detection algorithms. Journal of Process Control, 13(5).

Díaz, D. (2011). Sistema para el procesamiento de las series temporales de procesos industriales. Tesis. Universidad de Oriente. Facultad de Eléctrica.

Donoho, D. L (1992). Wavelet Shrinkage and W.V.D. - A Ten-Minute Tour. In International Conference on Wavelets and Applications Toulouse, France.

Donoho, D., Johnstone, I., Johnstone, I. M. (1993). Ideal spatial adaptation by Wavelet shrinkage. Biometrika 81, 425–455.

Donoho, D. L. Y I. M. Johnstone (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. J. American Statistical Association, 90(432), pp.1200-1224.

Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth y R. Uthurusamy (1996). Advances in Knowledge Discovery and Data Mining. AAAI Press/MIT, Menlo Park, California.

Fellegi, I. Y Holt, D. (1976). A systematic approach to automatic edit and imputation. Journal of the American Statistical Association. Vol 71, 353. 17-35.

Flehmig, F.; Watzdorf, R. V. Y Marquardt, W. (1998). Identification of trends in process measurements using the wavelet transform. Computers & Chemical Engineering 22 pp.: 491-496.

Geng, Z. Y Li, K. (2003). Factorization of posteriors and partial imputation algorithm for graphical models with missing data. Statistics and probability letters. 64, 369-379.

Ghael, S., A. M. Sayeed y R. G. (1997). Baraniuk. Improved wavelet de-noising via filtering in additive noise. IEEE Transactions of Automated Control, AC-13, pp.646-655.

Gleason, T. Y Staelin, R. (1975). A proposal for handling missing data. Psychometrika. Vol 40, 2. 229-252.

Goicoechea, P. (2002). Imputación basada en árboles de clasificación.

Goebel, M. Y L. Gruenwald (1999). A Survey of Data Mining and Knowledge Discovery Software Tools. SIGKDD Explorations, 2(1), pp.20-33.

Han, J. Y M. Kamber (2001).Data mining: concepts and techniques. Morgan Kaufmann, San Francisco, CA, USA.

Hand, D., H. Mannila y P. Smyth (2001). Principles of Data Mining. The MIT Press, Cambridge, Massachusetts.

Kosanovich, K. A. Y M. J. Piovoso (1997). PCA of Wavelet Transformed Process Data for Monitoring. Intelligent Data Analysis, 1(1-4), pp.85-99.

Little, R. Y Rubin, D. (1987). Statistical Analysis with Missing Data. Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc. New York.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transaction on Pattern Analysis and Machine Intelligence, 11, pp.674-693.

Mesa, D. (2004). Imputación y Árboles de Decisión. Caracas, Venezuela. Guía práctica. Postgrado en Estadística, Universidad Central de Venezuela.

Misiti, M., Y. Misiti, G. Oppenheim y J. M. Poggi (2004). Wavelet Toolbox. User's Guide. 3 edn. Natick, MA (USA).

Narasimhan, S. Y C. Jordache (2000). Data Reconciliation and Gross Error Detection, an intelligent use of process data. Gulf Publishing Company, Houston, TX (USA).

Nounou, M. N. Y B. R. Bakshi (1999). On-line multiscale filtering of random and gross errors without process models. Aiche Journal, 45(5), pp.1041-1058.

Optimization-Toolbox (2003). Optimization Toolbox for Matlab. User's Guide. 3 edn. Natick, MA (USA).

Oppenheim, V. Y Schafer, R. (1975). Digital Signal Processing, Prentice Hall, N.J.

Platek, R. (1986). Metodología y Tratamiento de la no respuesta. Seminario Internacional de Estadística en EUSKADI. Cuaderno 10.

Popivanov, I. Y Miller, R. J. (2002). Similarity search over time-series data using wavelets. ICDE '02 Proceedings of the 18th International Conference on Data Engineering. 212-221.

Rowe, A. C. H. Y P. Abbott (1995). Daubechies Wavelets and Mathematica. Computer in Physics, 9(6), pp.635-648.

Roy, M., V. R. Kumar, B. D. Kulkarni, J. Sanderson, M. Rhodes y M. Vander Stappen (1999). Denoising algorithm using wavelet transform. Aiche Journal, 45(11), pp.2461-2466.

