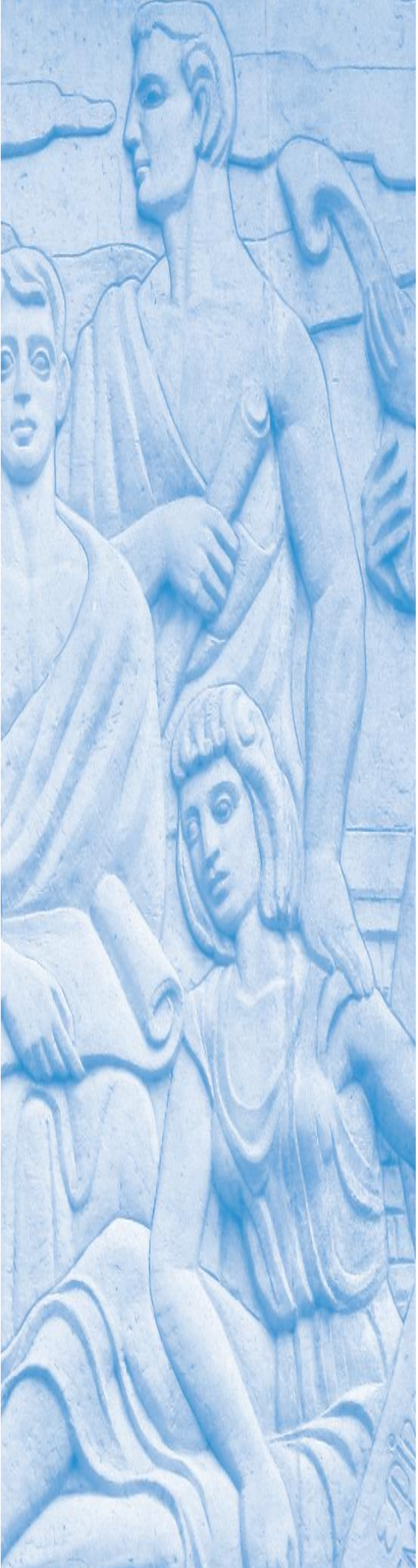


UNIVERSIDAD DE ORIENTE

Trabajo de Diploma

Autor: Pedro Pablo Luperón Bauzá.

Tutores: MSc. Ing David Díaz Martínez.

DSc. Ing Luis V. Seisdedos.

Santiago de Cuba

2017

UNIVERSIDAD DE ORIENTE | **Facultad de Ingeniería Eléctrica**
Departamento de Control Automático

Trabajo de Diploma

Título: Método para identificar estados estacionarios sobre registros de procesos de conversión de energía con correlación serie y mutua.

Autor: Pedro Pablo Luperón Bauzá.

Tutores: MSc. Ing David Díaz Martínez.
DSc. Ing Luis V. Seisdedos.

Santiago de Cuba
2017

Pensamiento:

Si quieres ser sabio, aprende a interrogar razonablemente, a escuchar con atención, a responder serenamente y a callar cuando no tengas nada que decir.

Johann Kaspar Lavater

Agradecimientos

Agradecimientos:

A lo largo de mis años de estudiante universitario han existido muchas personas que han hecho posible que pueda llegar hasta donde hoy me encuentro. Es por esto que sería imperdonable no expresar mi gratitud:

1. *A Dios, que me mostró el camino y me ha guiado hasta aquí. No me alcanzará la vida para exaltarlo y darle gracias.*
2. *A mis padres por su gran amor y por su apoyo en los momentos de dificultad. Porque gracias a sus consejos y a la forma de pensar que me inculcaron he alcanzado muchas de mis metas; y las que me falten serán por ellos y para ellos que las lograré.*
3. *A mis hermanos Ely y Juan por estar acompañándome siempre, física y mentalmente.*
4. *A mi novia Lisbette por amarme aún en los momentos más difíciles y a su excepcional madre.*
5. *A mi abuela Raquel Leyva y a mi tía Melissa Teruel Leyva, porque ambas han sido un pilar de la educación y del amor en mi vida.*
6. *A todo familiar que de una forma u otra me ha ayudado a cumplir este sueño.*
7. *A mi tutor, el MSc. David Díaz Martínez, por su peculiar y único punto de vista, su ayuda desinteresada y amistad.*
8. *A mi amiga, vecina y hermana Lida Beatriz Vázquez Blanco y a sus padres, por compartir conmigo cada curiosidad científica y por enseñarme a exigirme cada vez más.*
9. *A mis amigas Anita, Analia, Irelvis y a todos mis compañeros de cuarto por estar conmigo durante estos años de universidad.*
10. *Al excelente claustro de profesores de la carrera de Automática por haberme llenado de los conocimientos necesarios para la realización de este trabajo.*

Dedicatoria

Al igual que todas mis notas académicas y mis logros durante mi periodo de estudiante, quisiera dedicar este trabajo a mi madre; porque sin su amor, sus consejos, su sabiduría y su paciencia me hubiese resultado imposible la realización del mismo. Por cada momento vivido, por cada dificultad superada y su inmenso amor incondicional; mi más profundo, sincero y especial agradecimiento sólo para ti, madre.

RESUMEN

A medida que se ha ido avanzando en el campo del almacenamiento de la información, se ha hecho cada vez más evidente la importancia de explotar los datos de los procesos industriales implícitos dentro de las grandes minas de datos y en cómo éstos nos ayudan a mejorar el funcionamiento de las plantas de procesos actuales.

El siguiente trabajo constituye una herramienta que permite aprovechar, explotar y obtener los importantes conocimientos ocultos detrás del inmenso volumen de información de una mina de datos. Entre sus puntos principales se hará énfasis en la detección de Estados Estacionarios (EE) en procesos con correlación serie y mutua. La realización del mismo se centrará en la obtención y manipulación de una serie de datos, para así arribar a una serie de conclusiones que evidencien la importancia de este trabajo.

El SSITS (por sus siglas en inglés, *Steady State Identifier in Time Series*) permite hacer uso de un gran número de herramientas, todas dentro del software Matlab™, que se encuentran integradas en una única interfaz de usuario. Este software trabaja con los datos de una serie temporal y tiene como objetivo determinar el mejor método de detección de Estados Estacionarios; para que de esta manera se pueda realizar un diagnóstico preventivo de las dificultades que pueda presentar el proceso industrial mediante un estudio de la degradación de las variables de dicho proceso. También podrá visualizar cada una de estas variables y datos importantes de la misma. Se guardarán, además, los resultados, variables y configuraciones para su posterior estudio.

ABSTRACT

While data storage has continuously progressed, it has become increasingly evident the importance of exploiting the data of the industrial processes implicit within the large data mines and how they are help us to improve the current operations in process plants.

The following work is a tool that allows us to take advantage, exploit and get the important knowledge behind the immense volume of information in a data mine. Emphasis will be placed on the detection of Stationary States (SS) in processes with serial and mutual correlation. The development will be focus on the obtaining and manipulation of a series of data, in order to reach to conclusions that will raise awareness of the importance of this work.

The SSITS (Steady State Identifier in Time Series) allows us to use a large number of tools, contained within the Matlab ™ software, all of which are integrated into a single user interface. This software works with the data of a time series and aims to determine the best method of detection of Stationary States; so that a preventive diagnosis can be made of the main faults that the industrial process can present through a study of the degradation of the variables of said process. You can also view each of these important variables and data and the main results, variables and settings for further study will be also saved.

NOTACIONES GENERALES

ACF: Autocorrelation Function.

CRM: Customer Relationship Management.

CURE: Clustering Using Representatives.

DAS: Data Acquisition Systems.

DCS: Distributed Control System.

DDC: Direct Digital Control.

DR: Data Reconciliation.

EE: Estado Estacionario.

ERP: Enterprise Resource Planning.

ET: Estado Transitorio.

EWMA: Exponentially Weighted Moving Average.

GED: Gross Error Detection.

GUIDE: Graphical User Interface Development Environment.

KDD: Knowledge Discovery in Data.

MEWMA: Multivariate Exponentially Weighted Moving Average.

OLE: Object Linking and Embedding.

OPC: OLE for Process Control.

PCA: Principal Component Analysis.

PIT: Polynomial Interpolation Test.

PLC: Programmable Logic Controller.

RAT: Reverse Arrangements Test.

SCADA: Supervisory Control and Data Acquisition.

SS: Stationary States.

SSI: Steady State Identification.

CONTENIDO

INTRODUCCIÓN.

En muchos países las termoeléctricas son la fuente de electricidad más común, éstas suministran energía para satisfacer las necesidades de fábricas, industrias, sector residencial y consumo en general.

Para llevar a cabo la Revolución energética en Cuba, fue necesario romper con los esquemas tradicionales en la generación de energía eléctrica. Debido a la necesidad de revitalizar el sistema eléctrico nacional se optimizaron en el país todas las termoeléctricas que sincronizadas al Sistema Eléctrico Nacional apoyaron grandemente este cambio revolucionario.

El análisis de los factores que afectan el buen funcionamiento en las termoeléctricas es un aspecto importante a tener en cuenta. La determinación y descripción de los factores responsables para la falla de un componente, mecanismo o estructura, brindan una valiosa información para mejorar tanto el diseño, los procedimientos operativos y el uso de los componentes; como para evitar paradas de línea o pérdidas de producción en la industria.

Dentro de un proceso de generación de energía eléctrica existen diferentes variables del proceso (presión, temperatura, flujo, etc.), las cuales, estando en los valores correctos, determinan la calidad del proceso industrial. Cuando estas variables se encuentran en un valor correcto de operación y la planta no presenta problemas en su generación de electricidad, el sistema en cuestión alcanza un estado de operación estacionario y viceversa; cuando el sistema está en estado estacionario (EE) y la energía generada tiene la calidad requerida se puede afirmar que las variables del proceso se encuentran en su valor óptimo. Debido a lo antes explicado es posible prever, a partir del análisis del estado estacionario del sistema, hacer un estudio de la degradación de las variables del proceso; con el fin de detectar a tiempo las posibles fallas que pudieran ocurrir y planificar a tiempo las diferentes reparaciones y mantenimientos.

La competencia en continuo aumento, aparejada con rígidas normas medioambientales y de seguridad, ha conducido a los procesos industriales a una

automatización a gran escala. Esto, claro está, requiere monitoreo preciso de la operación de la planta. Las computadoras modernas con su poderoso hardware permiten que se realicen cientos de mediciones simultáneas cada segundo, dichas mediciones crean la base para futuras decisiones que favorecerán el proceso. Estas medidas, por otra parte, están a menudo corruptas con ruido aleatorio y algunas veces con errores sistemáticos. En presencia de estos errores, los pequeños cambios que están ocurriendo en un proceso pueden ser enmascarados, dificultando la mejoraría de dicho proceso o el alcance del objetivo propuesto. Por consiguiente, el analizar datos es un aspecto importante dentro de la automatización de procesos y también un problema potencial que necesita ser atendido antes de que sea intentada la optimización. Por la naturaleza aleatoria de las mediciones, los métodos estadísticos proveen herramientas eficientes y son utilizados eficazmente para solucionar este problema.

Existen tres aspectos importantes dentro del procesamiento de datos: el primero es la identificación de los estados estacionarios (SSI, del inglés *Steady State Identification*), lo cual involucra averiguar si la planta está trabajando bajo condiciones estacionarias; el segundo es la detección de errores gruesos (GED, del inglés *Gross Error Detection*), que ayuda a eliminar errores sistemáticos y el tercero es la conciliación de los datos (DR, del inglés *Data Reconciliation*), que involucra entre otros factores hacer estos datos libres de errores aleatorios.

Los modelos de estados estacionarios son a menudo utilizados para el monitoreo de procesos *online,* predicción de las propiedades del producto, la optimización *online*, etc. Los datos en línea se usan para ajustar estos modelos y para poder representar un sistema dinámico usando un modelo estático. La metodología empleada para DR también depende de la existencia de un estado estable. La detección adecuada de los intervalos de tiempo en los cuales un proceso se encuentra en estado estacionario es sumamente importante para el monitoreo y optimización del mismo. Si el proceso no está realmente en estado estacionario, o sea, se encuentra en estado transitorio (ET), la aplicación de modelos de estado estacionario al mismo puede provocar estimaciones incorrectas de parámetros, la toma de decisiones inapropiadas y operación inestable cuando el sistema está

operando a lazo cerrado en tiempo real. Detectar ventanas o intervalos de cuando un proceso continuo está operando en un estado de estabilidad es útil especialmente cuando se están utilizando modelos de estados estacionarios para optimizar el proceso o planta en tiempo real; por tanto, la identificación de los estados estacionarios es el paso de puesta en marcha de todas estas aplicaciones. Seguir con estas técnicas sin averiguar la existencia de un estado estable podría conducir a un total desorden en el proceso bajo consideración (Jeffrey D. & John D., 2012).

El término “estado estacionario” implica que un proceso está operando dentro de un punto estable o una región estacionaria, se debe asumir que la acumulación o la tasa de cambio de un material determinado, energía y el *momentum* son estadísticamente insignificantes o despreciables.

Una aplicación de la detección de estados estacionarios es determinar el criterio de parada para los métodos numéricos iterativos. Muchas técnicas matemáticas como la regresión no lineal y la optimización son iterativas y, por tanto, requieren un criterio de parada. En vez de iterar un número fijo de veces, el proceso debería detenerse cuando la función objetiva alcanza valores cercanos a los valores en estado estacionario (respetando siempre el número de iteraciones). El desarrollo de una técnica factible para la identificación de estados estacionarios ayudará a obtener todos los beneficios de los métodos basadas en modelos de control de proceso *online*. (Cao & Rhinehart, 1995).

Aplicar el modelo EE a un proceso que no lo está, resultaría en errores de Tipo I y de Tipo II (falsos positivos y falsos negativos), estimaciones de parámetros incorrectas y por último una incorrecta toma de decisiones acerca de cómo conducir el sistema para que sea más económico, eficiente y efectivo. Si se supone incorrectamente que un proceso se encuentra en EE se puede incurrir en una violación seria, ya que puede resultar en una operación inestable cuando se aplique una optimización a tiempo real o a lazo cerrado. Conocer exactamente cuando algunos procesos son estacionarios y cuando no lo son puede ayudar a identificar y a diagnosticar eventos anormales potenciales o síntomas de estos en otras áreas

de la planta como alarmas de poco flujo de vapor o emanación de líquido contaminado.

Hay diversas técnicas disponibles para detección de estados estacionarios (DEE). La mayor parte de las mismas se basan en seleccionar una ventana de datos, calculando ya sea el promedio, la varianza, o la pendiente de regresión de esta ventana de datos y comparar estos resultados con los anteriores usando las pruebas estadísticas apropiadas.

Los métodos de DEE vistos en la literatura presentan serias dificultades a la hora de considerar los efectos de la correlación serie y mutua en los datos por lo que en general terminan obviando este análisis. Además, tienen un carácter univariado y se ha hecho muy poco para traducir las metodologías teóricas en aplicaciones computacionales que puedan ser usadas en la práctica.

Debido a lo antes expuesto, nos planteamos como:

Problema de la investigación:
La ineficiencia de los métodos existentes para la DEE ante mediciones serie y mutuamente correlacionadas procedentes de unidades de generación de energía eléctrica.

Objeto de la investigación: Los registros históricos archivados a partir de mediciones instrumentales de variables correspondientes a sistemas de generación de electricidad.

Objetivo: Diseñar e implementar una aplicación computacional que permita identificar eficazmente los estados estacionarios en mediciones provenientes de plantas energéticas compensando los efectos de la correlación serie y mutua.

Campo de acción: Las estrategias de detección de estados estacionarios aplicados a las unidades de generación de electricidad utilizando el software Matlab™.

Hipótesis: Si se diseña una herramienta de software capaz de identificar satisfactoriamente los intervalos de EE en series temporales contaminadas con ruido en plantas de energía; entonces, se contará con datos fiables para pasar a un procedimiento de monitorización del desempeño de dichas plantas.

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

1. Caracterizar desde el punto de vista gnoseológico, histórico y actual los aspectos teóricos referidos al reconocimiento de estados estacionarios en procesos industriales.
2. Diagnosticar y fundamentar el problema de la insuficiencia de herramientas informáticas capaces de detectar estados estacionarios con enfoque multivariado.
3. Implementar algoritmos y métodos eficientes para la detección ante señales con correlación serial y mutua.
4. Construir una interfaz de usuario que sea capaz de integrar todas las prestaciones y utilidades necesarias para la resolución del problema antes planteado.
5. Comprobar la validez de los métodos empleados mediante señales patrones.

