ITEGAM
ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)
INSTITUTO DE TECNOLOGIA ITEGAM

Industrial Multivariate Explorer: A solution with Matlab for time series mining

Luis Vázquez Seisdedos[1], Alvaro Aguilera Castillo[1], David Diaz Martinez[1], Rolando Ramis Rosales[2], Jose Manuel Rodriguez Perez[3]

[1]Departamento de Control Automático, Facultad de Ingenieria Electrica, Universidad de Oriente, Santiago de Cuba, Cuba ({lvazquez, aaguilera,ddiaz}@fie.uo.edu.cu)
[2]Refineria "Hermanos Diaz", Santiago de Cuba, Cuba (rramiz@refscu.cupet.cu)
[3]Central Termoeléctrica Habana, Habana, Cuba (josemrod@cthabana.uni.cu)

ABSTRACT

The design and features of a Matlab's application to support applied researches for serial time computing is presented. The input data can be from historical record coming from chemical and thermal processes and also it can be generated by simulation. Up to 8 signals, linearly normalized and distributed can be visualized on an axis Matlab's object. By means of two cursors, the user can choose short windows of recorded signals. On this serial time sections, in this version, statisticians are computed and they facilitate the static modeling. They can be saved into an Excel file. It is an opened software application permitting to include new features. The Windows between 2 cursors command facilities the dynamic modeling. Its applicability is exemplified by times series from industry (from a 250 MW thermal power plant) and generated by simulation.

Keywords: process monitoring, data-driven modeling, quasi-stationary states, chemometric methods, steam generators.

Multi-variable Industrial: Una solución con Matlab para minar series temporales

RESUMEN

Se presenta el diseño y las prestaciones de una aplicación desarrollada en Matlab orientada a dar soporte de cómputo a la exploración y extracción de información a partir de series temporales. La data de entrada podrá ser de un registro histórico de un proceso químico y termo energético y aquellos generados de experimentación simulada. La visualización es de hasta 8 señales linealmente normalizadas y distribuidas a lo alto del objeto axis. El usuario podrá seleccionar ventanas cortas de registro mediante dos cursores. Sobre estas secciones de series temporales, en esta versión, se computan estadígrafos que facilitan el modelado estático. Estos podrán ser salvados en un fichero Excel. Es una aplicación abierta permitiendo la inclusión de nuevas prestaciones. El comando Windows entre dos cursores facilita el modelado dinámico. Su aplicabilidad se ejemplifica con series temporales de la industria (de una central térmica de 250 MW) y generadas por simulación.

Palabras Claves: Monitoreo de procesos, modelado basado en datos, estados cuasi-estacionarios, métodos químico-métricos, generadores de vapor.

1 INTRODUCCIÓN

La evolución de los sistemas de automatización hacia aquellos denominados de Control Distribuido (del inglés; DCS) y el software instalado en los computadores de supervisión [1], así como la facilidad de establecer compactos sistemas registradores de datos (del inglés; *Data logger*) conducen a la generación de grandes volúmenes de datos.

La supervisión de procesos se encarga de observar continuamente las variables del proceso en busca de la detección de anormalidades que puedan representar un problema operativo o de calidad. A esta se le destinan como sub-tareas la detección y diagnóstico de fallos y el análisis de procesos. Dependiendo del horizonte de tiempo con el que se trabaja, la supervisión se puede aplicar a 2 niveles:

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

DOI: https://dx.doi.org/10.5935/2447-0228.20150019

A largo plazo: En este nivel se analiza el comportamiento del proceso a largo plazo, a través de los datos históricos. La meta es identificar causas de bajo rendimiento y oportunidades de mejora. Los términos Análisis del proceso o Mejora del Proceso se utilizan para designar este tipo de supervisión [2] y [3].

2. REVISIÓN BIBLIOGRAFICA

En los procesos industriales controlados, las series temporales asociadas a cada variable medida constituyen las minas de datos históricos a partir de las cuales se podrá caracterizar el rendimiento, detectar los cambios de parámetros y servir de base a diagnosticar sus causas, etc.