Struzik, Z. R. Y Siebes, A. (1999). The Haar Wavelet Transform in the Time Series Similarity Paradigm. Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery, Springer-Verlag pp.: 12- 22.

Taswell, C. (2000). The what, how, and why of wavelet shrinkage denoising. IEEE Computing in Science & Engineering, 2(3), pp.12-19.

Anexos

A. Familias de transformada wavelet.

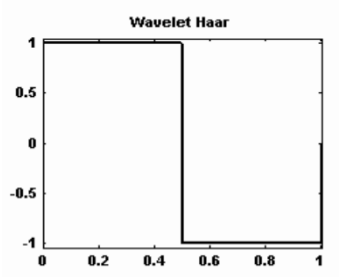

Figura A.1 Wavelet Haar.

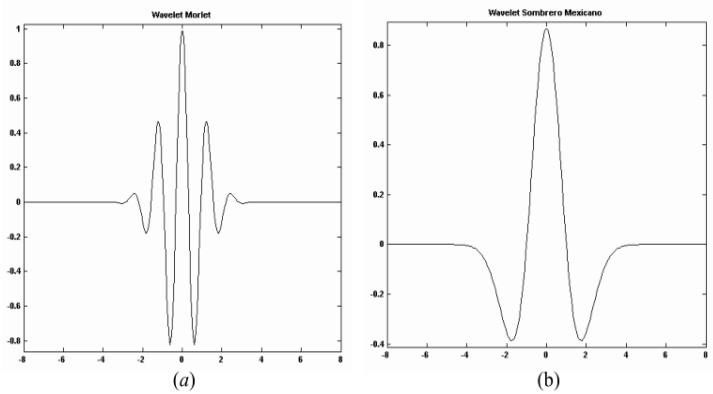

Figura A.2 (a) Wavelet Morlet y (b) wavelet Sombrero mexicano

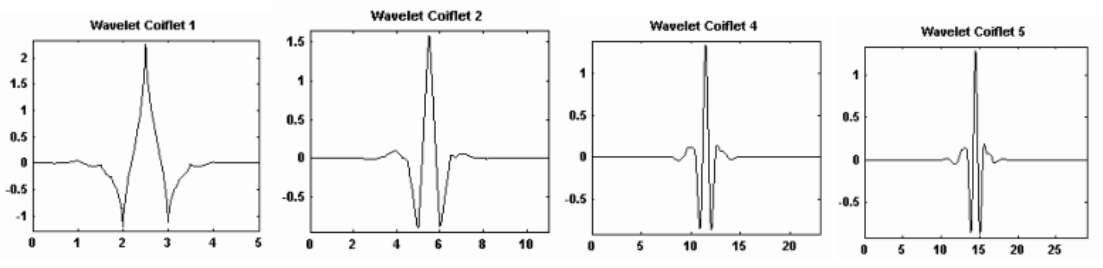

Figura A.3 Familia de wavelet Coilflet.

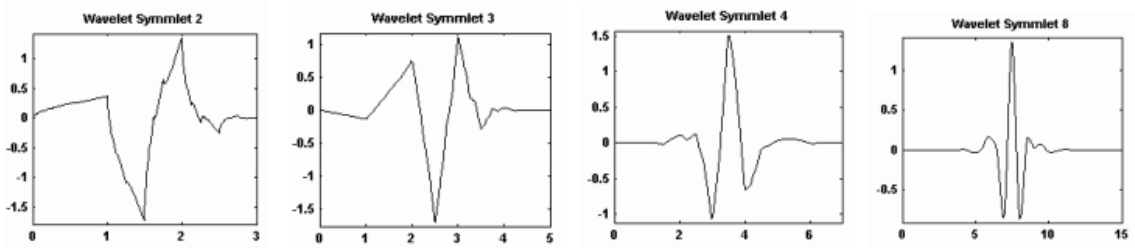

Figura A.4 Familia Symmlet.