Métodos y técnicas empleados en la investigación:

1. Análisis de fuentes documentales.

2. Técnicas y métodos empíricos: Observación y experimentación.

3. Método histórico lógico.

4. Método de análisis y síntesis.

5. Métodos experimentales: Programación.

Significación práctica:

La investigación realizada, en conjunto con la herramienta informática desarrollada, ofrece al sector industrial una herramienta computacional capaz de explotar todas las ventajas y beneficios de la detección de los estados estacionarios de un proceso de generación de energía eléctrica.

Este trabajo también sirve como un puente entre la Universidad de Oriente y aquellas empresas interesadas en el control de las variables que están vinculadas directamente a la producción y en la implementación de medidas correctivas en sus procesos productivos. Para el Departamento de Control Automático constituye un avance dentro del área del Control Estadístico de Procesos. Brinda además una base sólida para el trabajo en una nueva línea de investigación para ingenieros químicos, mecánicos, informáticos y automáticos.

El informe de la investigación se estructuró de la siguiente manera: Introducción, dos capítulos, conclusiones generales, recomendaciones, bibliografía y anexos.

En el capítulo I se realiza un estudio teórico del presente trabajo. Para ello se exponen las principales definiciones y principios básicos de la minería de datos y la detección de estados estacionarios.

En el capítulo II se exponen las características específicas del SSITS, su interfaz de usuario, principales componentes, facilidades que ofrece y los algoritmos computacionales empleados.

CAPÍTULO 1: Trabajo con Series Temporales y Detección de Estados Estacionarios.

Introducción.

Las industrias en la actualidad 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. Al procesar estos registros correctamente, se puede mejorar la operación de la planta mediante la identificación de aquellas variables que están influyendo sobre la eficiencia y el comportamiento de la misma y se puede realizar un diagnóstico preventivo de fallos y malos funcionamientos.

En el presente capítulo se realiza primeramente un estudio de los aspectos referidos a los sistemas encargados de registrar y almacenar estos importantes datos, su evolución, características, etc. También se aborda el tema de la minería de datos que se encarga de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Se analizan puntos referidos al procesamiento de las series temporales, que es un tema fuertemente relacionado con todo el trabajo que se plantea, y por último se exponen aspectos sobre la importancia que tiene la detección de los estados estacionarios del proceso ya que esa es la filosofía principal que se sigue en este trabajo.

1.1 Evolución de los sistemas de control, supervisión y almacenamiento de datos históricos del proceso.

El primer uso de las computadoras digitales fue en la adquisición de datos desde las plantas y su almacenamiento sin ninguna influencia sobre la operación del proceso. Estos sistemas, que fueron llamados inicialmente Sistemas de Adquisición de Datos (DAS, del inglés *Data Acquisition Systems*), evolucionaron y la información recolectada por ellos, posteriormente fue procesada para obtener los puntos de

ajuste más óptimos para cada lazo de control de los procesos. Estos valores de referencia calculados inicialmente eran enviados a los controladores analógicos y luego a los controladores digitales. Por lo que este método de operación se denominó control supervisor.

Con el paso de pocos años compañías de alto prestigio introdujeron sus versiones particulares de computadoras industriales para el control y supervisión de procesos entre las que se encontraban IBM, General Electric, Foxboro, Control Data Corporation y muchas otras más. Los segundos pasos se realizaron con el diseño de computadoras en las cuales se introdujeron memorias de alta velocidad, aunque el salto más importante ocurrió cuando se sustituyen los dispositivos de estado sólido (diodos, transistores, capacitores, resistores) por los circuitos integrados aumentando la velocidad de procesamiento de las computadoras industriales. El siguiente capítulo en la evolución de los sistemas de supervisión y control por computadoras aparece con el desarrollo del Control Digital Directo (DDC, del inglés *Direct Digital Control*), donde las funciones de los instrumentos analógicos (filtros, controladores analógicos PID) fueron incorporadas a las computadoras y se eliminan físicamente del proceso; motivo por el cual, la tendencia fue a reducir los costos de los sistemas de supervisión y control y a incrementar la flexibilidad en el diseño de nuevos sistemas. Sin embargo, los mayores avances llegaron en la década de 1970, cuando Honeywell anunció el primer Sistema de Control Distribuido (DCS, del inglés *Distributed Control System*) llamado TDC-2000. La base de este sistema estaba en la redundancia de los controladores digitales, las comunicaciones y las interfaces para los operadores del proceso. En ese mismo año la firma japonesa Yokogawa saca al mercado el sistema CENTUM. En el año 1980, Bailey (ahora parte de ABB) introduce el sistema NETWORK 90 y la Fischer & Porter Company (ahora también parte de ABB) introduce el DCI-4000 (DCI para la instrumentación distribuida del control).

El desarrollo de los DCS dio lugar a la adopción del sistema operativo predominante hoy en día: UNIX. Los años 80 también atestiguaron el primer PLC (del inglés, *Programmable Logic Controller*) integrado en la infraestructura DCS.

La invasión de Microsoft dio lugar al desarrollo de tecnologías tales como OPC (del inglés, *OLE for Process Control,* donde OLE significa Objeto de Enlace Empotrado), que ahora es un estándar de la conectividad de la industria. Los años 90 también se caracterizaron por “la Guerra Fieldbus”, donde las organizaciones rivales compitieron para definir lo que se convirtió en el estándar del Fieldbus IEC para la comunicación digital con la instrumentación de campo en vez de comunicaciones analógicas 4-20 mA.

Todos estos sistemas y los que se continúan desarrollando en la actualidad son los que han hecho posible la adquisición de los datos transmitidos por los sensores de campo, en principio para propósitos puramente de control, pero a medida que pasó el tiempo también se fueron almacenando para su uso posterior en muchas aplicaciones y estudios de la planta.

1.2 Sistemas y métodos para la historización de los datos.

La razón principal que debe impulsar a la historización de los datos, es que a partir de estos registros históricos obtenidos se puede comprender con mayor profundidad los datos con el objetivo de mejorar tanto en la eficiencia como en el factor económico del proceso.

Los datos históricos también ofrecen beneficios para la diagnosis y el mantenimiento de equipos tales como bombas y válvulas. Esta información puede permitir seguir la degradación de una pieza a lo largo del tiempo, de forma que pueda ocurrir el mantenimiento preventivo cuando sea necesario. Pueden prevenir fallos inesperados debido a piezas rotas, desgaste prematuro u otros problemas mecánicos inesperados.

En las últimas décadas han emergido productos especializados los cuales son muy eficaces para la historización de los datos. Estos productos soportan colecciones de datos para cientos de fuentes de datos como los PLC, sistemas de control distribuidos (DCS, del inglés *Distributed Control System*), Sistemas de Supervisión y Adquisición de Datos (SCADA, del inglés *Supervisory Control and Data Acquisition*) y Servidores OPC, proporcionan almacenaje eficiente y eficaz de datos, ofrecen funcionalidad de valor añadido tal como agregación, y suministran datos a

los clientes mediante una rica diversidad de herramientas que incluyen tendencias, *displays* e informes.

Actualmente puede plantearse una solución de historización a través de sistemas existentes DCS o SCADA. Cuando la solución se expande, la capacidad para soportar la captación de datos del proceso de múltiples fuentes es cada vez más importante. En la compañía se desplegarán diferentes sistemas DCS, PLC y SCADA, por lo que seleccionar la solución que los soporte es crucial.

La recolección de los datos del proceso se realiza a través de un programa de interface de datos o de colección de datos para cada fuente de datos específicos. Un cada vez mayor número de vendedores está haciendo que sus procesos estén disponibles a través de un OPC, así que la interfaz de colección de datos OPC está siendo el método más popular para obtener datos del proceso. Dentro de las grandes ventajas de un servidor OPC resaltan que los fabricantes de hardware sólo tienen que hacer un conjunto de componentes de programa para que los clientes los utilicen en sus aplicaciones y no tienen que adaptar los *drivers* ante cambios de hardware. Prácticamente todos los mayores fabricantes de sistemas de control, instrumentación y de procesos han incluido OPC en sus productos.

1.3 SCADAS y DCS para el manejo de las series temporales.

Un Sistema de Control Distribuido es un sistema de control aplicado, por lo general, a un sistema de fabricación, proceso o cualquier tipo de sistema dinámico, en el que los elementos del tratamiento no son centrales en la localización (como el cerebro), sino que se distribuyen a lo largo de todo el sistema con cada componente o sub-sistema controlado por uno o más controladores. Todo el sistema está conectado mediante redes de comunicación y de monitorización. (ver **Anexo 1**).

En instalaciones más complicadas y plantas en gran escala, los lazos de control se cuentan en el orden de los centenares. Para tales procesos grandes, los DCS son más apropiados. Hay muchos vendedores que proveen estos sistemas; entre los que destacan: Baily, Foxboro, Honeywell, Rosemont, Yokogawa, etc.

Conceptualmente, los DCS son similares a una red simple de PC. Sin embargo, hay algunas diferencias. En primer lugar, el hardware y software de los DCS están

hechos de manera tal que sean más flexibles, fácil para modificar y configurar, y ser capaz para maniobrar un gran número de lazos.

En segundo lugar, los DCS modernos son equipados con optimización, modelo de alto rendimiento y el software para su control como opciones. Por consiguiente, un ingeniero imaginativo que tiene un fondo teórico en sistemas de control modernos, puede configurar rápidamente la red de un DCS con el objetivo de implementar un control de alto rendimiento.

Las mayores ventajas de un DCS son flexibilidad en el diseño de sistemas, facilidad de expansión, confiabilidad y facilidad de mantenimiento. Una ventaja obvia de este tipo de arquitectura distribuida es que la pérdida completa de la carretera de datos no causará una pérdida completa de la capacidad del sistema. A menudo las unidades locales pueden continuar su operación sin pérdidas significativas de función en períodos de tiempo moderados o extendidos. (King Saud University, 2002)

1.4 Descripción del Toolbox GUIDE del Matlab™.

Matlab™ es el nombre abreviado de "*MATrix LABoratory*". Es un programa para realizar cálculos con vectores y matrices. Como caso particular puede también trabajar con números escalares, tanto reales como complejos. Una de las capacidades más atractivas es la de realizar una amplia variedad de gráficos en dos y tres dimensiones. Matlab™ tiene también un lenguaje de programación propio. Es, para muchos, el paquete más usado en ingeniería en la actualidad. (Arafet, Domínguez, & Chang, 2004)

Este software se ha ido perfeccionando con el transcurrir del tiempo logrando un alto avance dentro de su paquete de herramientas. Su principal avance es el Toolbox GUIDE que permite crear de modo interactivo la interface de usuario y ejecutar programas que necesiten ingreso continuo de datos. Tiene las características básicas de todos los programas visuales como Visual Basic o Visual C++, aunque todavía con unas posibilidades mucho más limitadas.

Su nombre GUIDE proviene de la contracción de las siglas GUI (*Graphical User Interface*) e IDE (*Integrated Development Environment*), es decir, es un ambiente de desarrollo integrado (IDE) para la confección de interfaces gráficas de usuario (GUI).

Desde el punto de vista de su aspecto, GUIDE, se parece a una moderna herramienta para el diseño de interfaces de usuario rápidas, que no son más que programas que facilitan la producción y puesta a punto de software en relativo poco tiempo. Hay que decir que antiguamente Matlab™ no proporcionaba esta herramienta y la tarea de realizar un programa en dicho lenguaje con ventanas y botones era bastante tediosa, pero con GUIDE esto deja de ser así, por tanto, si bien no es uno de los avances más importantes de Matlab™, sí ha sido uno de los más agradecidos por los usuarios.

1.5 Introducción a la minería de datos.

El *data-mining* (*minería de datos*), es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Básicamente, el *data-mining* surge para intentar ayudar a comprender el contenido de un repositorio de datos. Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales. (Lon-Mu, Siddhartha, Sclove, & Rong, 2001)

De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación que surge entre la información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento.

Sus dos retos fundamentales son: trabajar con grandes volúmenes de datos con los problemas que ello conlleva y el empleo de técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil.

La minería de datos es en sí misma, parte de un dominio aún más grande que se conoce como descubrimiento del conocimiento a partir de datos (KDD; del inglés *Knowledge Discovery in Data*), 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. Diversos autores (Wang, 2001), (Cios, Pedrycz, & Swiniarski, 1998), (Hand, Mannila, & Smyth, 2001) han adoptado como básica la definición de KDD propuesta por (Fayyad, Piatesky-Shapiro, & Smyth, 1996): “KDD es el proceso no trivial de identificar a través de los datos patrones noveles, potencialmente útiles y entendibles”.

Según (Hernández, Ramírez, & Ferri, 2004), para una mejor comprensión del concepto de Minería de Datos se divide a continuación este proceso en 6 etapas fundamentales, como sigue:

- Recolección de datos.
- Depuración y transformación de los datos.
- Generar el modelo de minería de datos.
- Evaluación del modelo.
- *Reporting.*
- Predicción.

Para obtener los mejores resultados de cualquier proyecto de minería de datos es muy importante investigar si realmente los datos proporcionan una representación exacta del problema, debe tenerse en cuenta que una gran cantidad de datos no siempre equivale a una gran cantidad de información. Estas dificultades podrán ser resueltas usando métodos de selección de propiedades tales como; agrupamiento (*clustering*) y análisis de las componentes principales, o la aplicación del conocimiento acerca del proceso.

Sin embargo, la adquisición de datos 'buenos' para la mina es una de las partes mayormente consumidoras de tiempo del proceso. Se estima que hasta el 80 por ciento de la duración de un proyecto de minería de datos podría ser localizado en la etapa de preparación de datos. De todos modos, el tiempo gastado en esta fase se refleja en los resultados que se alcancen posteriormente.

Dentro de la minería de datos existe un término llamado Quimiometría que trata, específicamente, de todos aquellos procesos que transforman señales analíticas y datos más o menos complejos en información. La Quimiometría utiliza métodos de origen matemático, estadístico y otros procedentes del campo de la lógica formal para conseguir sus fines. Por todo ello, la Quimiometría se sitúa en un campo interdisciplinar. Aunque sus métodos y herramientas provienen de otras disciplinas, claramente los fines de la Quimiometría están ligados a la química y su éxito depende de los problemas químicos que sea capaz de resolver.

Para caracterizar, evaluar, diagnosticar, pronosticar, etc. los procesos químicos, por ejemplo, la producción de energía, se requiere disponer de herramientas de cómputo para aplicar las técnicas de Quimiometría. La meta de la mayoría de las técnicas de la Quimiometría es derivar un modelo empírico, a partir de los datos, que le permita al investigador estimar una o más propiedades del sistema desde sus mediciones instrumentales.

1.6 La calidad de los datos. Detección y corrección de errores.

La calidad de datos es una medida importante que las empresas pueden usar para determinar el "estado de forma" de la información empresarial para su uso en la planificación de estrategias y toma de decisiones tácticas, y en las actividades operativas del día a día. La carencia de calidad de los datos continúa siendo uno de los principales problemas para muchas organizaciones de hoy, a medida que los entornos de información se hacen más y más sofisticados. La existencia de aplicaciones, bases de datos, sistemas, mensajes y documentos dispares hacen más difícil que nunca, la identificación y control de la calidad de datos de una forma constante.

Actualmente los datos de las mediciones realizadas en determinado proceso, por lo general, salen afectados por diferentes factores como errores aleatorios y errores gruesos; que dentro de estos últimos resaltan los valores picos y saltos en el valor de la medida de una variable. Además, se ven seriamente afectados por el ruido, pérdidas e inconsistencias. Estos errores se deben, por lo general, a fallos en la comunicación entre el sensor y la computadora (o algún otro componente

perteneciente al enlace entre estos elementos) donde se está visualizando determinada variable del proceso. También pueden ser encontrados errores que afectan la calidad de los datos cuando el sensor que está realizando las mediciones se encuentra defectuoso o necesita mantenimiento.

Se hace necesario entonces implementar métodos capaces de detectar y corregir estos errores que afectan la calidad de datos. Por lo tanto, se puede poner foco en:

- Detectar y corregir inconsistencias: Básicamente se trata de detectar registros que no cumplan con determinadas reglas, y luego modificar los datos. Una técnica para la localización de errores es la llamada *data editing*, la cual consiste en la definición de reglas (*edits*) que deben ser respetadas por cierto conjunto de datos, para lograr de esta manera la detección de inconsistencias.
- Detectar y corregir datos incompletos: En este caso si bien es muy simple detectar los datos incompletos, puede que corregir sea difícil (en el caso de no tener forma de obtener la información faltante). Aquí se distinguen dos tipos de fuentes de incompletitud: datos truncados y datos censurados.
- Detectar y corregir anomalías: Este es el caso de datos cuyo valor difiere en gran medida con respecto a los demás datos.