Para caracterizar, evaluar, diagnosticar, pronosticar, etc los procesos químicos, por ejemplo para la producción de energía, se requiere disponer de herramientas de cómputo para aplicar las técnicas de la Químico-métrica (del inglés, *Chemometrics*). Bajo esta denominación se considera la ciencia de las mediciones [4] realizadas sobre sistema químico para caracterizar su estado vía la aplicación de métodos matemáticos o estadísticos. Por ende es una ciencia basada en datos. La meta de la mayoría de las técnicas de la Químico-métrica es derivar un modelo empírico, a partir de los datos, le permita al investigador estimar una o más propiedades del sistema a partir desde las mediciones. Los sistemas químicos incluyen la dinámica del proceso

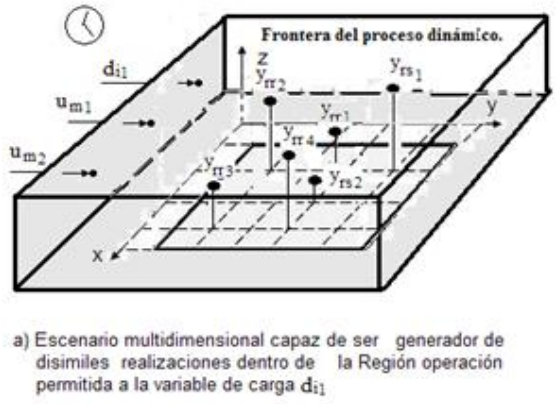
La obtención de conocimientos a partir de bases de datos es un campo de investigación y aplicación interdisciplinario, que se ha hecho relevante en los últimos 12-15 años. Intenta proponer soluciones al problema de cómo extraer información de grandes cantidades de datos. Diversos autores [2], [5], [6], [7], [8] y [9] han adoptado como básica la definición de KDD (*Knowledge Discovery in Databases*) propuesta por Shapiro: "KDD es el proceso no trivial de identificar a través de los datos patrones noveles, potencialmente útiles y entendibles". Se descompone en varias etapas, según como sigue:

- Definición del objetivo del análisis.
- Selección de datos: Se hace de acuerdo con los objetivos propuestos, muchas veces se asocia a aspectos informáticos relacionados a cómo acceder y almacenar los datos.
- Preprocesamiento de los datos: asegura la calidad de los datos en el sentido de que elimina ruidos aleatorios, *outliers* (datos atípicos o errores gruesos), manejo de datos ausentes o perdidos.
- Transformación de los datos: se refiere a cómo encontrar algún tipo de característica que ayude a mejorar la eficiencia y facilidad de identificación de patrones. Lo típico en esta etapa son los métodos de proyección y reducción de dimensionalidad de los datos.
- Minería de datos (MD): es el paso central del proceso KDD. La meta en esta etapa es identificar patrones bien definidos, válidos, novedosos, potencialmente útiles y significativos, de acuerdo con el objetivo del análisis.
- Interpretación y validación: se enfoca hacia la evaluación e interpretación de los resultados del paso MD.

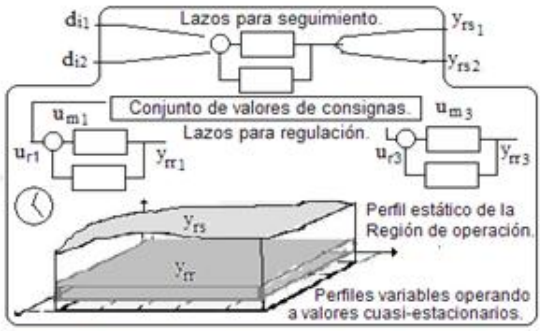
De este modo, para investigaciones aplicadas basadas en datos que resulten, ya sean de experimentaciones de laboratorios o sean de realizaciones industriales, se requiere disponer de una aplicación de software en una plataforma tal que soporte los cálculos computacionales de ingeniería necesitados como lo es el Matlab y que a su vez facilite; (i) la visualización multicanal con escalamiento, (ii) seleccionar tramas con cursores y crear una base de datos secundaria con resultados calculados en las ventanas cortas. El objetivo de este artículo es diseñar e implementar un explorador Multi-variado Industrial (IME) usando la plataforma MATLAB en su versión 7.7.

3. METODOS

El monitoreo de procesos es esencial para mantener una elevada calidad de la producción al igual que su seguridad. Sea, como se ilustra en figura 1, un proceso dinámico controlado, P_D, multi-variable, en general de parámetros distribuidos, con existencia de fenómenos de transporte.



a) Escenario multidimensional capaz de ser generador de disimiles realizaciones dentro de la Región operación permitida a la variable de carga d_{i1}



b) Proceso químico de parámetros distribuidos con soluciones de control por regulacion ante consignas fijas y seguimiento de la variable de carga en sus demandas.

