Rellenado óptimo de huecos en series de datos mediante modelo CEGH.

REPORTE TÉCNICO.

MSc. Ing. Ruben Chaer y Ing. Lorena Di Chiara.
Departamento de Potencia – IIE-FING-UDELAR.
Administración del Mercado Eléctrico – ADME.
Febrero 2016 - Montevideo, Uruguay.

***Resumen*—Este trabajo presenta un método para completar series de datos en forma óptima utilizando un modelo CEGH previamente calibrado para representar las relaciones de las series entre si y con sus pasados.**

***Palabras para indexado*—Rellenado de datos. Modelo estocástico CEGH.**

I. Introducción.

Dado un conjunto de señales (variables en el tiempo), los modelos CEGH están compuestos por un un conjunto de funciones no-lineales (Deformadores) que intentan transformar el proceso aleatorio de las señales en un proceso gaussiano.

En este trabajo se trabajará siempre sobre las señales en el espacio gaussiano; en dicho espacio, las correlaciones y autocorrelaciones que representan el proceso son modelada por un sistema lineal como se muestra en la ec. 1 donde X_k es el vector de N señales gaussianas salidas del proceso aleatorio, N_r es la cantidad de retardos considerados en el sistema lineal, R_k es un vector de ruidos blancos gaussianos independientes que atacan el filtro lineal que representa el proceso. Las matrices A_h y B determinan el sistema lineal (o filtro lineal).

$$X_{k+1} = \sum_{h=0}^{h=N_r} A_h X_k + B R_{k-h} \qquad \text{ec.(1)}$$

En la construcción del modelo CEGH se utilizan crónicas de las señales (también llamadas series históricas o realización histórica del proceso) para determinar Los Deformadores y las matrices del Filtro Lineal. En la plataforma SimSEE [2] se usa en forma extensiva los modelos CEGH para el modelado de los aportes hidráulicos, a las represas, velocidad de viento, radiación solar, precio del petróleo, etc. En el conjunto de utilidades que conforman dicha plataforma se suministra el programa AnalisisSerial que permite la creación de modelos CEGH a partir de series de datos.

El uso de los modelos CEGH antes referido es para la generación de series sintéticas para simulación por método de Monte Carlo. En este trabajo se desarrolla un uso diferente del modelo CEGH que es el de completar huecos en las medidas.

II. Relleno de medidas.

Supongamos que se dispone de un modelo CEGH calibrado con series históricas y que se quiere utilizar el mismo para completar los posibles huecos en las series que se continúan recibiendo y procesando. A modo de ejemplo (ejemplo que motivó este desarrollo) supongamos que se dispone de un CEGH que modela las correlaciones entre las velocidades de viento en diferentes parques eólicos distribuidos en el territorio nacional. En forma continua se reciben las señales de todos los parques y las mismas son usadas para calcular una estimación de la generación. Esporádicamente, se producen errores en las estaciones meteorológicas o en la recolección de datos por lo cual puede faltar la información (en una ventana de tiempo) de alguna de las posiciones de medida. Esto impediría la estimación de la generación de ese parque en particular, salvo que sea posible inferir una estimación información faltante a partir del resto de las medidas y del modelo CEGH que está calibrado sobre el conjunto.

III. El problema de "La entrada más probable".

Razonando sobre la ec.1, supongamos que se conocen los valores de las señales X_k y X_{k+1} el problema planteado es calcular el vector R_k más probable para esos valores del filtro.

En forma genérica, el problema es calcular la realización más probable de un vector R de ruidos blancos gaussianos independientes (distribución $N(0,1)$) que explique con máxima verosimilitud la ec2

$$A R = B \qquad \text{ec.(2)}$$

La función de densidad de probabilidad de un conjunto de variables aleatorias gaussianas es la que se muestra en la ec 3.

$$p_x(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-E(x))^T \Sigma^{-1} (x-E(x))} \qquad \text{ec.(3)}$$

La solución más verosimil del problema de la ec.1 se puede formular como un problema de optimización maximizar la probabilidad de la solución dado que la misma debe cumplir con la restricción lineal ec.2,

Considerando la ec.3 es sencillo mostrar que maximizar $p_R(R)$ es equivalente a minimizar $R^T R$.