La técnica de reconciliación de datos es para mejorar la precisión, consistencia y confiabilidad de los datos y tiene el objetivo de reducir el impacto de los llamados errores aleatorios. Se fundamenta en un ajuste óptimo de los datos medidos tales que estos valores ajustados deben satisfacer condiciones basadas en leyes de conservación (por ejemplo, de masa, de energía, de elementos químicos, etc.) u otras restricciones del proceso.

1.7 Técnicas de pre-procesamiento de los datos.

Antes de aplicar cualquier técnica de minería de datos es preciso realizar un análisis previo de los datos que se dispone. Al proceso anteriormente mencionado se le conoce como pre-procesamiento de datos y es muy importante puesto que garantiza la integridad y veracidad de los datos. En (Dapozo, Porcel, López, & Bogado, 2007) se describen las siguientes etapas principales:

Limpieza de datos: Tiene como objetivo reducir el ruido y las inconsistencias. Para ello, se seleccionará una muestra resumen de los datos interpretando la validez de algún valor para algún atributo y mejorar la calidad de los datos. Para el manejo de datos con ruido, uno de los métodos que existen es el *binning*, que permite reducir la numerosidad.

Integración de Datos: Se realiza para eliminar las redundancias que pueden ser detectadas por un análisis correlacional. Si se realiza la integración de datos con cuidado se puede evitar la inconsistencia y la redundancia entre los datos, además de mejorar la calidad de la información obtenida a partir de esos datos.

Transformación de Datos: Consiste en la normalización de los mismos. Este paso implica la transformación del tipo de algunos atributos, en caso que fuera necesario, teniendo presente que convertir el tipo de un atributo a otro puede cambiar la semántica de dicho atributo. Hay que considerar que la normalización cambia un poco los datos con los que se cuenta al principio.

Reducción de Datos: En este paso se disminuye el tamaño de los datos, eliminando características redundantes. Las diferentes técnicas de reducción de datos son utilizadas para obtener muestras o representaciones más pequeñas de los datos manteniendo la integridad de los mismos.

1.8 La supervisión de procesos.

Se entiende como *supervisión* de un proceso al conjunto de acciones desempeñadas con el propósito de asegurar el correcto funcionamieto del proceso incluso en situaciones anómalas. De hecho, se puede afirmar que la supervisión está presente en cualquier proceso productivo y que se realiza a través de encargados y operarios especializados, que detectan la presencia de comportamientos anómalos y actúan en consecuencia (ajustando parámetros, cambiando consignas y activando accionamientos para prevenir un mal superior o conservar la capacidad operativa del proceso). Se trata de dar al operador o encargado de control el máximo soporte, liberándolo de la tensión que supone una vigilancia constante y de las tareas rutinarias (elaborción de informes periódicos,

lectura y comparación de registros que garantizan el orden y la sistematización anhelada en los planes de calidad,etc.).

El propósito de la supervisión es la automatización de estas tareas. Para ello debe sacarse provecho de toda información y conocimento disponible sobre el proceso. La dificultad de tales sistemas reside en la diversidad de procesos existentes y las diferentes manifestaciones del conocimiento que sobre estos se dispone. Debido a estos y otros inconvenientes, hoy en día no es posible, todavía, cerrar el lazo que supone la supervición sin incluir al operario humano.

El objetivo de la supervisión es asegurar el orden aún cuando haya desviaciones no previstas en la automatización. Por este motivo se establece la supervisión en un nivel jeráquicamente superior a la automatización y con una tarea clara de vigilancia. Para ello deberá disponer de las siguientes capacidades:

- Registrar la evolución del proceso y detectar desviaciones en las variables.
- Analizar estas desviaciones y deducir el motivo.
- Elaborar un diagnóstio de la situación.
- Resolver situaciones conflictivas en línea.
- Tomar las medidas adecuadas para que no vuelva a suceder.

Dependiendo del horizonte de tiempo en el que se trabaja, la supervisión se puede aplicar a 2 niveles, (Seisdedos, Blanco, Peña, & Rodríguez, 2014):

A corto plazo: A este nivel, las variables de proceso se observan continuamente con la meta de detectar cualquier desviación respecto del estado normal del proceso y de reaccionar lo más rápidamente para asegurar la operación normal de la planta. El término monitorización se utiliza más frecuentemente para referirse a este nivel con un énfasis en la detección e identificación de fallos.

A largo plazo: A este nivel se analiza el comportamiento del proceso a largo plazo y a través de los datos históricos con la meta de identificar causas de bajo rendimiento y oportunidades de mejora. Los términos Análisis del Proceso o Mejora del Proceso se utilizan con cierta frecuencia en la literatura para designar este tipo de supervisión (Wang, 2001) y (MacGregor, 2004).

1.9 Estados estacionarios. Importancia.

El tiempo en que un proceso opera en estado estacionario es una indicación del rendimiento o la eficacia de su diseño del proceso, control, optimización y operación manual. Aunque algunos procesos operan en un estado de inestabilidad sostenida debido al carácter cíclico o caótico de la cinética de reacción subyacentes, transferencia de calor y mecánica de fluidos, se puede argumentar que el escenario operativo preferido de la mayoría de los procesos industriales continuos del mundo es estar en un estado de estabilidad.

El avance y la instalación de microprocesadores industriales basado en controladores lógicos programables, sistemas de control distribuido, dispositivos de medición inteligentes montados en el campo y control avanzado de procesos son otras evidencias de la necesidad de procesos continuos que funcionan en el estado estacionario ya que su objetivo principal es regular el proceso en torno a algún punto de operación predefinido.

La identificación del estado estacionario es una tarea importante para un control satisfactorio de muchos procesos. Los modelos de estado estacionario son muy utilizados en las funciones de evaluación del desempeño y rendimiento del proceso, identificación y ajuste del modelo, para seleccionar los segmentos de datos para modelado, en la optimización y control, detección de fallos, análisis de sensores, reconciliación de datos, análisis de procesos, formación de redes neuronales y para activar las intervenciones en línea. Además, en línea, la detección de los estados estacionarios en tiempo real se utiliza para activar la siguiente etapa de un plan experimental o fase del proceso.

Hasta la fecha, una de las razones más populares para la aplicación de un algoritmo de detección de estado estacionario en una planta es llevar a cabo la reconciliación de datos y optimizadores económicos que procedan cuando se observa un estado de equilibrio. Si una falsa indicación de una operación de estado estacionario se señala, esto podría potencialmente considerar errores gruesos y mediciones malas como aceptables, dejando la entrada al optimizador de tiempo real como cuestionable. En consecuencia, la descarga directa de la solución optimizada para

los puntos de ajuste de la capa de control avanzado y regulador podría tener implicaciones graves cuando las condiciones de estado estacionario son poco reconocidas. A pesar de que todos los optimizadores de proceso disponibles comercialmente tienen comprobación de límite simple y las funciones de rampa para impedir cambios excesivos en los puntos de ajuste, la comprensión de los efectos de la acumulación de material y energía en la fidelidad o la solidez de las ecuaciones del modelo a estado estacionario es débil (es decir, las estimaciones de parámetros ambiguos y ajustes de medición pueden afectar negativamente la solución para el optimizador económico). Además del trabajo de (Kao, Tamhane, & Mah, 1990), (Kao, Tamhane, & Mah, 1992) y (Forbes & Marlin, 1996) poco se ha hecho para evaluar la sensibilidad de la información serialmente correlacionada tanto en la detección de errores gruesos y la optimización económica de una planta real. Esto demuestra además la necesidad de un algoritmo de confirmación de estado estacionario preciso.

1.9.1 Técnicas para la identificación de estados estacionarios.

Existen diferentes técnicas para la detección de estados estacionarios, las cuales pueden ser clasificadas como las siguientes:

1. **Técnica basada en la regresión lineal** con el objetivo de determinar el mejor ajuste con tendencia lineal para N valores anteriores de la variable. Heurísticamente, si el proceso está en estado estacionario, entonces la pendiente de la línea de tendencia será idénticamente igual a cero. Sin embargo, debido a la presencia de ruido en el proceso, la pendiente puede fluctuar con valores cercanos a cero y consecuentemente, un valor de la pendiente desigual de cero, no es razón para rechazar la hipótesis de estado estacionario. Para aceptar o rechazar la hipótesis podría usarse una prueba basada en la estadística T de Student aplicada al cociente entre el valor de la pendiente y el error estándar de ese valor. Si el cociente mencionado excede el valor crítico puede considerarse que hay suficiente evidencia para rechazar la hipótesis de que el proceso está en estado estacionario.

2. **Técnica basada en la evaluación de los valores promedios en ventanas sucesivas de datos**. La misma se basa en seleccionar una ventana de datos, calculando ya sea el promedio, la varianza, o la pendiente de regresión de esta ventana de datos y comparando estos resultados con los mismos resultados anteriores usando las pruebas estadísticas apropiadas.

   Una estrategia basada en gráficas de medias móviles comunes para el control de proceso estadístico involucra el uso de la media móvil y límites superiores e inferiores de la desviación estándar para detectar al estado no estacionario. Todavía otro método en esta categoría se basa en la relación de varianza calculada por dos métodos diferentes en la misma ventana de datos. Durante un estado estable la relación de varianza estará cerca de la unidad; mientras que, en un estado no estacionario, esta relación tendrá un valor mucho mayor que la unidad. En todos estos métodos, necesitamos seleccionar una ventana de datos y realizar cálculos sobre el set de datos. Entre más grande sea la ventana de datos, mayor retardo existirá para detectar cambios de estado, mientras que si la ventana es pequeña se tiende a causar un efecto adverso en las pruebas estadísticas por el ruido asociado con las mediciones.

   Estos métodos necesitan la suposición específica de que, en el lapso de tiempo seleccionado, las variables del proceso están en estado estacionario; pero en la práctica actual es muy difícil escoger períodos de tiempos sucesivos satisfaciendo esta suposición.

3. **Técnica basada en el análisis de la varianza.** Se calculan valores de la media y la varianza exponencialmente ponderadas y se aplica una prueba derivada de la prueba de Fisher para aceptar o rechazar la hipótesis de estado estacionario. Dentro de esta técnica encontramos el método de (Cao & Rhinehart, 1995); que es el método más práctico y citado por muchos autores en la literatura de Control de Procesos.

El desarrollo de una técnica viable para la identificación de estados estacionarios ayudará a darse cuenta de los beneficios de las técnicas de control de procesos en línea o basado en un modelo. Debido a la importancia de lo antes mencionado, se

hace necesario el estudio de métodos robustos para lograr este objetivo. Escoger el método más robusto para nuestro proceso determinará en gran medida el éxito y la confiabilidad del sistema implementado. A continuación, se presenta una revisión de las técnicas más utilizadas para la identificación de estados estacionarios.

1.9.2 Estudios previos realizados en el área de DEE.

Prueba de Interpolación Polinomial. *[Polynomial Interpolation Test (PIT)*]

En (Savitzky & Golay, 1964) desarrollaron un algoritmo para un filtro para tratar información medida en los procesos con ruido, como la espectroscopia. Una serie experimental de medida es primeramente filtrada, escogiendo un tamaño de la ventana *n* (el cual debe ser un número impar). Cada ventana es interpolada usando un polinomio de grado *p*, con *p<n*. La información obtenida a partir del polinomio interpolado es menos ruidosa. Así, la primera derivada de cada polinomio en los puntos céntricos se calcula y el valor es utilizado como un estadístico para evaluar la estacionalidad del punto. Los parámetros del filtro son el tamaño de la ventana, *n*, y el grado de polinomio, *p*.

Prueba del rango de von Neumann. [*Rank von Neumann Test*]

Se aplica la modificación de rango de la prueba de von Neumann para la independencia de datos, como está descrita en (Bartels, 1982) y (Madansky, 1988). Aunque la identificación de estados estacionarios no es la meta original de esta técnica, indica si una serie temporal no tiene correlación de tiempo y así puede usarse para inferir que hay sólo ruido aleatorio añadido a un comportamiento estacionario. En esta prueba una relación *V* es calculada de la serie temporal, cuya distribución se espera sea normal con una desviación estándar conocida, para así confirmar la estacionalidad de un set específico de puntos.

Un artículo detallado dado por (Narasimhan, Mah, Tamhane, Woodward, & Hale, 1986) que describe el uso de una estadística multivariante fue uno de los primeros en proporcionar un enfoque teórico y estadístico para el problema de la DEE. Mediante la estimación de las medias y las matrices de covarianza en línea por períodos sucesivos que contienen N observaciones, que calculan una estadística Hotelling T^2 con el fin de detectar diferencias apreciables entre el período de dos

vectores de medias. Ellos utilizan una matriz de dispersión agrupada mediante la cual asumieron que las mediciones sólo contenían errores aleatorios (es decir, no hay correlación serial). Más tarde, (Narasimhan, Kao, & Mah, 1987) proporciona un enfoque alternativo basado en la teoría matemática de las evidencias, sin embargo, esto requiere que la matriz de covarianza sea diagonal lo cual parece ser una limitación grave. Dos parámetros que eran utilizados para definir la estadística, N y p (el tamaño de la ventana de detección y el número de mediciones para incluir en la prueba) fueron analizados para mostrar cómo podrían mejorar la capacidad o la potencia de la prueba para detectar sucesivos períodos de estado estacionario. Su método está destinado a procesos cuasi estacionarios con movimiento lento debido a los armónicos de frecuencias más bajas, tales como las variaciones diurnas y estacionales. Su estudio debe aplicarse a los procesos cuyas variables exhiben estacionalidad sincrónica; es decir, los procesos cuyas mediciones avanzan juntas en el tiempo y pueden o no ser mutuamente correlacionadas. De hecho, éstos son los tipos de procesos que nos ocupan en este estudio ya que se cree que son la mayoría. Dada la naturaleza relativamente a largo plazo derivada de muchos parámetros del proceso tales como composiciones de materia prima, coeficientes de ensuciamiento de intercambiadores de calor, actividades catalizadoras, la limpieza de los filtros en línea, la frescura de tratadores de productos y el uso de sellos mecánicos y cubiertas, no hay ninguna duda de que este tipo de comportamiento del proceso son realistas y representativos. Del mismo modo, estas tendencias de deriva lenta pueden ser consideradas como componentes estacionales desconocidos (Box & Jenkins, 1976) superpuestas sobre las respuestas de mediciones ya correlacionadas en serie. Fundamentalmente, estas derivas causales constituyen un segundo armónico, por así decirlo, en los datos de series de tiempo que puede analizarse adicionalmente utilizando una técnica similar con un intervalo de control adecuado (es decir, la aplicación de un segundo programa de detección del estado estacionario).

En (Holly, Cook, & Crowe, August, 1989) destacaron un manejo práctico y univariado de la cuestión que implicó la realización de las pruebas *T*-Student en las medias y pruebas de hipótesis sobre las varianzas de cada una de las mediciones

de proceso a través de una ventana de detección predefinida que también incluía un test de pendientes por períodos sucesivos para ayudar a asegurar que los períodos de estado estacionario aparente fueran señalados. Se utilizaron test de pendientes para señalar las derivas lentas en el proceso, pero la palabra lenta es un término relativo, dado que se requiere un conocimiento a priori de la longitud de una deriva con el fin de determinar el número de ventanas adyacentes para incluir en la prueba. Además, su técnica no tuvo en cuenta la existencia de observaciones relacionadas en el tiempo y mediciones correlacionadas entre sí, lo que reduce el poder de las pruebas y artificialmente aumenta la probabilidad de errores de Tipo I.

Prueba F modificada*. (Modified F Test)*

En (Cao & Rhinehart, 1995) describen un enfoque para identificar estados estacionarios usando estadística *F* siendo la relación de dos varianzas filtradas derivadas para cada medición. Estas variaciones se calculan apropiadamente pasando las respuestas de medición en alguna forma a través de tres ecuaciones de promedio móvil de tipo filtro exponencialmente ponderado (EWMA, por sus siglas en inglés, *Exponentially Weighted Moving Average*) en la que cada ecuación requiere un factor de filtro para ser seleccionada. Una cuestión práctica surge con este tipo de enfoque, dado que no está claro en su método en cuanto a cómo manejar eficaz y apropiadamente sistemas con más de una medición (es decir, para los tres ejemplos descritos en su documento, sólo se utilizó una medición para cada uno de los tres procesos) aunque potencialmente el uso de técnicas de promedio móvil de tipo cartas de control (MEWMA, por sus siglas en inglés*, Multivariate Exponentially Weighted Moving Average*) que recientemente se han hecho disponibles podrían ser aplicadas (Lowry & Woodall, 1992); (Prabhu & Runger, 1997). Sin embargo, los métodos MEWMA también explotan el uso de la estadística de Hotelling T^2 y también requieren los errores del sistema sean independientes e idénticamente distribuidos. En última instancia, cualquier tratamiento general de detección del estado estacionario debería ser fácilmente extensible para incluir la estructura multivariable del sistema subyacente ya que es la naturaleza colectiva del proceso bajo pruebas y no una sola medición.