Figura 1 – Fronteras para la supervisión y monitorización de variables reguladas y de seguimiento. Entradas: d_i: Señales de carga o perturbación. U_r: Señales de referencia o consigna. Salidas: y_{rs}: Señales de salida de variables en seguimiento. y_{rr}: Señales de salida de variables reguladas u_m: Señales de control.

La figura 1a delimita sus fronteras por el cubo rectangular y el carácter multidimensional de la señales asociadas a cada variable medida. La Figura 1b muestra una representación estática, la cual es sólo con fines ilustrativos de un caso hipotético bi-variado. Sobre una superficie aproximadamente plana paralela a los x-y se ubican a las variables reguladas para distinguir su valor cuasi-constante e independiente a dos posibles entradas de carga se ubican y otra superficie en la que se ubica aquellas variables de seguimiento de la carga o demanda

P_D podrá ser cualquier proceso de tiempo continuo, dinámico, al cual se le regulan variables dentro de determinadas condiciones de tolerancia. A este subconjunto se le denota mediante y_{rr} y al mismo tiempo, existen otras variables que responden a un problema de seguimiento de la carga o demanda y se le denota mediante y_{rs}. Las señales de carga o de demanda se le denota mediante di y las acciones correctivas de control se le denota mediante u_m.

El conjunto de variables sometidas a regulación, y_{rr}, acotan sus señales dentro de subconjuntos de valores de rango estrecho y aquellas variables sometidas a seguimiento de la carga, y_{rs}, sus señales están cuantificadas dentro de un subconjunto de valores que rango amplio.

Sea, como se indica en la figura 2, una variable de carga denotada por d_{i1} que puede tomar valores en la región comprendida entra el límite inferior y un valor nominal.

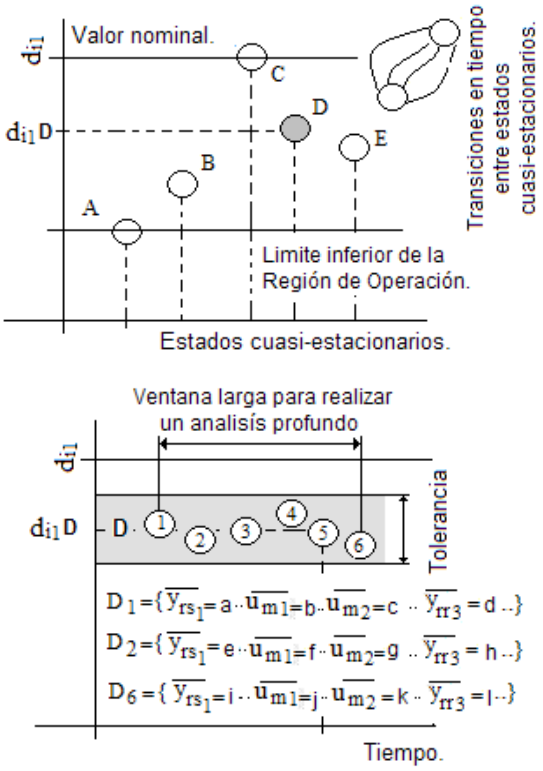


Figura 2 – Estados cuasi-estacionarios: comportamiento multivariable y multidimensional en función de una variable de carga y su variabilidad temporal en días: (a) d_{i1} vs conjunto de estados cuasi-estacionarios {A, B, C, D, E} seleccionados en ventanas cortas; (b) evolución del estado cuasi-estacionario "D" a lo largo de la ventana temporal larga.

En un proceso químico de producción continua se podrá transitar por diferentes estados cuasi-estacionarios, por ejemplo; {A, B, C, D, E}, Figura 2a, cada uno de los cuales definido por el conjunto de valores centrados en un valor medio constante dentro de una tolerancia establecida.

Estos estados fueron estimados en tramas de series temporales a lo ancho de una ventana corta de tiempo, por ejemplo, durante las ventanas cortas; {1, 2, 3, 4, 5, 6}, Figura 2b.