$$\min_R R^T R \quad @ | A R = B \qquad \text{ec.(4)}$$

Para la solución del problema de la ec.4 nos plantemos la función Lagrangiana de la ec

$$L(R,\lambda) = R^T R + \lambda^T (AR - B) \qquad \text{ec.(5)}$$

El gradiente de la $L(R,\lambda)$ debe anularse en la solución del óptimo de la ec.4. lo que lleva a las expresión $\nabla_R L = 2R + A^T \lambda = 0$ y a la expresión $\nabla_\lambda L = AR - B = 0$ que refleja que se deben cumplir el conjunto de restricciones del problema. Despejando R de la primera de las expresiones se obtiene la ec.6 y sustituyendo R de la ec.6 en la segunda expresión se obtiene la ec.

$$R = \frac{-1}{2} A^T \lambda \qquad \text{ec.(6)}$$

$$\frac{-1}{2} A A^T \lambda = B \qquad \text{ec.(7)}$$

De la ec.7 se puede obtener: $\lambda = (A A^T)^{-1} 2 B$ que sustituyendo en la ec.6 se llega a la solución buscada ec.8.

$$R = A^T (A A^T)^{-1} B \qquad \text{ec.(8)}$$

Chequeo: En el caso en que A sea una matriz cuadrada invertible, la solución debiera ser: $R = A^{-1} B$ con lo cual se debiera cumplir que: $A^{-1} = A^T (A A^T)^{-1}$ para satisfacer la ec.8. Multiplicando por la izquierda, ambos lados de la igualdad por A se tiene: $I = A A^{-1} = (A A^T)(A A^T)^{-1} = I$ lo que confirma que en el caso de que el sistema de restricciones es determinado la solución de la ec.8 es la correcta.

Existencia: Con respecto a la existencia de una solución del problema, cabe acotar que si el rango de A es igual que la dimensión del espacio de B entonces las columnas de A expanden el espacio en que se encuentra B y la ec.8 puede usarse para determinar la solución de máxima verosimilitud de los vectores R que combinan las columnas de A en un vector igual a B .

Si el rango de A es inferior a la dimensión del espacio de B corresponde chequear si el vector B se encuentra o no en el sub-espacio expandido por las columnas de A para lo cual hay que determinar una base de dicho espacio y calcular la proyección de B en el espacio ortogonal a dicho sub-espacio. Si dicha proyección es nula, quiere decir que se pueden quitar restriccions (hay restricciones redundantes). Si se aplica el método de escalerización tradicional al sistema (A, B) , obtiene el conjunto no redundante seguido de las filas NULAS al final.

Núcleo de A : En el caso en que el vector de ruidos R pertenece a un espacio de dimensión superior al espacio del vector B , la transformación representada por la matriz A tendrá un núcleo de dimensión mayor que cero. Supóngase un valor de R cualquiera y descompongamos dicho valor en un vector P perteneciente al núcleo y otro Q perteneciente sub-espacio ortogonal al núcleo.

Se tiene entonces que $AR = AP + AQ = AQ$ por construcción. El vector AQ es entonces la imagen del vector R en el espacio del vector B . Ahora supongamos que aplicamos la transformación de la ec.8 al vector AQ con lo cual se tiene: $Y = A^T (A A^T)^{-1} AQ$, siendo Y la imagen del vector AQ en el espacio de R dada por la transformación $A^T (A A^T)^{-1}$. Bien, si ahora aplicamos la transformación A al vector Y se obtiene $AY = (A A^T)(A A^T)^{-1} AQ = AQ$ lo que muestra que la transformación $A^T (A A^T)^{-1}$ es la inversa de la transformación A y viceversa entre vectores del sub-espacio ortogonal al núcleo de A (en el espacio de R) y el espacio de B . Dado un vector R y una transformación A la proyección de R sobre el sub-espacio ortogonal al núcleo de A se puede calcular como: $Q = A^T (A A^T)^{-1} R$ y la proyección de R en el núcleo de A como $P = R - Q = (I - A^T (A A^T)^{-1}) R$

IV. Completando series de datos.

Supóngase que se dispone de un CEGH que representa el modelo de correlaciones de un conjunto de series temporales y que se quiere utilizar dicho CEGH para completar huecos en las medidas de las series.