Un enfoque interesante por (Tong & Crowe, 1997) ha acuñado el término estado estacionario de "acumulación cero" y utiliza la técnica de análisis secuencial de los residuos de las limitaciones de materiales del proceso o balances tanto individual como colectivamente para determinar si se ha producido alguna acumulación significativa de materiales en el proceso. Dado que tanto los estados estacionarios como los procesos estacionarios son efectivamente procesos de acumulación cero, la detección de la acumulación cero es una alternativa muy viable donde se ha demostrado por (Zhang & Pollard, 1994) que la función de auto-correlación de los desequilibrios de materiales para procesos simples puede mostrar menor correlación que para las propias mediciones individuales. Si no se ha detectado ninguna acumulación (o incluso acumulación constante), entonces el proceso se considera adecuado para la reconciliación de datos de estado estacionario y la detección de errores gruesos. Como las pruebas se calculan secuencialmente cuando un nuevo conjunto de mediciones de tamaño p está disponible en cada intervalo de muestreo, la detección de estado estacionario es más rápida en comparación con otros métodos que utilizan un período fijo y es general para mediciones dependientes del tiempo (Stoumbos & Reynolds, 1997); (Tong & Crowe, 1997). No obstante, la frecuencia a la que se toman las muestras en última instancia, influye en la probabilidad de errores de tipo I y de su poder cuando la correlación serie está presente (Zhang & Pollard, 1994). Sin embargo, este procedimiento tal como se aplica, requiere un conocimiento previo de la matriz de covarianza de los residuales de balance que puede cambiar de un estado estacionario a otro debido a variaciones en las mediciones, aunque parece relativamente sencillo para modificar su prueba para incluir una estructura de covarianza cambiante. Por otra parte, se requiere un modelo de estado estacionario de balance de materiales del proceso, que en muchos casos demanda de conocimiento hasta ese momento de las densidades de flujo, ya sea a partir de la información de laboratorio o de costosa instrumentación de campo. Por otra parte, si se cuenta con mediciones de temperatura y presión precisas, el método de (Tong & Crowe, 1997) requiere una extensión para permitir balances de materia y energía simultáneos (limitaciones bilineales) para permitir mediciones no fluidas en el esquema de detección.

También se ha prestado atención reciente a la reconciliación y la supervisión de los sistemas dinámicos (Albuquerque & Biegler, 1996); (Bagajewicz & Jiang, 1997), que por supuesto no requieren una evaluación de estado estacionario antes de su ejecución. En la literatura se ha reportado también interés relacionado en el control estadístico de los procesos de los sistemas dinámicos multivariables (Ku, Storer, & Georgakis, 1995); (Negiz & Cinar, 1997) con el fin de tener en cuenta la presencia de y auto-correlaciones y relaciones cruzadas en el seguimiento de los procesos continuos en tiempo real. Aunque no vamos a discutir la conveniencia de conciliación dinámica de datos con respecto a cuestiones tales como el aumento de los requisitos de complejidad, de incertidumbre y de información de los materiales y ecuaciones energéticas del modelo dinámico, vemos como un inconveniente, que estas técnicas requieren todavía una estimación a priori de la matriz de covarianza de las mediciones con el fin de llevar a cabo el análisis posterior crucial de los resultados de la reconciliación (detección de errores gruesos e identificación). Por esa razón, la clasificación de EE de la operación de proceso es todavía importante para proporcionar estimaciones razonables de la matriz de covarianza de las mediciones que no sólo deben incluir la incertidumbre de los dispositivos de medición, sino también la variabilidad intrínseca del proceso en sí (Holly, Cook, & Crowe, August, 1989). Curiosamente, (Chen, Bandoni, & Romagnoli, 1997) proponen un método para estimar robustamente la matriz de covarianza de las mediciones indirectamente utilizando los residuos de balance de materiales. Robusto significa que el efecto de los valores atípicos (desviaciones de estado estacionario y/o errores gruesos en las mediciones) se atenúan mediante la distancia de Mahalanobis. Sin embargo, este enfoque podría ser mejorado si los datos se recogieran durante períodos de estado estacionario, no obstante, no se considera el efecto de las observaciones correlacionadas de las mediciones de procesos estacionarios.

Prueba de los Arreglos Inversos *[Reverse Arrangements Test (RAT)]*

Una prueba no paramétrica es la Prueba de los Arreglos Inversos, en la cual un estadístico, designado *A*, se calcula para evaluar la tendencia de una serie temporal. El procedimiento exacto de cálculo, así como también las tablas conteniendo

intervalos de confianza está descrito en (Bendat & Piersol, 2000). Si *A* es demasiado grande o demasiado pequeña comparada con estos valores estándar podría significar que hay una tendencia significativa en los datos, por consiguiente, el proceso no debería ser considerado en estado estable. La prueba es aplicada secuencialmente para ventanas de datos de un tamaño dado.

Narasimhan, Chen y Liao presentaron un método de pruebas en EE basado en técnicas estadísticas (Chen & Liao, 2002). El método asume que sólo los errores al azar existen con media cero, y los períodos de tiempo sucesivos definidos de cada una de ellas se asumen constantes. En la primera etapa, una prueba se aplica para determinar si las matrices de covarianza de dos períodos consecutivos son iguales, entonces las medias de los dos períodos se analizarán usando la prueba de Hotelling. Un método propuesto por (Narasimhan, Mah, Tamhane, Woodward, & Hale, 1986), utiliza una Estadística H-T^2, que se define mediante la comparación de la diferencia de medias entre los períodos de la variabilidad dentro de los plazos.

Conclusiones Parciales.

1. Se realizó un recuento histórico acerca de la evolución de los sistemas de control, así como los métodos para la supervisión y almacenamiento de las series temporales.
2. Se abordó el tema de los Sistemas de Control Distribuidos, de la importancia de la implementación de los mismos y de las ventajas obtenidas de esta implementación.
3. Se exponen algunos aspectos de los servidores OPC donde se hace énfasis en los tipos de servidores existentes y en su utilidad.
4. Se presentan algunas generalidades del Toolbox “GUIDE” de Matlab™ y de su utilización para realizar interfaces gráficas.
5. Se dan a conocer aspectos relevantes de la minería de datos, la Quimiometría y de cómo realizar correctamente la detección y corrección de errores; para luego llegar a las diferentes técnicas para pre-procesar datos.
6. Finalmente se describe en qué consiste la detección de estados estacionarios y su importancia para la industria, se explican diferentes métodos matemáticos para detectarlos y se explica el método implementado en este trabajo.

CAPÍTULO 2: SSITS y su trabajo con series temporales.

Introducción.

En toda industria de generación de electricidad es importante mantener un estricto control de las variables que están vinculadas a dicho proceso productivo. Toda planta de generación de electricidad debe trabajar en un buen estado técnico y dentro de los rangos correctos de operación con el fin de cumplir con su objetivo garantizando la rentabilidad económica. Para esto, se necesita una herramienta capaz de analizar los parámetros fundamentales de dichas plantas y que permita estudiar detalladamente las variables del proceso que se encuentran en estados estacionarios; así como los aspectos claves de las mismas. La herramienta informática aquí desarrollada tiene como objetivo presentar una técnica que compense la existencia de correlación serial y mutua en los ruidos de medición y del proceso, dado que esto puede afectar negativamente a la cantidad de errores de Tipo I en la detección de EE. A continuación, se describen sus aspectos técnicos, fundamentos teóricos y sus ventajas con respecto a otros métodos empleados para la detección de EE.

2.1 Definición del método para detección de estado estacionario.

Para la detección de los intervalos a estado estacionario se implementa un método que se puede resumir de la siguiente manera; en cada serie sucesiva de N observaciones de las p mediciones, agrupar la matriz de covarianza de la muestra con la del período anterior, y luego llevar a cabo una compensación en la matriz de covarianza agrupada con el fin de superar el problema de la inflación de la estadística T^2 debido a la existencia muy probable de auto-correlación en las mediciones (Kao et al., 1990, Harris y Ross, 1991). Para los períodos en que se detecta un cambio de estado estacionario, se lleva a cabo una factorización de Cholesky de la matriz de covarianza para evaluar la contribución de la medición. Al transformar las diferencias en las medias entre los dos períodos adyacentes utilizando el inverso del factor Cholesky inferior, se llega a p nuevos vectores latentes o puntuaciones que son simplemente combinaciones lineales de las

diferencias de medias entre los períodos sucesivos. Estos vectores latentes son más rápidos de calcular y contienen casi la mitad de los que no son ceros que el popular Análisis de Componentes Principales (PCA, del inglés *Principal Component Analysis*) con los valores propios-vectores propios de la matriz de covarianza. Aplicando un análisis "modificado" de la contribución de (Kourti & MacGregor, 1996) en las puntuaciones, seríamos capaces de proporcionar alguna indicación del peso de cada variable como contribuyente potencial global de la operación no estacionaria.

En este tratamiento, se trata de identificar la causa de una violación de la prueba T^2 atribuyéndola a mediciones de procesos sospechosos cuyas medias han cambiado bruscamente. En este nuevo enfoque se asume la hipótesis nula de que el proceso es estacionario acerca de su media, sometido a errores aleatorios independientes e idénticamente distribuidos; contra la hipótesis alternativa de que es un proceso no estacionario con pendiente diferente de cero, detectable y determinística, así definimos:

Hipótesis nula----------> Proceso estacionario

Hipótesis alternativa-----> Proceso no estacionario

Como se ha indicado, la intención con esta detección es indicar los períodos en que el proceso no presenta ninguna acumulación apreciable de material y energía. Es una condición necesaria, al menos macroscópicamente, no observar ninguna diferencia estadísticamente entre las medias de las p mediciones de proceso calculadas a partir de un período al siguiente, lo que implica un proceso que se encuentra en estado *cuasi estacionario*. Con el fin de plantear esto en términos estadísticos, se deben hace algunas suposiciones tomadas directamente de (Narasimhan, Mah, Tamhane, Woodward, & Hale, 1986):

1- Todas las mediciones de proceso se miden directamente.

2- Las mediciones sólo contienen errores aleatorios, que son normalmente distribuidos con media 0.

3- Se supone conocida la matriz de covarianza de los errores de medición y se le permite cambiar de un período al siguiente.

4- Los vectores de medición sucesivos son independientes entre sí.

Con las consideraciones anteriores, el modelo de medición puede ser caracterizado como sigue:

$x(t) = \mu(t) + \eta(t) + \omega(t)$……..……[1]

Aquí, la realización de tiempo continuo de cada medición del proceso es función de tres términos:

$\mu(t)$ el valor esperado o verdadero para el estado de la medición que puede o no contener las tendencias de movimiento lento como se ha mencionado, pero si está presente una no estacionalidad se asume que cambia bruscamente.

$\eta(t)$ la naturaleza probabilística o estocástico de los ruidos de medición y de proceso y superpuestas.

$\omega(t)$ la manifestación de los errores sistemáticos o gruesos presentes en la medición debido a un mal funcionamiento y similares. Para este tratamiento de detección de estado estacionario, se supone que las mediciones están libres de errores gruesos (es decir, $\omega(t) = 0$).

$x(t) = \mu(t) + \eta(t)$……..…[2]

2.1.1 Generación de datos para validación.

Para la comprobación de los algoritmos propuestos, se generan una serie de ‘’señales’’ de las cuales se conocen todas sus características y comportamiento. Básicamente consisten en unos patrones escalonados (**Figura 2.1**) que representan los intervalos de EE y los ET (de aquí que ya sean conocidos) y a éstos se le añade ruido con las características descritas más adelante. De esta manera se conoce cuál debería ser la salida de cualquier método de detección de EE y se puede evaluar si el planteado en este trabajo devuelve resultados correctos.

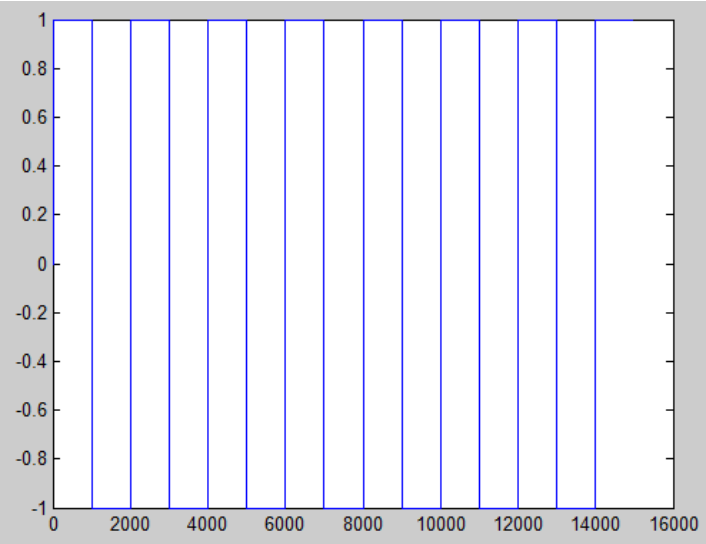


Figura 2.1: Ejemplo de señal para detectar los EE.

Dado que los procesos estacionarios pueden exhibir dependencia entre observaciones sucesivas (Vasilopoulos & Stamboulis, 1978); (Harris & Ross, 1991), una base más realista es suponer que se pueden modelar como un proceso de media móvil autorregresiva de orden *r, q* (es decir, ARMA (*r, q*) en la notación de (Box & Jenkins, 1976)). Por consiguiente, $\eta(t)$ puede representarse en forma discreta en el tiempo por:

$$\eta_t = \frac{1-\theta_1 z^{-1}-\cdots-\theta_q z^{-q}}{1-\phi_1 z^{-1}-\cdots-\phi_r z^{-r}}\varepsilon_t \quad \text{.....[3]}$$

donde el numerador corresponde a la dinámica de media móvil y el denominador corresponde a la dinámica autorregresiva.

Se consideran procesos con un máximo de 5 variables donde se aplica la correlación cruzada de los datos a través de la matriz simétrica *H*. Un ejemplo de un set de datos generados con los siguientes parámetros para cada una de las 5 variables, es mostrado a continuación:

Tabla 2.1. Ejemplo de parámetros necesarios para la generación de datos.

i	$\theta_{i,1}$	$\phi_{i,1}$	$\sigma^2{}_{\varepsilon,i}$
1	0.70	0.95	1.00

2	0.60	0.90	0.90
3	0.50	0.85	0.80
4	0.40	0.80	0.70
5	0.30	0.75	0.60

Donde $\phi_{i,1}$ representa el parámetro autorregresivo de primer orden y $\sigma^2{}_{\varepsilon,i}$ la varianza del ruido blanco.

Para ilustrar la cantidad de correlación conjunta simulado en este ejemplo, se calcularon dos matrices de covarianza típicas no compensadas agrupadas $\bar{\mathbf{S}}_{k+1}$ para el mismo período de muestreo a través de 150 puntos (*N*=150) con ***H*** dado en el **Anexo 2** y ***H=0*** . Las dos matrices se muestran a continuación, respectivamente, y se han convertido a matrices de correlación con el fin de resaltar las diferencias.

$$H1=\begin{bmatrix} 1 & -0.65 & 0.56 & -0.52 & 0.53 \\ -0.65 & 1 & -0.57 & 0.53 & -0.58 \\ 0.56 & -0.57 & 1 & -0.60 & 0.55 \\ -0.52 & 0.59 & -0.60 & 1 & -0.56 \\ 0.53 & -0.58 & 0.55 & -0.56 & 1 \end{bmatrix} \quad H2=\begin{bmatrix} 1 & -0.37 & -0.03 & -0.17 & 0.15 \\ -0.37 & 1 & 0.16 & -0.02 & -0.12 \\ -0.03 & 0.16 & 1 & -0.05 & -0.02 \\ -0.17 & -0.02 & -0.05 & 1 & -0.02 \\ 0.15 & -0.12 & -0.02 & -0.02 & 1 \end{bmatrix}$$

Claramente, la primera matriz de correlación tiene correlación mutua más significativamente que la segunda, como se puede ver mediante la comparación de los elementos fuera de la diagonal.