4. RESULTADOS Y DISCUSIÓN

El IME es una aplicación abierta sobre Matlab que ha surgido con fines docentes y para minar datos de procesos termo-energéticos. En su núcleo presenta dos funcionalidades básicas; la visualización y la disponibilidad de selección tramas en ventanas cortas a partir de ventanas largas de series temporales. Es sobre los vectores existentes en la ventana corta que queda abierta la posibilidad de extender las aplicaciones de valor añadido que se requieran ir incorporando.

Con vistas a facilitar el diseño de algoritmos se ha incluido en el IME la posibilidad de enlazar un modelo Simulink (aplicaciones de matlab con extensión de fichero ".mdl") diseñado por el usuario y que le permita una fase de ensayo y validación previa.

4.1 Prestaciones actuales del IME.

El diseño del IME, como se ilustra en la figura 3, posee actualmente las siguientes prestaciones:

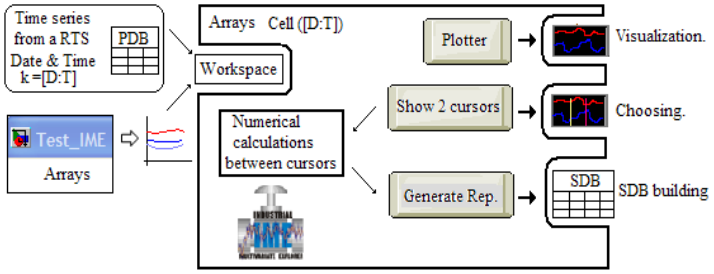


Figura 3 – Prestaciones del Explorador Multi-variado Industrial (IME)

1. Visualizar sobre el mismo osciloscopio (objeto Axis de la programación grafica de Matlab) hasta 8 vectores. Este tipo de dato ha sido previamente cargado sobre su espacio de trabajo (Ventana *Workspace*). Sí el juego de datos dispone de un arreglo de celdas con las fechas y tiempo real absoluto, entonces la manipulación de estas series temporales están acuñadas con su tiempo real.
2. Seleccionar a un subconjunto de elementos de los vectores largos delimitados entre dos cursores (según una ventana corta) que posiciona el usuario sobre el graficado.
3. Construir una Base de Datos Secundaria (SDB) a partir de una Base de Datos Primaria (PDB). En esta versión del IME, la SDB resulta de calcularle los estadígrafos a cada

trama seleccionada entre cursores. De este modo, la meta es estimar el conjunto de valores asociados a cada variable que caracteriza cada estado cuasi-estacionario del sistema dinámico bajo estudio. Los siguientes cinco comandos de la estadística descriptiva son aplicadas: "**mean**", "**std**", "**range**", "**min**" y "**max**".

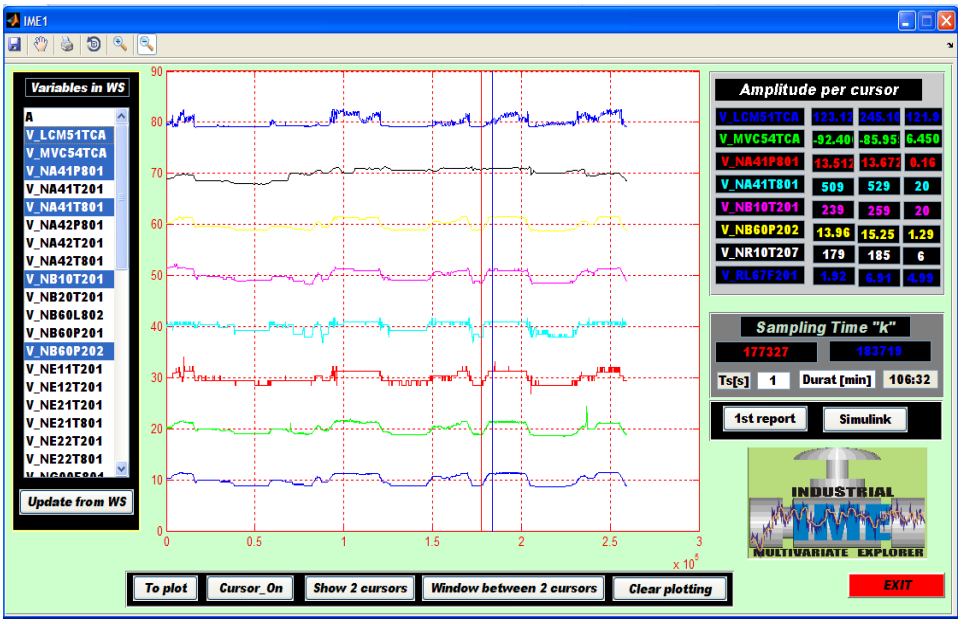
4.2. Interfaz de usuario del IME.