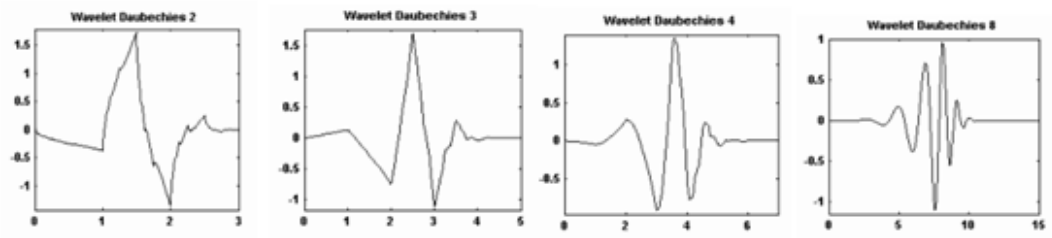

Figura A.5 Familia de wavelets Daubechies.

B. Descomposición de señales usando transformada wavelet.

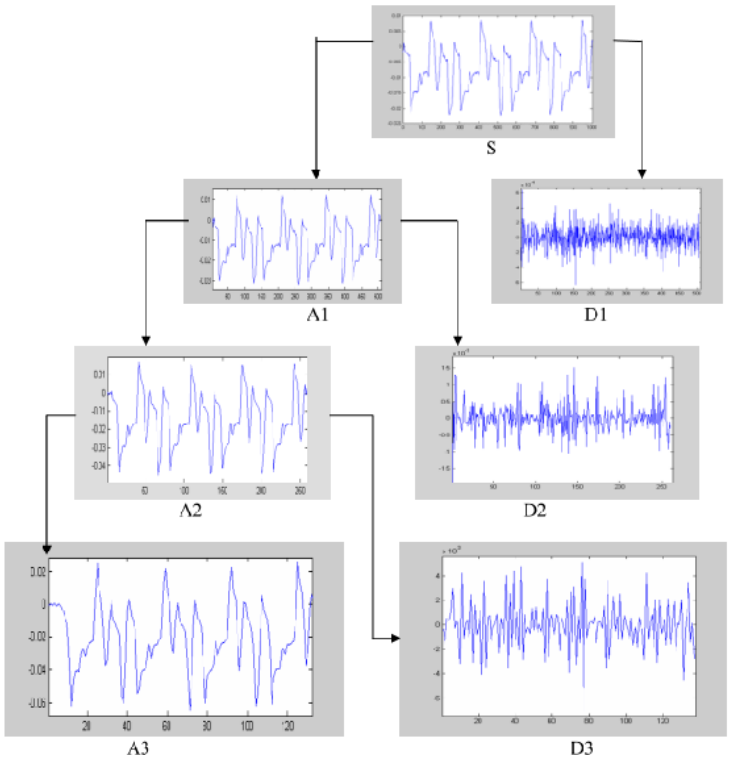

Figura B.1 Descomposición en tres niveles de la misma señal.

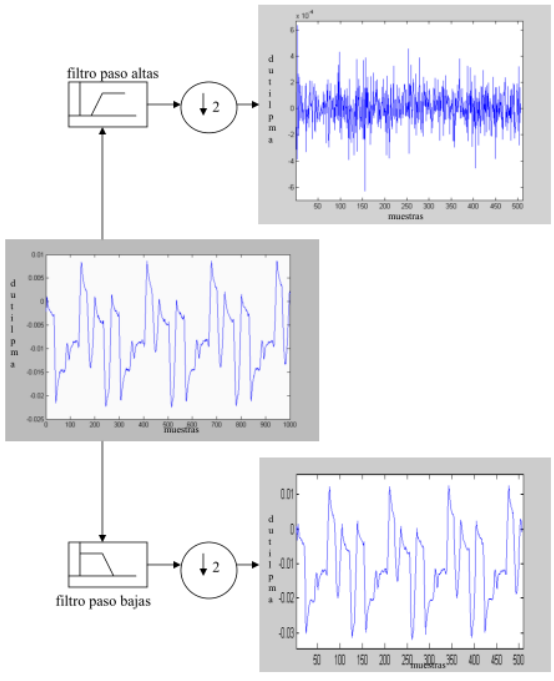

Figura B.2 Descomposición de señal en alta y baja frecuencia con reducción de muestras dadas por los coeficientes wavelets.