2.1.2 Enunciación matemática del método de detección de EE.

Si definimos $\bar{\mathbf{x}}_{k+1}$ como el vector medio para las p mediciones tomadas sobre *N* muestras en el *k*+1th período, luego pueden denotarse dos hipótesis para detectar cambios en el medio entre dos períodos sucesivos, k y *k*+1th, como:

$H_0: \bar{x}_{k+1} = \bar{x}_k$

$H_1: \bar{x}_{k+1} \neq \bar{x}_k$ ..... [4]

siendo necesaria la aceptación de H_0 para que un proceso sea declarado en estado estacionario del período *k* a *k+1*. Cuando la matriz de covarianza se calcula a partir de los datos, se puede realizar una prueba estadística multivariable mediante el cálculo de la conocida estadística de Hotelling (Narasimhan, Mah, Tamhane, Woodward, & Hale, 1986)

$$T_{k+1}^2 = \frac{N}{2}(\bar{x}_{k+1} - \bar{x}_k)^T \bar{S}_{k+1}^{-1}(\bar{x}_{k+1} - \bar{x}_k) \quad \text{.....[5]}$$

donde *N* es el número de mediciones en un período y $\bar{S}_{k+1}$ es la matriz de covarianza agrupada de tamaño *p x p* calculada como:

$$\bar{S}_{k+1} = \frac{(S_{k+1}+S_k)}{2} \quad \text{.....[6]}$$

La covarianza de la muestra o matriz de dispersión para el período k+1 se calcula:

$$S_{k+1} = \frac{1}{N-1}\sum_{n=1}^{N}([X_{k+1}]_n - \bar{x}_{k+1})([X_{k+1}]_n - \bar{x}_{k+1})^T \quad \text{.....[7]}$$

donde $[X_{k+1}]_n$ se define como la n^{th} columna de la matriz X_{k+1} que es de tamaño *pxN* que contiene las *N* muestras para las *p* mediciones con S_k calculada del mismo modo. Ya que T_{k+1}^2 se distribuye bajo la hipótesis nula como *2p(N-1)/(2N-p-1)* veces una variable aleatoria *F* con *p* y 2*N*-*p*-1 grados de libertad (Narasimhan, Mah, Tamhane, Woodward, & Hale, 1986) , luego el nivel superior de significación α bajo la hipótesis nula está dado por:

$$T_\alpha^2 = \frac{2p(N-1)}{2N-p-1}F_{\alpha,p,2N-p-1} \quad \text{.....[8]}$$

En consecuencia, lo anterior es nuestra estimación a priori del valor umbral de la estadística de Hotelling desde la que se declara que el proceso es inestable si se excede este valor (es decir, rechazar la hipótesis nula y aceptar la alternativa).

2.2 Compensando el efecto de la correlación serial.

El uso de la estadística T^2 asume errores de medición independientes y aleatorios lo que implica que no existe correlación temporal, aunque la correlación espacial o mutua fácilmente se manipula dado que T^2 es una estadística multivariada. Por otra parte, debido a los elementos inerciales tales como tanques, reactores y otros

recipientes (Harris & Ross, 1991), la correlación serial está casi siempre presente en los sistemas y puede ser exagerada por la elección del intervalo de muestreo. Sin embargo, dado que los procesos estacionarios están siendo detectados por los cambios estadísticamente significativos en sus medias, en un sentido estricto, el uso de T^2 (o en realidad cualquier procedimiento de control de gráficos como el Shewhart, CUSUM y EWMA) no es apropiado, como plantean (Vasilopoulos & Stamboulis, 1978); (Harris & Ross, 1991); (Wardell, Moskowitz, & Plante, 1994); (Ku, Storer, & Georgakis, 1995). El efecto indeseable de correlación serial de T^2 es que arbitrariamente infla su valor y produce un aumento de los errores de Tipo I, haciendo la detección propensa a falsos positivos.

Con el fin de eludir esto y para hacer T^2 más útil, se emplea un método de compensación de covarianza utilizando como un componente clave la matriz de auto correlación V_{k+1}, cuyos elementos $[V_{k+1}]_{m,i}$ son el *m*th coeficiente de auto correlación de retraso para la *i*th medición y se calcula como sigue:

$$[V_{k+1}]_{m,i} = \frac{\sum_{n=m+1}^{2N}([X_{k\dots k+1}]_{i,n-m} - [\bar{x}_{k\dots k+1}]_i)([X_{k\dots k+1}]_{i,n} - [\bar{x}_{k\dots k+1}]_i)}{\sum_{n=1}^{2N}([X_{k\dots k+1}]_{i,n} - [\bar{x}_{k\dots k+1}]_i)([X_{k\dots k+1}]_{i,n} - [\bar{x}_{k\dots k+1}]_i)} \forall\, m = 0,1,\dots M$$

……..[9]

Este estudio sigue el trabajo de (Kao, Tamhane, & Mah, 1990), quienes para la comprobación estadística univariada del error grueso frente a la correlación serial, compensan la estimación univariada de la varianza para cada medición para atenuar el efecto de la correlación serial. Su técnica de compensación de la varianza fue tomada directamente de (Vasilopoulos & Stamboulis, 1978) que ajusta la estimación de la varianza de la muestra para cada medición como una función de su respectiva función de auto correlación. Esto forma la base de nuestra técnica de compensación también. A continuación, se presenta la expresión necesaria para calcular los elementos del vector W_{k+1} que representa el término de compensación obtenido por (Vasilopoulos & Stamboulis, 1978) para la i-ésima medición en el período *k*+1

$$[W_{k+1}]_i = 1 + \frac{2}{N}\{(N-1)[V_{k+1}]_{1,i} + (N-2)[V_{k+1}]_{2,i} + \cdots + (N-m)[V_{k+1}]_{m,i} + (N-M)[V_{k+1}]_{M,i}\}.$$[10]

donde $[W_{k+1}]_i$ es el i^{th} valor de compensación de la variable y su cálculo es general para cualquier orden ARMA. Los $[W_{k+1}]_i$ sirven para ajustar las varianzas de las muestras y, o bien inflar o desinflar la función de los valores de θ y ϕ .

A partir de lo anterior, existen expresiones para la compensación univariada de las varianzas de las muestras de medición individuales (las diagonales de $\bar{S}_{k+1}$), que se pueden aplicar cuando existen errores de medición no independientes. Sin embargo, los procesos analizados aquí son de naturaleza multivariable. El dilema es ¿cómo compensar la matriz de covarianza agrupada $\bar{S}_{k+1}$? Dado que no parece haber ninguna literatura disponible en la actualidad (hasta donde se conoce) que nos guíe con esta compensación para sistemas multivariables, el enfoque pragmático adoptado aquí es pre y post multiplicar $\bar{S}_{k+1}$por una matriz diagonal que contiene las raíces cuadradas de W_{k+1} mostrada a continuación:

$$\tilde{S}_{k+1} = diag(W_{k+1}^{1/2})\bar{S}_{k+1}diag(W_{k+1}^{1/2}) \quad \text{.....[11]}$$

donde la estadística de Hotelling compensada es re-calcula como, (ver **Anexo 3**):

$$\tilde{T}_{k+1}^2 = \frac{N}{2}(\bar{X}_{k+1} - \bar{X}_k)^T S_{k+1}^{-1}(\bar{X}_{k+1} - \bar{X}_k) \quad \text{.....[12]}$$

Para ilustrar la ventaja de utilizar esta estrategia que utiliza W_{k+1}, consideramos un ejemplo muy simple que implica un proceso con 5 variables de medición donde se aplica solamente ruido blanco aleatorio normal usando las varianzas que se proporcionan en la **Tabla 2.1** (para $\sigma_{\varepsilon,i}^2$). Para las simulaciones, la intención es comparar la probabilidad el error Tipo I calculado frente al nivel de significación teórica de 5,0% seleccionado (es decir, un intervalo de confianza del 95%) para variar el tamaño de las muestras de *N*. La probabilidad de error Tipo I se define a continuación como:

$$\Pr\{ErrorTipo1\} = \Pr\{H_0 es\ rechazada\ cuando\ las\ p\ variables\ están\ en\ EE\} \text{.....[13]}$$

Para cada fila de la **Tabla 2.2**, 1.000 veces *N* determina el número total de muestras simuladas para un intervalo de muestreo arbitrario. Es decir, para *N*=150, se generaron 150.000 puntos de muestra en el que el número de T_{k+1}^2 mayor que el

valor crítico teórico, que se muestra en la columna 2, se divide por 1.000 y el porcentaje convertido para estimar α . Debe tenerse en cuenta que se utilizó un ajuste heurístico de *M* igual a 10% de *N* para todas las simulaciones (por ejemplo, si *N*=150 entonces *M*=15). Este enfoque fue elegido con el fin de limitar el número de variables de configuración.

Tabla 2.2. Errores de Tipo I y los valores críticos Hotelling para un simple proceso de ruido blanco (no hay períodos inestables) para diversos N.

N	T_α^2	$\mathbf{W}$ (Teórico)	$T_\alpha^{2,NM}(\mathbf{W})$	$\mathbf{W}_{\mathbf{k+1}}$ (Calculado)	$T_\alpha^{2,NM}(\mathbf{W}_{\mathbf{k+1}})$
50	12.05	5.4	12.70	8.4	14.07
100	11.54	4.8	11.63	7.4	15.60
125	11.44	5.5	11.69	7.4	13.28
150	11.38	4.7	11.40	7.5	13.22
175	11.33	3.8	10.83	7.6	18.35
200	11.30	5.6	11.24	6.7	14.05
225	11.27	4.2	10.88	7.8	16.28
250	11.25	4.1	10.86	8.5	15.61

De los resultados, se observa que para la columna 3 que utiliza la estimación teórica de W (W igual a un vector de unos), en comparación con el nivel de significación del 5% especificado, estos resultados son muy similares. La columna 4 resalta la T^2 calculada usando el método de (Nomikos & MacGregor, 1995), que también es comparable a la validación teórica de T^2 que su enfoque es potencialmente una alternativa viable a T_α^2 . Usando el W_{k+1}, calculado de la **Ecuación 10**, nos damos

cuenta que los resultados de la columna 5 son más altos de lo esperado en un 5%, pero siguen siendo razonables. La columna 6 informa el $T_{\alpha}^{2,NM}$ calculado que es algo más grandes que el W teórico como era de esperar dado los resultados de la columna 5. En consecuencia, si el proceso no está serialmente correlacionado, no se recomienda ninguna compensación. Esto tiene sentido teniendo en cuenta que no se requeriría una compensación en absoluto. Sin embargo, el uso de compensación (W_{k+1}) no infla la estimación del error de Tipo I sustancialmente, (se observa un ligero sesgo) que es una buena característica del método cuando se desconoce en cualquier punto en el tiempo si el proceso contiene correlación temporal. Es decir, el método debe desenvolverse bien para el caso límite cuando no existe ninguna correlación serie.

2.3 Consideración del efecto de la correlación mutua.

La naturaleza multivariable del método viene implícita por el propio hecho de que no se trabaja cada señal independiente, sino que, en cada período se calcula la matriz de covarianza de todas ellas y se concentran en la llamada matriz de covarianza agrupada (**Ecuación 6**). Aquí es necesario destacar que, como concepto estadístico, la covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias. Es el dato básico para determinar si existe una dependencia entre ambas variables y además es el dato necesario para estimar otros parámetros básicos, como el coeficiente de correlación lineal o la recta de regresión. Dicho esto, en el método, todos los procedimientos posteriores se realizan a partir de esta matriz lo cual implica que en todo momento se está considerando la correlación mutua que existe entre las variables bajo estudio.

2.4 Identificación de variables inestables a través de análisis de contribución.

Determinar cuáles de las variables que comprenden la estadística han definido la alarma es una tarea difícil (Mason, Champ, Tracy, Wierda, & Young, 1997a), (Mason, Tracy, & Young, 1997b). Por lo tanto, se requiere una técnica para ayudar en la enumeración de las variables causantes. Aquí se utiliza el análisis de la contribución de (Kourti & MacGregor, 1996) con dos modificaciones.

La primera modificación es no utilizar la técnica de Análisis de Componentes Principales para determinar las variables latentes; en su lugar, se calcula una factorización de Cholesky de la matriz de dispersión compensada que sirve como la matriz de transformación lineal para determinar los nuevos vectores latentes. Esta modificación se elige ya que la factorización de Cholesky es computacionalmente muy eficiente en comparación con el análisis valor propio-vector propio.

La Factorización de Cholesky de $\tilde{S}_{k+1}$ se da como

$$\tilde{S}_{k+1} = \tilde{R}_{k+1}{\tilde{R}_{k+1}}^{T} \quad \text{.....[14]}$$

donde $\widetilde{\mathbf{R}}_{\mathbf{k+1}}$ es el factor de Cholesky triangular inferior de $\tilde{S}_{k+1}$ en el período *k+1*. Un nuevo vector latente entero puede derivarse como

$$t_{k+1} = \sqrt{\frac{N}{2}}{\tilde{R}_{k+1}}^{-1}(\bar{x}_{k+1} - \tilde{x}_k) \quad \text{.....[15]}$$

donde la estadística original de Hotelling se puede calcular fácilmente

$${\tilde{T}_{k+1}}^{2} = {t_{k+1}}^{T} t_{k+1} = \sum_{i=1}^{p} {[t_{k+1}]_i}^{2} \quad \text{.....[16]}$$

donde t_{k+1} es la nueva puntuación del vector de Cholesky. Dado que los vectores de puntuación son combinaciones lineales de las variables originales, la tarea en cuestión es proporcionar alguna indicación de las variables originales cuando el proceso se declara inestable. Con este fin, (Kourti & MacGregor, 1996) han ideado un análisis de la contribución sobre la base de PCA, como se ha mencionado, la cual es capaz de aislar la contribución de cada variable original a $\tilde{T}^2_{k+1}$. La base de su técnica consiste en calcular primero los contribuyentes individuales a los puntajes t_{k+1} como se muestra.

$$[cont_{k+1}]_{a+1} = \sqrt{\frac{N}{2}}\left[{\tilde{R}_{k+1}}^{-1}\right]_{a,i}([\bar{x}_{k+1}]_i - [\bar{x}_k]_i) \quad \forall\, a, i = 1, \ldots, p \quad \text{.....[17]}$$

donde $[cont_{k+1}]_{a,i}$ es la contribución a una a^{th} puntuación, de las que hay *p* de ellas, desde la i^{th} variable de medición (es decir, $cont_{k+1}$ es una matriz de tamaño *p* x *p*). Resumiendo, estas contribuciones individuales a todas las *p* mediciones será igual al valor de una a^{th} puntuación, es decir:

$[t_{k+1}]_a = \sum_{i=1}^{p}[cont_{k+1}]_{a,i} \quad \forall\, a = 1, \ldots, p$.....[18]

El siguiente paso es usar estas $cont_{k+1}$ para aislar la contribución global a la estadística de Hotelling. El enfoque de (Kourti & MacGregor, 1996) es incluir sólo aquellos $[cont_{k+1}]_{a,i}$ que tienen el mismo signo que $[t_{k+1}]_a$ (es decir, donde $[t_{k+1}]_a * [cont_{k+1}]_{a,i} > 0 \; \forall\, i = 1, \ldots, p$) por el que los valores opuestos marcados tienen su $[cont_{k+1}]_{a,i}$ re-ajustados a cero. Esto se razonó por el hecho de que sólo aquellos $[cont_{k+1}]_{a,i}$ que tengan el mismo signo que $[t_{k+1}]_a$ hacen que la estadística se infle; los valores de signo opuesto no son de interés ya que sirven para desinflar los $[t_{k+1}]_a$. Por lo tanto, la contribución global a la inflación de $\tilde{T}_{k+1}^2$ para cada variable de medición, después que su opuestos $[cont_{k+1}]_{a,i}$ se han fijado en cero, es

$[CONT_{k+1}]_i = \sum_{a=1}^{p}[t_{k+1}]_a \, [cont_{k+1}]_{a,1} \;\; \forall\, i = 1, \ldots, p$.....[19]

donde $[CONT_{k+1}]_i$ es la contribución general de la i^{th} variable de medición que puede ser comparada en tamaño con la otra medición $CONT_{k+1}$ para determinar las más grandes.

La segunda modificación con el método de (Kourti & MacGregor, 1996) es una sutil alteración. En (Kourti & MacGregor, 1996) se propuso considerar sólo las puntuaciones en el análisis de la contribución que supere un cierto valor umbral basado en el nivel de significación ajustado por Bonferroni (es decir, idénticos a valor crítico de $t_{u,k+1}$). En este trabajo, no se incluyó este paso basado en que la comprensión de las puntuaciones pequeñas (los que no superen el valor crítico) también se puede incluir en la **Ecuación 19** ya que los valores más pequeños de $[t_{k+1}]_a$ servirán por consiguiente para escalar proporcionalmente el $[CONT_{k+1}]_i$.