En principio, es posible pensar directamente en series gaussianas con distribución N(0,1) y en el CEGH como un sistema lineal con el de la ec.1.

Sea X_k el vector de n valores que toman las series para el instante k . La ec.1 nos permite estimar el valor siguiente de las series. Ahora supongamos que en el siguiente valor (instante $k+1$) hay un hueco en la series de datos y que de las componentes de X_{k+1} hay un subconjunto en los que no se tiene el valor (por ejemplo por rotura del instrumento de adquisición de una medida) y así en los sucesivos pasos de tiempo hasta que se llega a un instante

$k+m$ en el que se logra nuevamente tener información completa del vector X_{k+m}. Esto nos permite plantear un sistema de restricciones lineales y calcular la series de ruidos $R_k \dots R_{k+m-1}$ más verosímil para la evolución observada (en las variables en las que hay información) desde X_k hasta X_{k+m}.

Por simplicidad razonaremos sobre un CEGH de 1 retardo.

Para el primer paso de tiempo, se tiene: $X_{k+1} = A X_k + B R_k$ donde X_k está completo y en X_{k+1} falta información en algunas de las componetes. Supongamos que por ej. la componente j no tiene información.

Observar que para las componentes j de X_{k+1} en las que hay información, se establecen restricciones lineales de la forma de le ec.9.

$$\sum_{h=1}^{n} b_{jh} r_j^k = x_j^k - \sum_{h=1}^{n} a_{jh} x_h^k \qquad \text{ec.(9)}$$

Para las componentes j de X_{k+1} en las que no hay información, se puede expresar el valor de la información faltante como se estable en la ec.10.

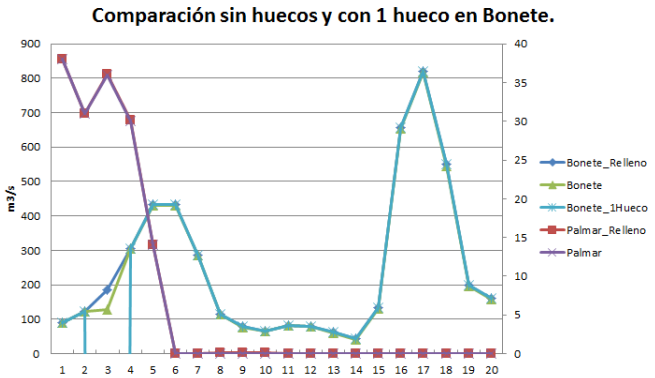
$$x_j^k = \sum_{h=1}^{n} a_{jh} x_h^k + \sum_{h=1}^{n} b_{jh} r_j^k \qquad \text{ec.(10)}$$

Observar que la ec.10 permite calcular la información faltante como una combinación lineal de la información pasada tanto en las variables x como en las entradas de ruido r. Esto permite en cada paso de tiempo, establecer nuevas restricciones (ecuaciones como la ec.9) y expresar la información faltante como combinación del pasado. El procedimiento continúa paso a paso agregando restricciones al conjunto de restricciones hasta llega a un paso en el que la información esté completa. Llegado a ese punto, se tendrá un problema del tipo resuelto en la sección III.