Una vez que se ejecuta el comando "IME", se podrá pasar a la ventana "IME1". El diseño ergonómico de su interfaz de usuario permite armonizar todo lo que éste necesita para trabajar en sus dos primeras prestaciones y al mismo tiempo se permite abrir el fichero correspondiente al modelo Simulink (que ha sido previamente diseñado y construido).

4.2.1 Visualización y selección.

El IME podrá tomar las dos siguientes fuentes de datos: (i) un conjunto de vectores unidimensionales (que representan Series Temporales Reales, RTS) con un vector de celdas que contiene la fecha y tiempo (D, T) de cada muestra y (ii) un conjunto de vectores unidimensionales que resultan de la exportación de resultados durante un tiempo de corrida de un modelo Simulink (ejemplificado en la figura 3 mediante "Test_IME.mdl").

Las figuras 4a y 4b muestran las variables industriales obtenidas a partir de señales reales y sintéticas, respectivamente. Estas últimas fueron generadas usando el modelo Simulink de la figura 5, el cual es abierto y su tiempo de simulación fue configurado mediante la interfaz mostrada en la figura 6.



(a)

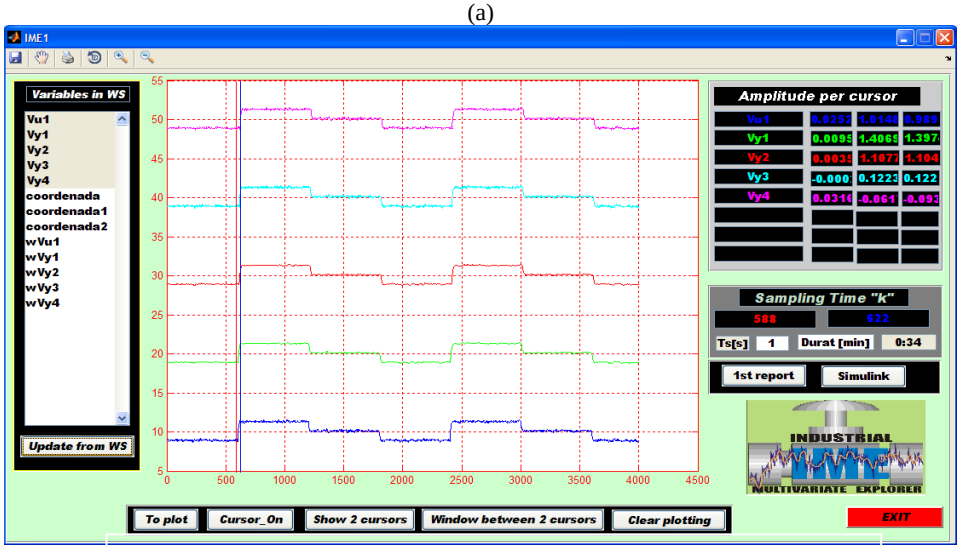


Figura 4. Prestaciones del Explorador Multivariado Industrial (IME)

(b)

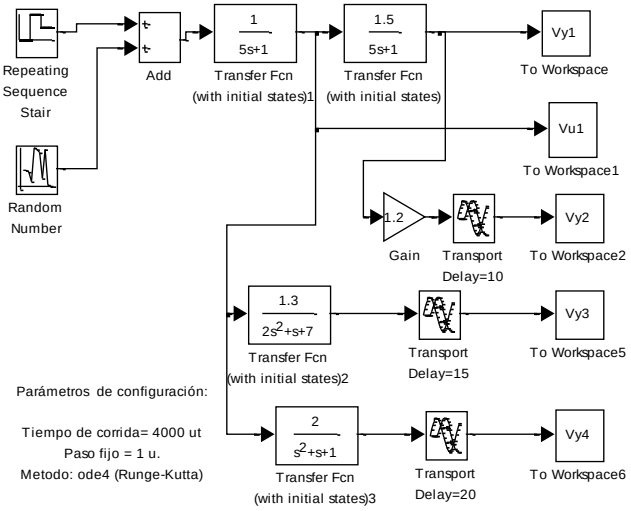


Figura 5. Modelo Simulink para generar señales sintéticas.