Con respecto al rendimiento de las pruebas de detección de estado estacionario y el procedimiento de identificación, se utiliza de nuevo la simulación que se encuentra en el **Anexo 2**. El nivel de rendimiento es dictado por dos simples estadísticas llamadas poder de detección e identificación como se describe a continuación:

$$Poder\ de\ Detección = Pr\left\{\begin{matrix} H_0\ es\ rechazada\ cuando\ una\ o\ más\ de\ las\ p \\ variables\ han\ cambiado\ su\ estado. \end{matrix}\right\}$$.....[20]

$$Poder\ de\ Identificación = Pr\left\{\begin{matrix} H_0\ es\ rechazada\ cuando\ una\ o\ más\ de\ las\ p \\ variables\ han\ cambiado\ su\ estado\ y\ todas \\ las\ variables\ han\ sido\ correctamente \\ identificadas \end{matrix}\right\} \ldots[21]$$

donde el poder de identificación será siempre menor o igual que el poder de detección.

La **Ecuación 22** determina el cambio en la media del proceso, cuando *sólo una* medición es perturbada de la media del período anterior donde δ_i es el cambio requerido en la *i*th media de la medición para desencadenar exactamente la estadística Hotelling. Del mismo modo, la **Ecuación 23** determina el cambio simultáneo en todas las medias de las mediciones, donde δ es el valor de cambio para cada una de las cinco mediciones. Cabe mencionar que, con todas las estimaciones de potencia, que son una función del tamaño del cambio en la media δ_i en cada medición donde δ_i se hace grande, cualquier método debería funcionar bien. No obstante, es la detección precisa y la identificación de las desviaciones sutiles de estado estacionario que diferencian a los métodos de baja potencia de los de alta potencia.

$$\delta_i = \sqrt{\frac{2T_a^2}{N[\tilde{S}_{k+1}^{-1}]_{i,i}}} \ldots\ldots[22]$$

$$\delta = \sqrt{\frac{2T_a^2}{N\sum_{i=1}^{P}\sum_{j=1}^{p}[\tilde{S}^{-1}{}_{k+1}]_{i,j}}} \ldots\ldots[23]$$

A partir de los resultados, se observa que la columna 2 contiene la potencia total de detección para detectar un cambio en el estado de equilibrio para cada perturbación de una única o múltiples medias de mediciones del proceso para covarianzas compensadas solamente; tener en cuenta que el poder de detección es el mismo tanto para la factorización de Cholesky como para los métodos PCA. Como se indica, el poder de detección para el estadístico multivariado es bastante razonable a pesar de que se estrecha direccionalmente en la medida que el número de mediciones con comportamiento inestable aumenta. La columna 5 es la contra-parte a la columna 2, que detalla el poder de detección utilizando las estadísticas

univariadas. Evidentemente, la estadística univariada no funciona tan bien como la estadística multivariada, lo cual no sorprende, dado que se ha introducido una estructura correlacionada de forma conjunta en el sistema. Las columnas 3 y 4 muestran el poder de identificación para la factorización de Cholesky y los métodos de análisis de contribución valor-propio/vector-propio con la segunda modificación en efecto. Claramente, las dos potencias están muy cerca una de la otra para todos los casos tratados y ninguno es un líder superior. Por lo tanto, parece que cualquier método podría ser utilizado sin pérdida aparente de potencia resultante. Sin embargo, ya que la factorización de Cholesky es más fácil de calcular, parece ser la elección más lógica para el cálculo en línea de detección de estado estacionario y la identificación. De hecho, a partir de las simulaciones en esta quinta medición de un sistema ARMA(1,1), la factorización Cholesky es en promedio 15 veces más eficiente en términos de las operaciones de punto flotante que se requieren en comparación con el método de valor-propio/vector-propio (es decir, el método PCA se tarda 15 veces más "lazos" MatLab para calcular $CONT_{k+1}$). Por lo tanto, se concluye que el método multivariado empleando el análisis de la contribución modificado puede ser utilizado como una herramienta eficaz para aislar variables potencialmente inestables con una precisión razonable. Desde la perspectiva de la facilidad de uso, es probable que sólo la medición con el mayor $CONT_{k+1}$ sea investigada más a fondo. Una vez que una explicación se ha racionalizado para esa variable y el problema posiblemente arreglado o eliminado entonces la medición correspondiente al segundo $CONT_{k+1}$ más grande sería analizada y así sucesivamente. Por lo tanto, dado el alto poder de identificación para cada simulación inducida de estado inestable, una a la vez, esta parece ser una estrategia operacional eficaz para la técnica.

En otra aplicación práctica, un aspecto del método que debe ser mencionado concierne al ejemplo cuando una o más de las variables de medición no está disponible (es decir, un instrumento fallido, por ejemplo). Si esto ocurre entonces esas mediciones no disponibles o malas pueden excluirse fácilmente de la media del período k+1th y las estimaciones de covarianza y la detección del estado estacionario pueden continuar normalmente. Sin embargo, ya que la información

del período anterior puede contener la medición justo antes de que dejó de estar disponible, es sencillo hacer caso omiso de la fila y la columna correspondiente a la mala medición en $\tilde{S}_k$ y proceder a calcular la estadística T^2 de Hotelling.

2.5 SSITS v1.0.

Vale decir que, si bien los métodos anteriores fueron cuidadosamente estudiados y enriquecidos, el grueso del contenido de este trabajo radica en la creación de la herramienta SSITS v1.0 (ver **Anexo 4**) para la integración, configuración y simulación de todos estos algoritmos. Para su creación se utilizó el software MATLAB™ R2014a (8.3.0.532) de MathWorks Inc., y más específicamente su Toolbox para desarrollo de interfaces gráficas conocido como GUIDE. Cuando se trata con métodos matemáticos complejos que necesiten configuración de muchos parámetros, numerosas simulaciones para diferentes entradas, visualización y comparación de resultados, etc.; se vuelve casi indispensable que todas estas operaciones se puedan realizar desde la comodidad, flexibilidad, re-escalabilidad, y otras potencialidades que puede ofrecer una interfaz gráfica abarcadora como la creada aquí. Es necesario precisar que el método científico que se plantea, todavía está en fase de prueba y validación, y en función de eso está la aplicación Matlab™, creada para ofrecer una serie de facilidades que permitan hacer las pruebas necesarias con señales patrón que conlleven a resultados satisfactorios.

Esta aplicación cuenta con un espacio de trabajo dividido en varias secciones que están organizadas de acuerdo a la lógica y secuencia de trabajo descrita a lo largo de este capítulo. Esta herramienta informática sigue como directriz ser lo más intuitiva y entendible posible capaz de llegar al más amplio público profesional y estudiantil. Para cumplir con esto, fue dotada de una serie de facilidades y opciones; se puede operar en su totalidad en los idiomas español e inglés, se redujeron al mínimo los parámetros iniciales que deben ser introducidos por el usuario, todos los resultados se muestran numérica y gráficamente, se pueden guardar resultados de sesiones de trabajo para luego ser cargados posteriormente y continuar donde se había dejado, existen menús contextuales en cada uno de los botones con información de su funcionamiento y se cuenta además con una extensa ayuda con

herramientas de búsqueda que facilita el entendimiento de cada uno de los procedimientos y métodos. (Ver **Figura 2.2**)

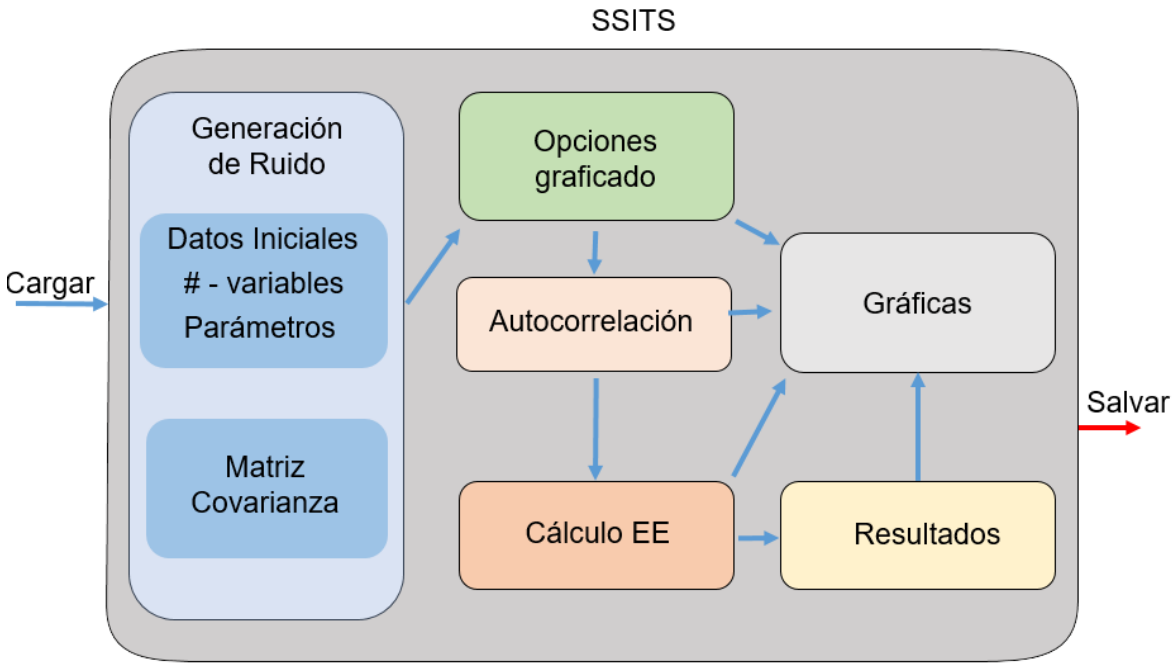


Figura 2.2: Concepción General del SSITS.

Existe una secuencia lógica de trabajo que se describe a continuación y puede ser vista en el **Anexo 5**:

Para iniciar el trabajo, el usuario debe seleccionar la cantidad de variables (n) con las que desea trabajar (esta herramienta en su primera edición cuenta con una selección de hasta cinco variables) para pasar después a una tabla donde se configuran los parámetros necesarios para la generación de la señal de ruido. También en esta primera parte se deberán introducir la Longitud de Simulación (NN) y el Número de Ventanas (MM). Como el número de muestras por ventanas (N) es un parámetro que depende de la división de estos dos anteriores se debería garantizar que dicha división resulte en un número entero.

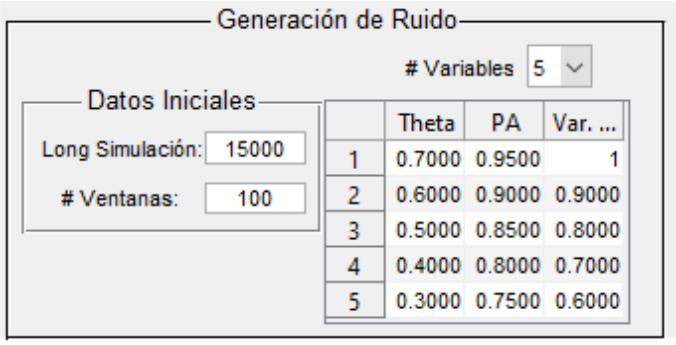


Figura 2.3: Panel de Generación del Ruido.

Si son escogidos los datos mostrados, cada ventana tendrá una longitud de 150 muestras, dividiendo así toda la señal de 15000 muestras en 100 ventanas. Un ejemplo de lo anteriormente mencionado se muestra en la **Figura 2.4**.

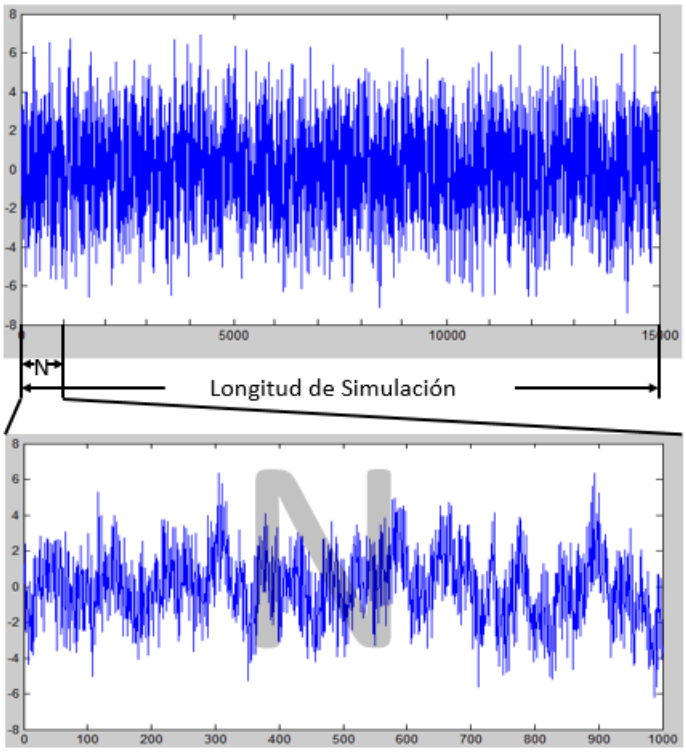


Figura 2.4: Ejemplo de ventanas de simulación.

El ultimo parámetro a ser introducido por usuario es la matriz de covarianza *H*, la cual en dependencia del número de variables escogidas cambiará el orden resultando una matriz idéntica de (*n x n*). Esta matriz será multiplicada por la señal de ruido para añadir de esta manera una covarianza que tenga el fin de simular el efecto de la correlación mutua, obteniéndose así el ruido "final" al cual se le

aplicaran todos los procedimientos. Los datos finales serán generados *click* en el botón “Generar”. Vale decir que todos estos parámetros mencionados están en función de la manera en la que está concebida actualmente la herramienta; diseñada para trabajar con una serie de señales patrones generadas con características conocidas para ver la eficacia de los algoritmos, en fase de pruebas. Una vez validado, en futuros trabajos se aplican a señales reales obtenidas de procesos termo-energéticos donde ya no se conocerían las características del ruido sin que hubiera que aplicar procedimientos de filtrado y extracción de estos rasgos distintivos.

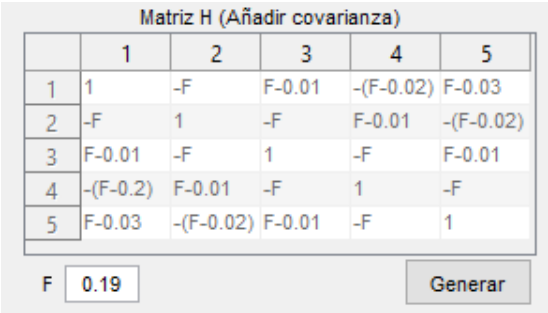


Figura 2.5: Introducción de matriz de covarianza H.

Cada una de las “*señales*” generadas en cada uno de los pasos, así como los resultados obtenidos para cada una de ellas se podrán visualizar gráficamente. Para ofrecer diferentes opciones de graficado se cuenta con un menú (**Figura 2.6**) que haciendo uso de las funciones de “*plot*” y “*subplot*” del Matlab™, se podrán elegir varias opciones en dependencia de las necesidades en cada momento de trabajo (visualización en la propia ventana o en una externa creada al efecto, ruido normal o con covarianza añadida, graficado simultáneo o independiente). Estas gráficas serán visualizadas en la parte derecha de la ventana donde también serán visibles algunos datos importantes del set de datos en cuestión.

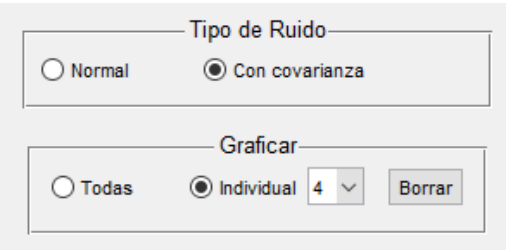


Figura 2.6: Panel de opciones de graficado.

En el panel que a continuación se describe, se encuentra la mayor carga computacional de esta herramienta, puesto que es donde se realiza como tal la DEE y evaluación de resultados. En dependencia de los datos que le sean introducidos se requerirá mayor o menor cantidad de operaciones de coma flotante lo cual influye directamente en el tiempo necesario para realizar todos los cálculos que le son solicitados (**Figura 2.7**).

Primeramente, se calcula la autocorrelación de las señales en cada período mediante la fórmula:

$$R_{(k)} = \frac{E[(X_i - \mu)(X_{i-k} - \mu)]}{\sigma^2} \quad \text{.....[24]}$$

Donde E es el valor esperado y k el desplazamiento temporal considerado (normalmente denominado desfase). Esta función varía dentro del rango [−1, 1], donde 1 indica una correlación perfecta (la señal se superpone perfectamente tras un desplazamiento temporal de k) y −1 indica una anticorrelación perfecta o lo que es lo mismo que no existe correlación alguna.

El resultado de la operación se podrá visualizar en una gráfica donde se observa la función de autocorrelación que es una representación visual bastante clara del grado de correlación que tiene cada una de estas variables.