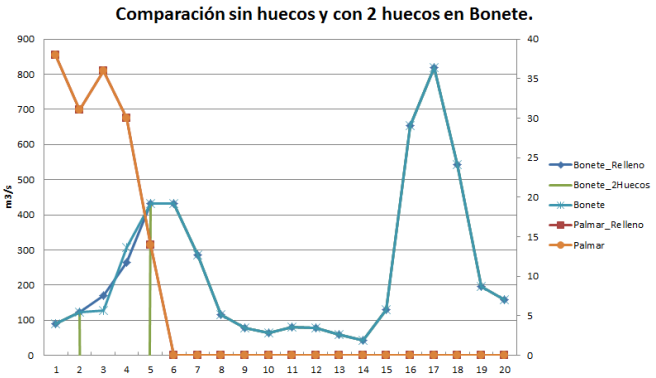
La Fig.1 sirve para explica el algoritmo para armar el problema. En el paso k la información X_k está completa (encabezado de las primeras n columnas marcado en celeste en la Fig.1) y se han colocado en las primeras n filas las matrices A y B que establecen como calcular X_{k+1} en función del pasado. El resto de los valores de las pimeras n filas son nulos. Así se prosigue, paso a paso hasta llega al final (últimas n filas) en que se escriben la evolución que lleva a X_{k+m} que se supone que el pimer vector con información completa posterior a X_k. Luego de este primer barido que completa la matriz de la Fig.1 se realiza otro barrido desde la columna correspondiente a la última componente de X_{k+m-1}^T hasta la columna correspondiente a la primer componente de X_k^T buscando componentes marcadas como huecos. Cuando se encuentra un hueco, se elimina la columna sustituyendo la variable por la su expresión dada por la fila asociada a la variable. En le Fig.1 a modo de ejemplo, se marcó una columna (línea punteada roja vertical) y la fila asociada ala variable. Al realizar esta sustitución, como cada variable solo depende del pasado, al llegar a la primer columna se habrán eliminado todas las referencias a los huecos y se ha logrado un sistema de ecuaciones como el planteado en la sección III. Durante el proceso de sustitución de variables, las ecuaciones usadas para la sustitución (las del tipo ec.10) deben ser almacenadas en una lista de ecuaciones para que una vez resuelto el problema de entrada más probable se puedan utilizar para calcula efectivamente los huecos y completar así la serie.

V. Ejemplos de aplicación.

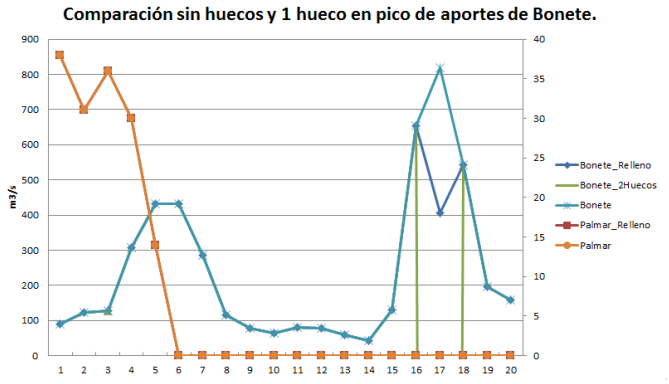
A. Caso señal bi-variada un hueco.



B. Caso señal bi-variada dos huecos.



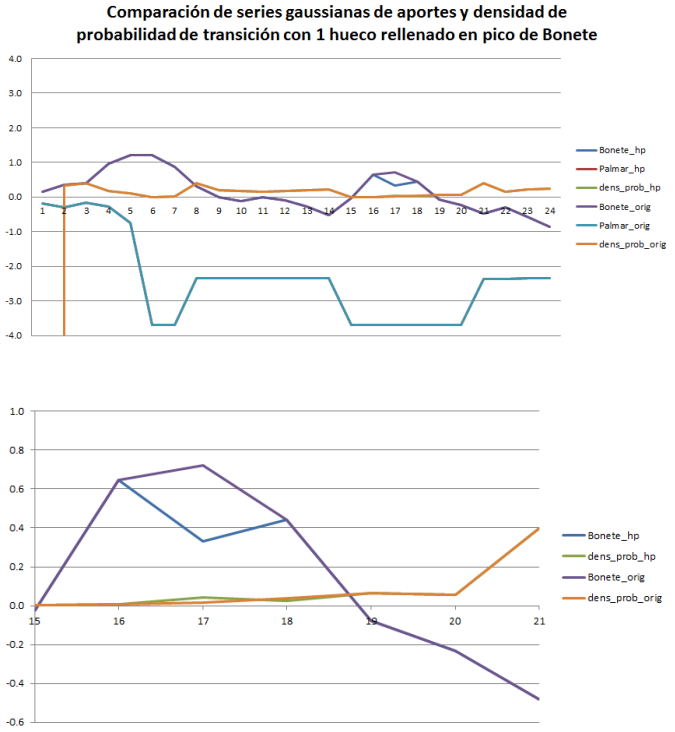
C. Caso señal bi-variada hueco en un pico.