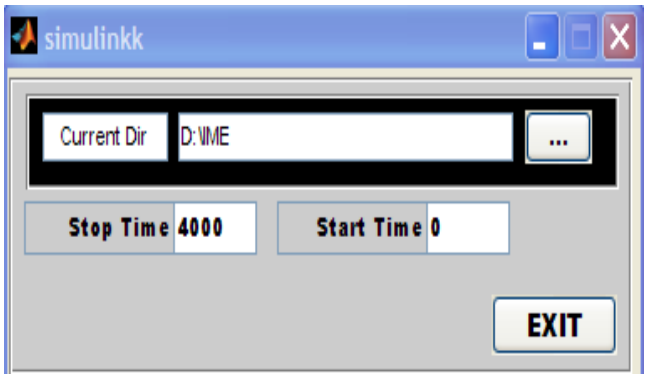


Figura 6. Interfaz de conectividad con el modelo Simulink.

De las señales exportadas del registro histórico y actualmente existente en el espacio de trabajo del Matlab se pueden visualizar a la vez hasta 8 de ellas. En la figura 4 se han seleccionado las siguientes: LCM51TCA, MW, Potencia activa en demanda, MVC54TCA, kPa, Presión de vacío en el condensador, NA41P801, MPa, Presión del vapor a la salida del sobrecalentador # 4 (lado izquierdo), NA41T801 ºC Temperatura del vapor salida del sobrecalentador # 4 (lado izquierdo), NB10T201, ºC, Temperatura a la salida del Economizador (lado izquierdo),NB60P202, MPa, Presión del domo, NR10T207, ºC, Temperatura gases a la salida de los Calentadores de Aire Regenerativo, CAR, (lado izquierdo) y RL67F201, Ton/h, es el Flujo de agua de atemperamiento # 2 (lado izquierdo). Al momento de realizar el graficado la primera en registrarse, V_LCM51TCA es la que se observa de color azul más abajo. Los valores de amplitud de cada señal se corresponden con el intercepto del cursor de color rojo y aparecen en la segunda columna del cuadro "**Amplitude per Cursor**". De modo análogo aparecen en la tercera columna los correspondientes al cursor en azul y en la cuarta columna se muestra la diferencia entre cada uno de ellos respectivamente. Asimismo se muestra el tiempo de muestreo de cada uno de ellos y se brinda una diferencia en tiempo (de formato en minutos: segundos) ya que se toma como base de cálculo que la unidad de muestreo es el segundo. Para la visualización están los comandos "**plot**" que grafica las señales de las variables seleccionadas y "**clear plotting**" que las borra del objeto axis. Para seleccionar se dispone de "**cursor on**" para situar el primer cursor y segundo cursor en secuencia y "**show 2 cursors**" para fijar las coordenadas de cada uno de ellos y determinar los valores de amplitud y tiempo de cada una de las variables existentes en el espacio de trabajo. De ellas sólo se muestran las que fueron seleccionadas.

El comando "**Window between 2 cursors**" crea nuevos vectores a las variables seleccionadas y le antepone la etiqueta "w". Los botones "**1st report**" y "**Simulink**" permiten pasar a la ventana de cálculo de estadígrafos y dar paso a la apertura de la aplicación Simulink respectivamente.

4.2.2 Cálculo entre cursores y creación de base de dato secundaria.

El comando "**1st report**" conduce al usuario a la interfaz de: (i) cálculo (sobre los vectores dentro de la ventana corta) y (ii) la funcionalidad de crear de base de dato secundaria con cada estadígrafo por hojas de Excel. Ya que respecto al tiempo, las series pueden venir con y sin arreglo de celdas (que contienen la fecha y tiempo real), es por lo que se disponen de dos plantillas para salvar los nuevos datos calculados. Las figuras 7 y 8 muestran la operatividad sobre la data industrial y sobre la data generada por simulación respectivamente.