C. Diagramas de flujo.

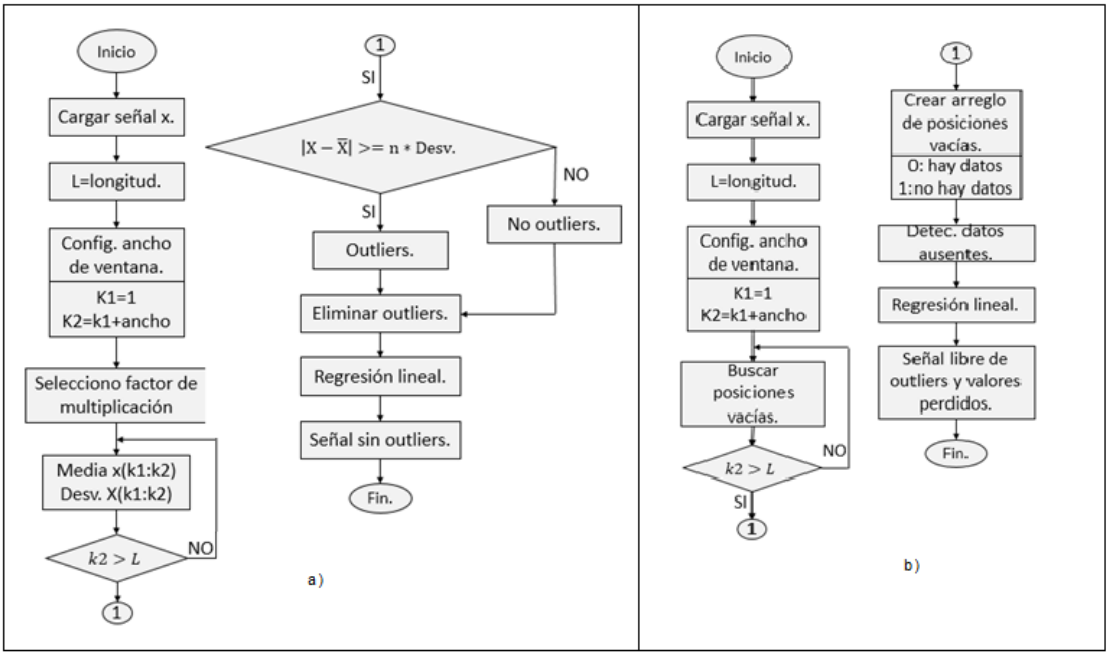

Figura C.1 Métodos de detección y eliminación de: a) outliers, b) datos ausentes.

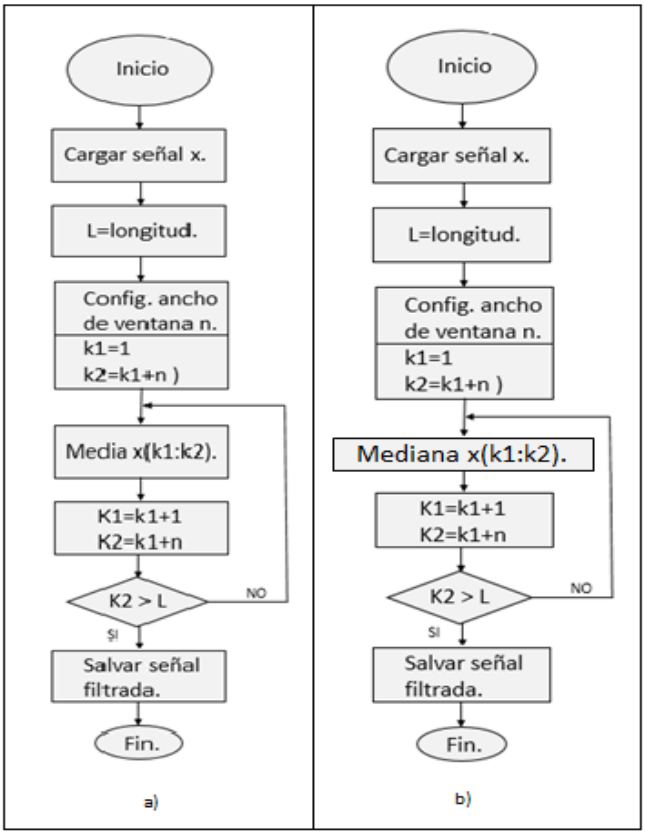

Figura C.2 Filtros de: a) media móvil, b) mediana.