El botón *Calcular EE* es el encargado de calcular la media de cada señal para cada ventana, compensar el efecto de la correlación serial, calcular los valores de T^2_crit y T^2, identificar la contribución de cada variable a la no-estacionalidad, calcular el número de aciertos y el poder de identificación por diferentes métodos. También se graficarán los valores de T^2 por el método de los contribuyentes individuales y el de Kourti-MacGregor; estas gráficas estarán acompañadas de sus respectivas leyendas.

Todos los resultados anteriores serán mostrados, además de forma gráfica, en tablas que pueden ser vistas en la **Figura 2.8**.

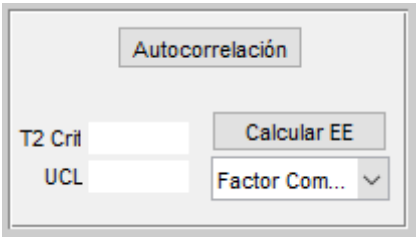


Figura 2.7: Función de Auto-correlación y detección de EE.

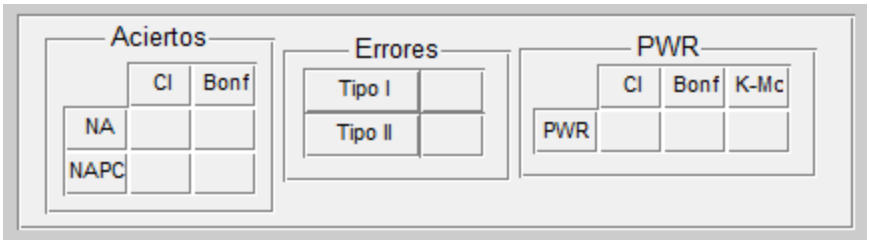


Figura 2.8: Tablas de resultados.

Observando los resultados de la **Figura 2.8** se puede establecer una comparación entre el método de CI y otros métodos; haciendo énfasis en la cantidad de aciertos predichos correctamente contra el número total de aciertos de cada método. En su segunda parte se aprecian los errores de Tipo I y II cometidos por el algoritmo implementado, como índice de desempeño del mismo y por último se establecen comparaciones acerca del comportamiento del poder de identificación de cada método. La programación de estos cálculos puede ser vista en **Anexo 6** y **Anexo 7**.

Como concepto, el error Tipo I o falso positivo, es el error que se comete cuando el investigador no acepta la hipótesis nula siendo esta verdadera. Es equivalente a encontrar un resultado falso positivo, porque el investigador llega a la conclusión de que existe una diferencia entre las hipótesis cuando en realidad no existe. Se relaciona con el nivel de significancia estadística.

El error Tipo II, también o falso negativo, se comete cuando el investigador no rechaza la hipótesis nula siendo esta falsa. Es equivalente a la probabilidad de un resultado falso negativo, ya que el investigador llega a la conclusión de que ha sido incapaz de encontrar una diferencia que existe en la realidad.

Si el algoritmo rechaza la hipótesis nula (período en EE) y por tanto indica que en determinado período la señal no está en EE, cuando en realidad si lo está, se

comete un error Tipo I. Cuando se detecta un período en EE cuando en realidad no lo está, se comete un error de Tipo II.

Todos los parámetros de configuración, los datos generados y resultados obtenidos podrán ser guardados en un archivo con extensión "*.mat". De esta manera, se puede almacenar toda una sesión de trabajo para ser continuada en futuras sesiones.

2.6 Experimentos y resultados.

La gran mayoría de los métodos de DEE estudiados basan su óptimo funcionamiento en la correcta selección por parte del investigador de algunos parámetros de entrada. Este no es la excepción, para un mejor desempeño es necesario escoger valores adecuados de longitud de simulación y número de ventanas.

Las pruebas realizadas estuvieron enfocadas en dos objetivos fundamentales:

1. para un mismo set de parámetros de entrada, ver cuál método obtenía mejores resultados y
2. encontrar cuáles son los valores apropiados de estos parámetros para nuestro método.

Los resultados de estas pruebas, que pueden ser vistos en la **Tabla 2.3**, arrojan que este algoritmo de DEE es muy eficiente cuando se estipula un número de ventanas superior a 500. Comparando los resultados con los de otros métodos de DEE, ante iguales valores de Longitud de Simulación y de Número de Ventanas, se observa que el implementado, Contribuciones Individuales (CI); es superior en casi todos los casos a los métodos tradicionales de Bonferroni y de Kourti-McGregor debido a que el número de aciertos predichos correctamente presenta en el peor de los casos un mínimo de 93 % ante un 82 % del método propuesto por Bonferroni, presentando además un mejor poder de identificación que el algoritmo de Kourti-McGregor. Resalta además que el algoritmo de CI alcanza su mejor resultado si se establece, para este caso, una Longitud de Simulación=150000 y un Número de Ventanas=500, comprobando de esta manera que dicho método se comporta de manera eficiente ante este proceso. En cuanto al poder identificación, se puede

observar como los valores del método de Contribuciones Individuales se mantiene siempre a la par de los métodos tradicionales, superándolos en varias ocasiones.

Tabla 2.3: Resultados para diferentes valores de parámetros de entrada.

		Aciertos (NA/NAPC)		Poder de Identificación		
Long Simul	# Ventanas	CI	Bonferroni	CI	Bonferroni	Kourti-McG
15000	100	6/6	20/20	1	5	1
16000	100	7/7	20/20	4	10	4
17000	100	6/6	18/18	3	6	3
18000	100	8/8	14/14	1	3	1
19000	100	5/5	14/14	0	3	0
20000	100	4/4	17/17	3	6	3
15000	100	6/6	20/20	1	5	1
15000	200	11/19	24/49	5	9	5
15000	300	16/42	27/65	4	8	4
15000	400	116/140	118/146	59	52	60
15000	500	494/494	465/465	461	415	458
15000	800	529/529	454/454	309	246	310
75000	100	6/6	18/18	0	1	0
76000	100	5/5	15/15	0	3	0
77000	100	4/4	7/7	0	0	0
78000	100	6/6	10/10	0	1	0
79000	100	3/3	10/10	0	0	0
80000	100	4/4	12/12	0	0	0
75000	100	6/6	18/18	0	1	0
75000	200	3/6	14/24	0	2	0
75000	300	3/6	16/37	0	4	0
75000	400	53/56	67/81	38	46	38
75000	500	498/498	452/452	494	444	494
75000	800	404/404	355/355	349	285	346
100000	100	4/4	11/11	0	1	0
110000	100	2/2	6/6	0	0	0
120000	100	6/6	14/14	0	0	0
130000	100	8/8	6/6	0	0	0
140000	100	8/8	13/13	0	0	0
150000	100	3/3	9/9	0	0	0
150000	100	3/3	9/9	0	0	0
150000	200	4/9	15/25	0	0	0
150000	300	1/5	9/19	0	0	0
150000	400	40/46	51/63	24	27	23
150000	500	499/499	489/489	499	486	499
150000	800	378/381	304/318	348	266	346
150000	1000	499/535	453/549	493	442	491

CI: Método de Contribuciones Individuales. Bonferroni: Método de Bonferroni.

Kourti-McG: Método de Kourti-McGregor. NA: Número de Aciertos.

NAPC: Número de Aciertos Predichos Correctamente.

Conclusiones Parciales

Se ha descrito un método relativamente sencillo que detalla un enfoque para la DEE en señales de procesos serial y mutuamente correlacionados donde se ha dado una breve panorámica de la literatura para resaltar los enfoques anteriores. El método ha sido diseñado para ser rápido en el cálculo y está destinado a ser ejecutado en línea dentro de la capa de aplicación de un Sistema de Control Distribuido. Se detalla una novedosa compensación serie de la matriz de covarianza que se basa en el cálculo directo de la función de auto-correlación para cada medición. Se demostró que esta compensación mejora en gran medida el error de Tipo I y la capacidad del método para detectar con precisión los períodos de estado no estacionario cuando se producen. También se detalló un análisis modificado de la contribución de cada variable basado en el inverso del factor de Cholesky, que permite la identificación de las mediciones que contribuyen a los períodos detectados de estado no estacionario. En consecuencia, el método puede ser usado en línea dada su eficiencia numérica cuando se aplica a sistemas grandes. También se ha logrado establecer una aplicación de software en Matlab™ con diferentes facilidades, cuya base matemática radica en la implementación del método anterior; que bajo una interfaz sencilla permitirá a especialista e investigadores detectar los períodos estacionarios de determinado proceso industrial basado en sus series temporales.

Conclusiones Generales

A manera de conclusiones de este trabajo de diploma podemos mencionar:

- Se realizó un estudio de los mecanismos de explotación de las minas de datos, el procesamiento de señales y los métodos de detección de estados estacionarios.
- Se detalló la implementación matemática de un nuevo método para la detección de EE en señales serial y mutuamente correlacionadas.
- Se diseñó e implementó el SSITS como herramienta informática para la interacción con el investigador en la ejecución de ensayos y experimentos del método.
- Se realizaron pruebas de rendimiento del método frente a otros existentes en la literatura obteniéndose resultados favorables.

Recomendaciones

Una vez desarrolladas las tareas propuestas en este trabajo de diploma, sugerimos:

- Pasar a una segunda fase donde se aplica el método de DEE aquí propuesto a registros reales provenientes de sistemas de generación de energía.
- Utilizar el informe redactado como material de apoyo por los estudiantes en temas de estadística aplicada a la Automática.
- Implementar mecanismos que le permitan al SSITS conectarse en tiempo real con procesos industriales y ejecutar acciones inmediatas capaces de detectar anomalías en el comportamiento de variables.
- Crear herramientas informáticas que, a partir de los resultados obtenidos, estimen importantes datos para la valoración del estado de la planta y su futuro funcionamiento.

Bibliografía

Albuquerque, J., & Biegler, L. (1996). Data Reconciliation and Gross-Error Detection for Dynamic Systems. AIChE Journal, 42, 10.

Arafet, P., Domínguez, H., & Chang, F. (2004). Una introducción a MATLAB. Santiago de Cuba: Facultad de Ing. Eléctrica Universidad de Oriente.

Bagajewicz, M., & Jiang, Q. (1997). Integral Approach to Plant Linear Dynamic Reconciliation. AIChE Journal, 43, 10.

Bartels, R. (1982). The Rank Version of von Neumann's Ratio Test for Randomness. Journal of the American Statistical Association, Vol. 77, No. 377, 40-46.

Bendat, J., & Piersol, A. (2000). Random data : analysis and measurements procedures. John Wiley & Sons.

Box, G., & Jenkins, G. (1976). Time Series Analysis: Forecasting and Control. Oakland: Holden-Day.

Cao, S., & Rhinehart, R. (1995). An efficient method for on-line identification of stady state. Journal of Process Control 5 (6), 363-374.

Chen, J., & Liao, C. (2002). Dynamic process fault monitoring based on neural network and PCA. Journal of Process Control, 12(2), 277-289.

Chen, J., Bandoni, A., & Romagnoli, J. (1997). Robust Estimation of Measurements Error Variance/Covariance from Process Sampling Data. Computers chem. Engng. 21, 6.

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

Dapozo, G., Porcel, E., López, M., & Bogado, V. (2007). Técnicas de preprocesamiento para mejorar la calidad de los datos en un estudio de caracterización de ingresantes universitarios. Corrientes.Argentina.

Fayyad, U., Piatesky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery: An overview. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1-34.

Forbes, J., & Marlin, T. (1996). Design Cost: A Systematic Approach to Technology Selection for Model-Based Real-Time Optimization Systems. Computers chem. Engng.

Grupo de Quimiometría y Cualimetría de Tarragona; Dpto de Química Analítica y Química Orgánica Universitat Rovira i Virgili (Tarragona). (2002 Junio, 24). Quimiometría, una disciplina útil para el análisis químico. Técnicas de Laboratorio (272), 412-416.

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

Harris, T., & Ross, W. (1991). Statistical Process Control Procedures for Correlated Observations. Can. J. Chem. Eng.

Hernández, J., Ramírez, M. J., & Ferri, C. (2004). Introducción a la Minería de Datos. Cataluña: Pearson.

Holly, W., Cook, R., & Crowe, C. (August, 1989). Reconciliation of Mass Flow Rate Measurements in a Chemical Extraction Plant. Can. J. Chem. Eng.

Jeffrey D., K., & John D., H. (2012). A Steady-State Detection (SSD) Algorithm to Detect Non-Stationary Drifts in Processes. Journal of Process Control.

Kao, C.-S., Tamhane, A., & Mah, R. (1990). Gross Error Detection in Serially Correlated Process Data. Ind. Eng. Chem. Res.

Kao, C.-S., Tamhane, A., & Mah, R. (1992). Gross Error Detection in Serially Correlated Process Data. 2. Dynamic Systems. Ind. Eng. Chem. Res.

King Saud University. (2002). Process Control in the Chemical Industries.

Kourti, T., & MacGregor, J. (1996). Multivariate SPC Methods for Process and Product Monitoring. J. Qual. Tech.

Ku, W., Storer, R., & Georgakis, C. (1995). Disturbance Detection and Isolation by Dynamic Principal Components Analysis. Chemometrics Intelligent Lab Syst., 30.

Lon-Mu, L., Siddhartha, B., Sclove, S., & Rong, C. (2001). Data mining on time series:an illustration using fast-food restaurant franchise data.

Lowry, C., & Woodall, W. (1992). A Multivariate Exponentially Weighted Moving Average Control Chart. Technometrics, 34, 1.

MacGregor, J. (2004). Data-Based Latent Variable Methods for Process Analysis, Monitoring and Control. Lisbon, Portugal: Elsevier.

Madansky, J. (1988). Prescriptions for working statisticians. New York: Springer - Verlag.

Mason, R., Champ, C., Tracy, N., Wierda, S., & Young, J. (1997a). Assessment of Mulitvariate Process Control Techniques. J. Qual. Tech., 29, 2.

Mason, R., Tracy, N., & Young, J. (1997b). A Practical Approach for Interpreting Multivariate T2 Control Chart Signals. J. Qual. Tech., 29, 4.

Narasimhan, S., Kao, C., & Mah, R. (1987). Detecting Changes of Steady States using the Mathematical Theory of Evidence. AIChE Journal 33, 11,.

Narasimhan, S., Mah, R., Tamhane, A., Woodward, J., & Hale, J. (1986). A Composite Statistical Test for Detecting Changes of Steady States. AIChE Journal.

Negiz, A., & Cinar, A. (1997). Statistical Monitoring of Multivariable Dynamic Processes with State-Space Models. AIChE Journal, 43, 8.

Nomikos, P., & MacGregor, J. (1995). Multivariate SPC Charts for Monitoring Batch Processes. Technometrics, 37, 1.

Prabhu, S., & Runger, G. (1997). Designing a Multivariable EWMA Control Chart. J. Qual. Tech., 29, 1.

Savitzky, A., & Golay, M. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem., Vol 36, 1627-1630.

Seisdedos, L. V., Blanco, J. M., Peña, F., & Rodríguez, J. M. (2014). Monitorización y diagnóstico de centrales térmicas: desarrollo de un detector visual de estados estacionarios. Memoria Investigaciones en Ingeniería, No 12, 17-29.

Stanley, G., & Mah, R. (1977). Estimation of Flows and Temperatures in Process Networks. AIChE Journal, 23, 5.

Stoumbos, Z., & Reynolds, M. (1997). Control Charts applying a Sequential Test at Fixed Sampling Intervals. J. Qual. Tech., 29, 1.

Tong, H. (1995). Studies in Data Reconciliation using Principal Component Analysis. Ph.D. Dissertation. McMaster University, Hamilton, Canada.

Tong, H., & Crowe, C. (1997). Detecting Persistent Gross Errors by Sequential Analysis of Principal Components. AIChE Journal 43, 5.

Vasilopoulos, A., & Stamboulis, A. (1978). Modification of Control Chart Limits in the Presence of Data Correlation. J. Qual. Tech.