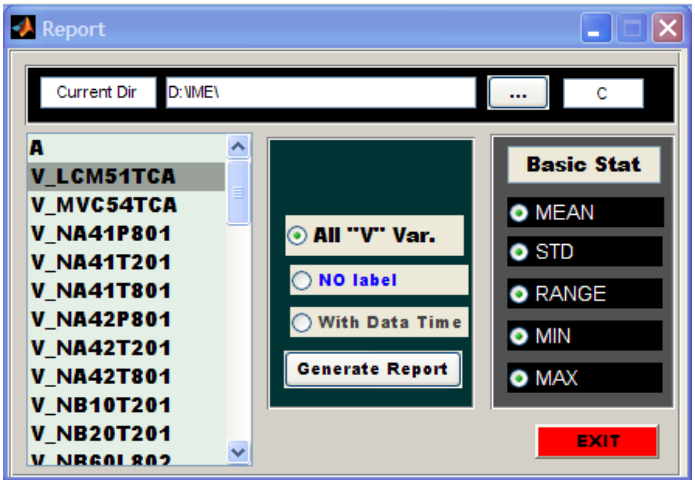


Figura 7. Interfaz para enlace de fichero Excel y selección de estadígrafos.

Ed. 002 VOL 001 – ISSN 2447-0228 (online)

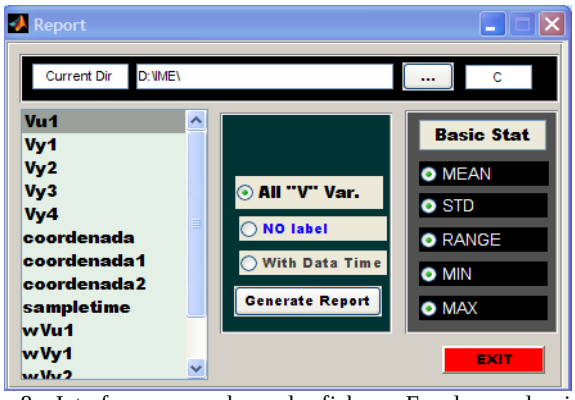


Figura 8. Interfaz para enlace de fichero Excel y selección de estadígrafos.

De modo análogo, las figuras 9 y 10 muestran las plantillas de salva para la industrial en la que se dispone de la fecha y el tiempo real absoluto y aquella para tiempo real relativo.

	A	B	C
1		Industrial Multivariate Explorer	
2			
3		Mean Value Report	
4			
5		Initial Date	06 Feb 2010 03:35:21:000
6		End Date	06 Feb 2010 05:10:13:000
7		Duration[min]	94.86666667
8		Time Coordenate of cursor 1	48922
9		Time Coordenate of cursor 2	54614
10		V_LCM51TCA	126.9989364
11		V_MVC54TCA	-92.02215394
12		V_NA41P801	13.45660566
13		V_NA41T201	465
14		V_NA41T801	517

MEAN STD RANGE MIN MA

Figura 9. Plantilla para salva de cálculo de estadígrafos. Formato de fecha y tiempo.

	A	B	C	D
1		Industrial Multivariate Explorer		
2				
3		Maximum Value Report		
4				
5				
6				
7		Duration[min]	1.566666667	
8		Time Coordenate of cursor 1	746	
9		Time Coordenate of cursor 2	840	
10		Vu1	1.080294326	
11		Vy1	1.536751069	
12		Vy2	1.844101282	
13		Vy3	0.202029414	
14		Vy4	2.048327852	

MEAN STD RANGE MIN MA

Figura 10. Plantilla para salva de cálculo de estadígrafos. Formato de fecha y tiempo.

4.3 Perspectiva del uso del IME.

El IME es una aplicación abierta que, en primer término facilita la visualización multicanal. De la figura 3 se infiere que existe un elemento clave como lo es la delimitación de la ventana corta, bajo alguna condicional preestablecida. Esta ponencia muestra como solución la pericia del usuario experto en localizar que las variaciones de las señales que representan a las variables sean mínimas de modo tal que el proceso no exhiba tendencia y por ende se encuentre en un estado cuasi-estacionario. Las perspectivas del IME están enmarcadas en sus facilidades actuales y en los objetivos que se planteen dentro de los conceptos de ventanas corta y larga de procesos industriales sometidos a cambios paramétricos dentro de las variadas cajas de herramientas del paquete Matlab.