VI. Verosimilitud de una trayectoria.

Al analizar los casos del apartado anterior, surge que los resultados de los casos A y B coinciden con el relleno que manualmente podría hacer un humano mirando las figuras ??? y ??? respectivamente; pero no así en caso C en que el rellenado realizado por el algoritmo no parece mantener la suavidad que los humanos preferimos. Obviamente estas observaciones son vagas y es necesario tener un criterio para evaluar la bondad de las soluciones. Por tal motivo, se agregó a la herramienta la posibilidad de generar la probabilidad de la trayectoria del estado del sistema para poder tener así la Verosimilitud de cada trayectoria definiendo la misma como la densidad de probabilidad asignada a su ocurrencia.

Ejemplo:



	Bonete_hp	Palmar_hp	dens_prob_h	Bonete_orig	Palmar_orig	dens_prob_orig
15	-2.53E-02	-3.69E+00	4.55E-04	-2.53E-02	-3.69E+00	4.55E-04
16	6.46E-01	-3.69E+00	5.29E-03	6.46E-01	-3.69E+00	5.29E-03
17	3.33E-01	-3.69E+00	4.27E-02	7.20E-01	-3.69E+00	1.78E-02
18	4.43E-01	-3.69E+00	2.26E-02	4.43E-01	-3.69E+00	3.79E-02
19	-7.93E-02	-3.69E+00	6.31E-02	-7.93E-02	-3.69E+00	6.31E-02
20	-2.34E-01	-3.69E+00	5.64E-02	-2.34E-01	-3.69E+00	5.64E-02
21	-4.79E-01	-2.36E+00	3.97E-01	-4.79E-01	-2.36E+00	3.97E-01

Producto de la densidad de probabilidad de las transiciones donde se encuentra el hueco:

serie sin huecos: 6,75e-4

serie completada: 9,66e-4

Como era de esperar el resultado del producto de la densidad de probabilidad de la serie rellenada es mayor que el de la serie original ya que los datos con que fue rellenada la serie son los de máxima verosimilitud.

VII. Referencias

[1] R.Chaer. Fundamentos de modelo CEGH de procesos estocásticos multivariables. SimSEE. Reporte Técnico , IIE-FING-UDELAR–2011.
http://iie.fing.edu.uy/publicaciones/2011/Cha11/Chaer%20SimSEE.pdf

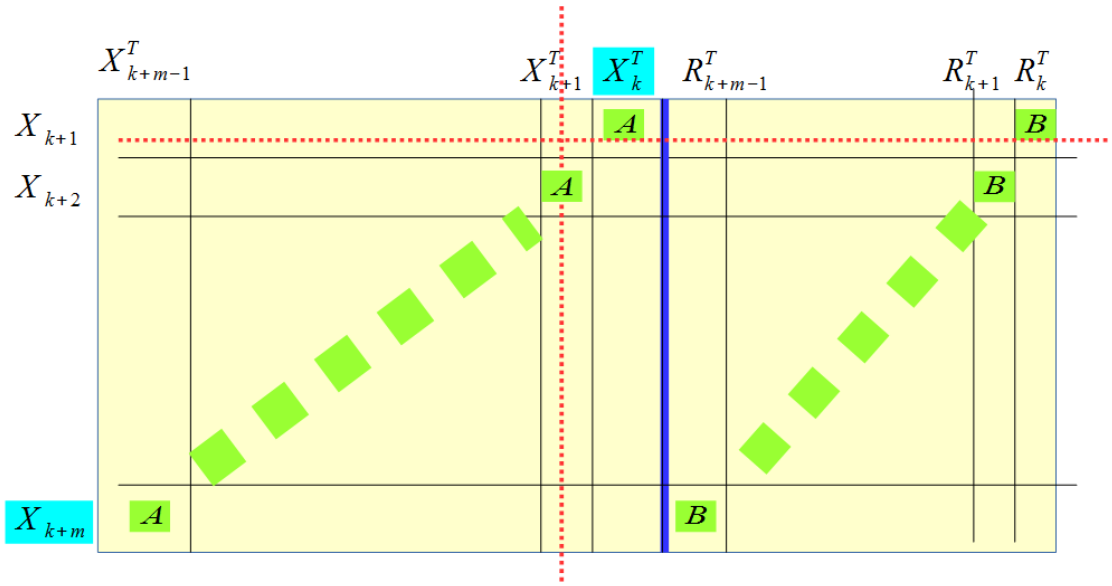


Fig. 1: Armado del problema.