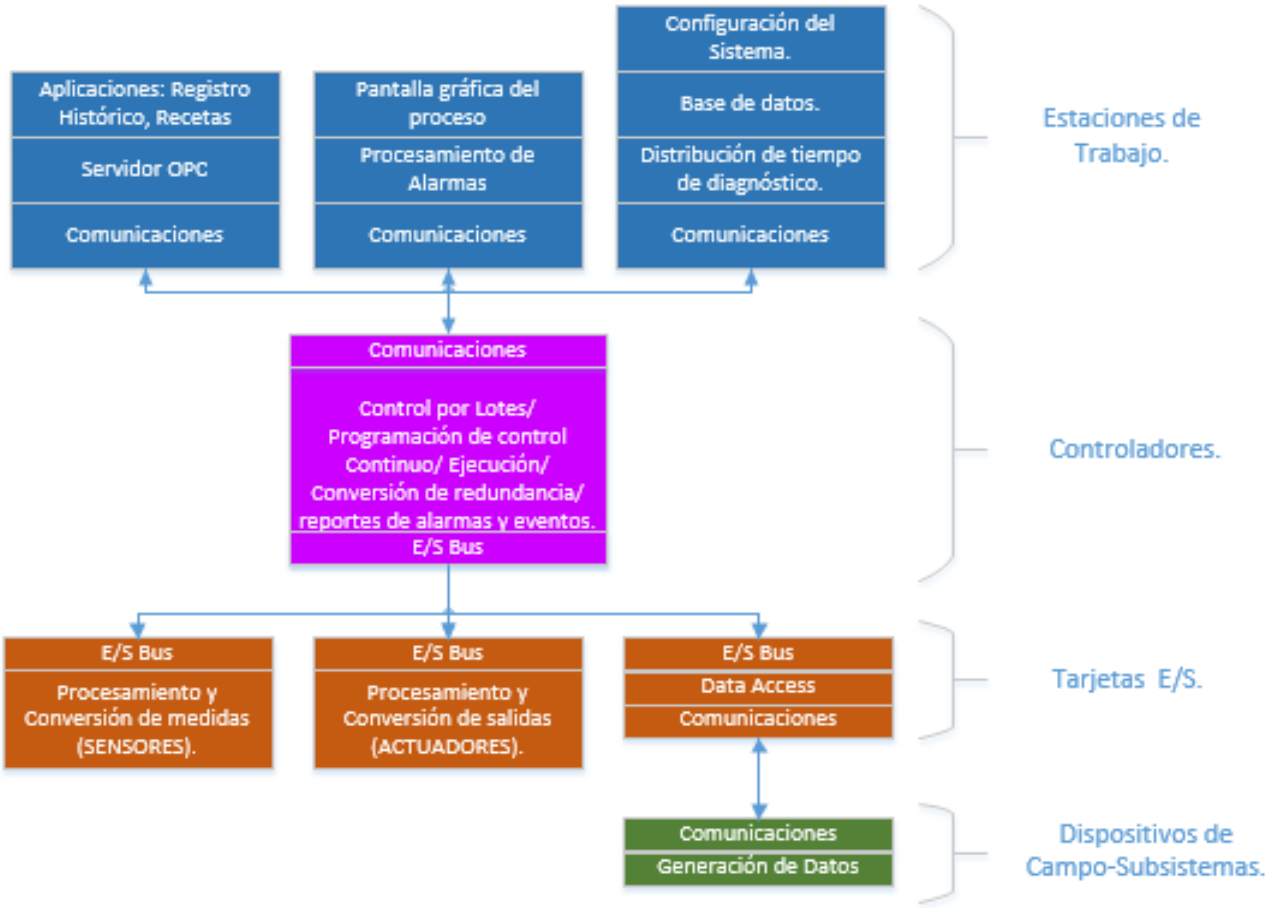
Wang, X. (2001). En Knowledge Discovery through Mining Process Operational Data. In Application of Neural Networks and other Learning Technologies, in Process Engineering (págs. 287-327). London: Imperial College Press.

Wardell, D., Moskowitz, H., & Plante, R. (1994). Run Length Distributions of Residual Control Charts for Autocorrelated Processes. J. Qual. Tech.

Zhang, N., & Pollard, J. (1994). Analysis of Auto-correlations in Dynamic Processes. Technometrics, 36, 4.

Anexos

Anexo 1: Elementos de un DCS.



Anexo 2: Parámetros del modelo para el proceso ARMA (1,1).

El modelo de proceso estocástico simulado contiene 5 medidas que se implementan mediante la siguiente expresión:

$$[n_t]_i = \emptyset_{i,1}[n_{t-1}]_i + [H]_i \underline{\varepsilon}_t - \theta_{i,1}[\underline{\varepsilon}_{t-1}]_i \quad \text{[A1]}$$

donde $[n_t]_i$ es el ruido estocástico superpuesto para la *i*th medición y los valores de los parámetros utilizados, incluyendo las varianzas del ruido blanco ($\sigma^2_{\varepsilon,i}$), se muestran en la **Tabla 2.1**.

Con el fin de introducir un poco de covarianza en el sistema, la matriz simétrica H se utiliza, como se muestra en la **Ecuación A2** para simular el efecto de la correlación cruzada.

Tabla 2.1. Ejemplo simulado de modelado de datos estocásticos.

i	$\theta_{i,1}$	$\phi_{i,1}$	$\sigma^2{}_{\varepsilon,i}$
1	0.70	0.95	1.00
2	0.60	0.90	0.90
3	0.50	0.85	0.80
4	0.40	0.80	0.70
5	0.30	0.75	0.60

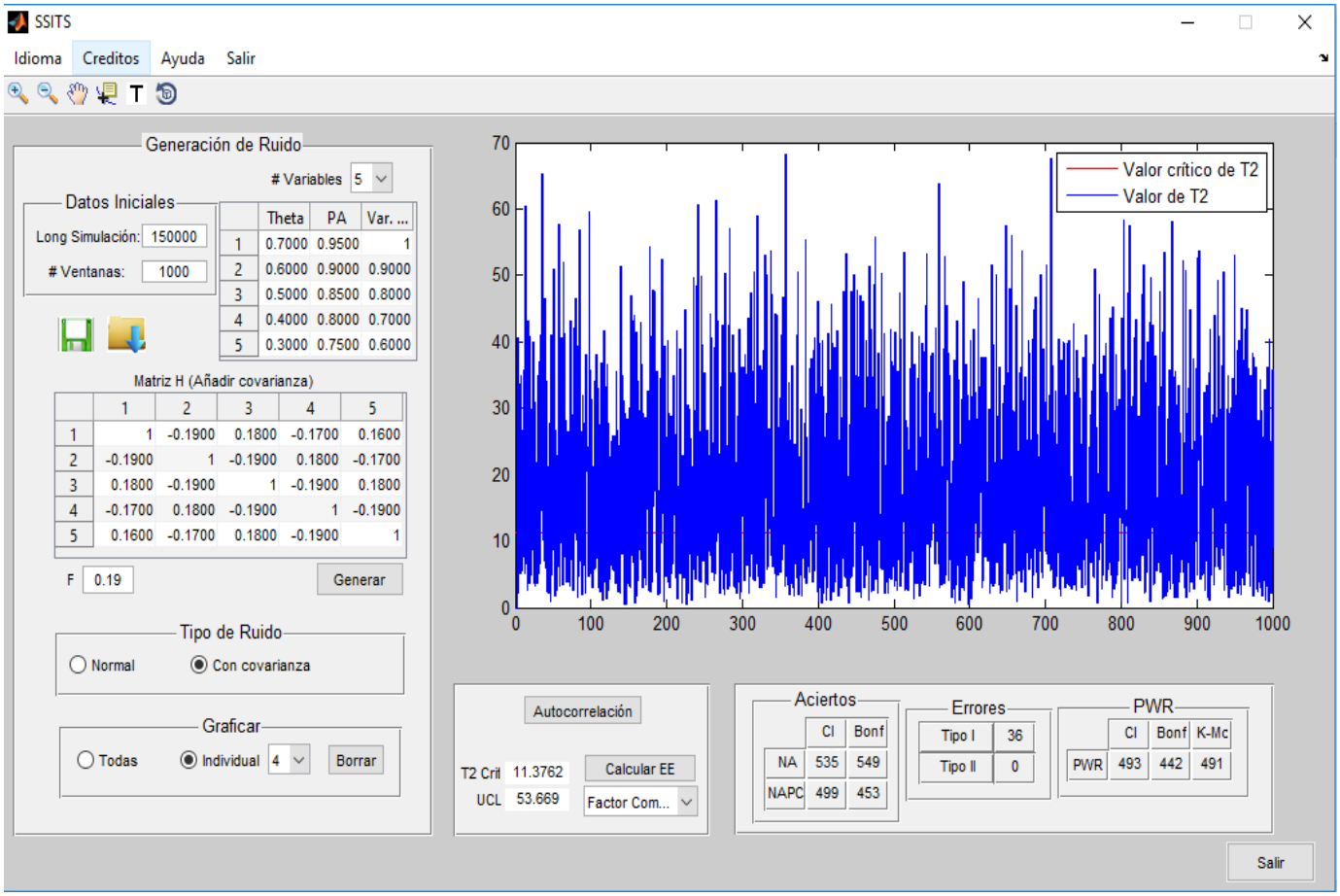
El uso de H simula covarianza a través de la adición de ruidos blancos en el tiempo t de las otras mediciones (es decir, $[H]_j * \left[\underline{\varepsilon}_t\right]_j \forall\, j = 1 \ldots p, j \neq i$).

$$\mathbf{H} = \begin{bmatrix} 1 & -0.19 & 0.18 & -0.17 & 0.16 \\ -0.19 & 1 & -0.19 & 0.18 & -0.17 \\ 0.18 & -0.19 & 1 & -0.19 & 0.18 \\ -0.17 & 0.18 & -0.19 & 1 & -0.19 \\ 0.16 & -0.17 & 0.18 & -0.19 & 1 \end{bmatrix} \quad \text{[A2]}$$

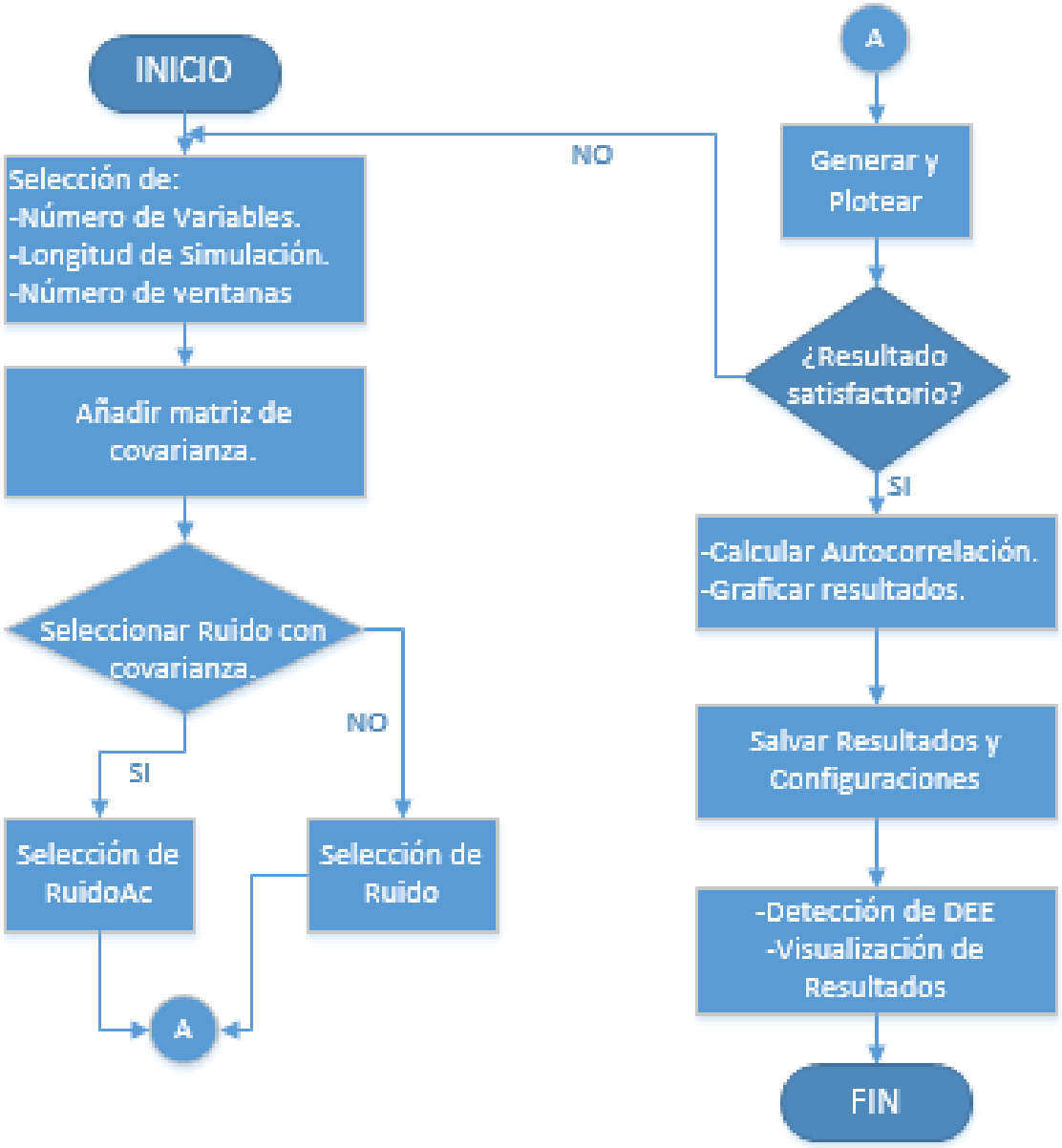
Anexo 3: Código para compensar el efecto de la correlación serial y calcular los valores de T^2.

```
   W(:,m) = zeros(Nvars,1);
    for i = 1:Nvars
      for k = 2:nl+1
        W(i,m) = W(i,m) + (N-k-1)*acRuido(k,i);
      end
    end
    W(:,m)=ones(Nvars,1)+2/N*W(:,m); %término de compensación
    A2=cov(ruidoC(N*(m-1)+1:N*m,:)); %matriz de covarianza
    x2_var = diag(A2);     %diagonal de la matriz de covarianza
    x1_var = diag(A1);

    %-----------------------------------------------------
    % Compensacion del efecto de la correlación serial.
    S = diag(abs(W(:,m)).^(0.5))*S*diag(abs(W(:,m)).^(0.5));

% ESTADISTICA HOTELLING. Compensado el efecto de la correlación
serial.
      T2(m)=0.5*N*d(:,m)'/S*d(:,m);
%T2(m)=0.5*N*d(:,m)'*Si*d(:,m);



%--------Calcular T2 critico-----------------------------------
---------------------------------
if m == 2
flag = 3;
[teta,xx,ff] = txf(2.0*alpha,fa-Nvars+1,Nvars,flag);%calcular
la estadística T, F y Chi-Square de Student
F = ff;
end
T2_crit(1) = 1/((fa - Nvars + 1)/(Nvars * fa)) * F;
```

Anexo 4: Interfaz Gráfica Creada (SSITS).



Anexo 5: Diagrama de flujo general del SSITS.



Anexo 6: Cálculo del número de aciertos y errores de Tipo I y II.

```
for m = 2:MM
    if T2(m) > T2_crit(1)
      if abs(DSS(m)~=0)
        n_hits_corr = n_hits_corr + 1;
      end
      n_hits = n_hits + 1;
    end

    if abs(DSS(m)==0)     %Error de Tipo I.
      if T2(m)> T2_crit(1)
         errorT1=errorT1+1;
      end
    end

    if abs(DSS(m)~=0)  %Error de Tipo II.
      if T2(m)< T2_crit(1)
         errorT2=errorT2+1;
      end
   end

    if T2_u(m) > T2_crit(1)  %limite Bonferroni
      if abs(DSS(m)~=0)
        n_hits_corr_u = n_hits_corr_u + 1;
      end
      n_hits_u = n_hits_u + 1;
    end
  end
```

Anexo 7: Cálculo del poder de identificación de cada método.

```
for m = 2:MM
  if T2(m) > T2_crit(1) %Usando CI
    if abs(DSS(m)~=0)
      [h,hh] = sort(-abs(CONT(:,m)));
      for i = 1:Nvars
        for j = 1:n_ge
          if SF(i) > 0.0 & hh(j) == i
            Dummy = Dummy + 1;
          end
        end
      end
      if Dummy == n_ge
        n_hits_pwr_e = n_hits_pwr_e + 1;
      end
    end
  end
    if T2(m) > T2_crit(1)      %Usando Kourti and MacGregor
    if abs(DSS(m)~=0)
      [h,hh] = sort(-abs(CCONT(:,m)));
      for i = 1:Nvars
        for j = 1:n_ge
          if SF(i) > 0.0 & hh(j) == i
            Dummy = Dummy + 1;
          end
        end
      end
      if Dummy == n_ge
        n_hits_pwr = n_hits_pwr + 1;
      end
    end
  end
    if T2_u(m) > T2_crit(1)         %Usando Bonferroni
    if abs(DSS(m)~=0)
      [h,hh] = sort(-abs(z_u(:,m)));
      for i = 1:Nvars
        for j = 1:n_ge
          if SF(i) > 0.0 & hh(j) == i
            Dummy = Dummy + 1;
          end
        end
      end
      if Dummy == n_ge
        n_hits_pwr_u = n_hits_pwr_u + 1;
      end
    end
  end
```