Las señales correspondientes a las variables supervisadas de los procesos químicos portan ruidos y perturbaciones que están fuertemente auto-correlacionadas y son, en la mayoría de las plantas, de naturaleza no lineal. Las técnicas de monitoreo de procesos que han sido ampliamente empleadas tales como: el Análisis de Componentes Principales (PCA) y las Mínimos Cuadrados Parciales (PLS) [10], [11] caen en modelos estáticos, las cuales asumen que las observaciones son independientes del tiempo y siguen una distribución Gaussiana. Las extensiones al PCA y al PLS, también denominados DPCA y DPLS han estado desarrolladas a direccionar este problema. El Análisis Canónico Variado (CVA), [12], como herramienta de monitoreo basado en el espacio de estado es más apropiado que estos últimos referidos métodos de monitoreo dinámico.

Sin embargo, para aquellos procesos termoenergéticos que se suceden en las centrales termoeléctricas, y que, por ejemplo, como el generador de vapor opera bajo un régimen de transferencia de calor de modo distribuido, y que regularmente opera en régimen cuasi-estacionario, la supervisión a largo plazo, dígase en ventanas de tiempo largas, con asistencia del IME permitirá investigar sobre la distribución de estos estados con el transcurso de los días. Mediante la determinación de modelos regresivos se podrá examinar los cambios parametricos e implementar métodos para la inferencia de sus causas.

5. CONCLUSIONES.

Mediante esta aplicación se facilita notablemente la exploración y extracción de información de procesos químicos y termo-energéticos a partir de las señales asociadas a sus variables. Una vez visualizado un registro largo, la detección de tramas, correspondientes a estados de operación cuasi-estacionarios, en ventanas de tiempo cortas permitirá caracterizarlos. A su vez, este fichero Excel constituye una base de dato secundaria, portadora de largos periodos de trabajo y sometidos a diferentes regímenes que propiciaran el diagnostico de la operación y la aplicación de técnicas de planificación de mantenimientos preventivos.

Ed. 002 VOL 001 – ISSN 2447-0228 (online)

6. REFERENCIAS BIBLIOGRÁFICAS

1. Creus A. (1999). **Instrumentación Industrial.** Editora Alfaomega-S.A, Colombia.

2. Wang, X. Z. (2001). **Knowledge Discovery through Mining Process Operational Data.**, In Application of Neural Networks and other Learning Technologies, in Process Engineering (Mujtaba, I. M. y M. A. Hussain eds.). Imperial College Press, London. pp. 287-327.

3. Macgregor, J. F. (2004) **Data-Based Latent Variable Methods for Process Analysis, Monitoring and Control**., in European Symposium on Computer Aided Process Engineering - 14 (Barbosa Póvoa, A. y Matos, H. eds.). Elsevier, Lisbon, Portugal.

4. Barry M. Wise and Neal B. Gallagher. **The process chemometrics approach monitoring and fault detection.** J Proc. Cont Vol. 6, No. 6, pp. 329 348, 1996

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

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

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

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

9. U.M Fayyad., G. Piatesky-Shapiro, P. Smyth. **From Data Mining to Knowledge Discovery:** An overview". Advances in Knowledge Discovery and Data Mining, p.p: 1-34, AAAI/MIT Press 1996.

10. Manabu Kano, Shinji Hasebe, Iori Hashimoto, Hiromu Ohno. **A new multivariate statistical process monitoring method using principal component analysis.** Computers & Chemical Engineering, Volume 25, Issues 7-8, 15 August 2001, Pages 1103-1113, ISSN 0098-1354,

11. John F. MacGregor, **Data-based latent variable methods for process analysis, monitoring and control, In:** A. Barbosa-Povoa and H. Matos, Editor(s), Computer Aided Chemical Engineering, Elsevier, 2004, Volume 18, European Symposium on Computer-Aided Process Engineering-14, 37th European Symposium of the Working Party on Computer-Aided Process Engineering, Pages 87-98, ISSN 1570-7946, ISBN 9780444516947,

12. Odiowei, P.-E.P.; Yi Cao; "**Nonlinear Dynamic Process Monitoring Using Canonical Variate Analysis and Kernel Density Estimations**,"Industrial Informatics, IEEE Transactions on , vol.6, no.1, pp.36-45, Feb. 2010.