Revista de Matemática: Teoría y Aplicaciones 2009 **16**(1) : 60–75
cimpa – ucr issn: 1409-2433

DESCRIPCIÓN DE DOS MÉTODOS DE RELLENADO DE DATOS AUSENTES EN SERIES DE TIEMPO METEOROLÓGICAS

Eric J. Alfaro* F. Javier Soley†

Recibido/Received: 20 Feb 2008 — Versión revisada/Revised version: 6 Nov 2008 — Aceptado/Accepted: 8 Dic 2008

Resumen

Se presentan dos metodologías para el rellenado de datos ausentes, enfocadas hacia su uso en series de tiempo geofísicas. La primera se basa en la descomposición en componentes principales de la matriz de correlación de datos de una misma variable entre estaciones cercanas y en periodos de tiempo comunes. Este método multivariable permite incorporar en los valores rellenados los fenómenos de mayor escala a partir de la información de las estaciones cercanas. El segundo método es para ser utilizado cuando no hay estaciones cercanas y el rellenado se debe hacer con la información de la misma estación. Consiste en ajustar un modelo autoregresivo a la serie de tiempo y utilizar ese modelo como estimador de los datos ausentes. Se evaluaron dos algoritmos para calcular los coeficientes autoregresivos: el estimador de Burg y el propuesto por Ulrych y Clayton. El primero es apropiado para procesos estocásticos y el segundo para series determinísticas. Las dos metodologías descritos en este trabajo son recursivas: se hace una primera estimación a los datos ausentes corriendo el algoritmo ignorando los datos ausentes si es posible ó aproximándolos de forma grosera. El algoritmo se continúa ejecutando con los nuevos valores sustituyendo los valores estimados en la corrida anterior. La ejecución termina cuando la diferencia máxima de los valores estimados entre dos corridas sucesivas es menor que un valor fijado de antemano por el usuario. Los datos rellenados conservan la media y la desviación estándar original de la serie de tiempo. Estos algoritmos se adaptaron y modificaron, por medio del uso de Interfaces Gráficas con el usuario, para su uso en SCILAB, que es una plataforma

*Centro de Investigaciones Geofísicas, Escuela de Física y Centro de Investigaciones en Ciencias del Mar y Limnología, Universidad de Costa Rica. 2060-Ciudad Universitaria Rodrigo Facio, San José, Costa Rica. E-Mail: erick.alfaro@ucr.ac.cr

†Centro de Investigaciones Geofísicas, Universidad de Costa Rica. 2060-Ciudad Universitaria Rodrigo Facio, San José, Costa Rica. E-Mail: fjsoley@racsa.co.cr

similar a MATLAB pero de fuente abierta y que corre indistintamente en Windows y Linux. Ellos fueron elaborados como una actividad de extensión de la Universidad de Costa Rica.

Palabras clave: datos faltantes, control de calidad, filtros auto regresivos, análisis de componentes principales, aplicaciones de software libre.

Abstract

Two methods for filling missing data gaps in geophysical time series are presented. The first one is based on the principal component decomposition of the correlation matrix built for close spatial stations with common time series records of the same variable. This multivariate method allows the incorporation in the estimated values of large scale features based on the information shared by the stations. The second method could be used when there are no close station and the missing data must be calculated from the same station historical information. This method adjusts an auto-regressive model to the time series which is then used to estimate the missing data. Two algorithms were used to calculate the auto-regressive coefficients: the Burg estimator and the one proposed by Ulrych and Clayton. The first one is appropriate for stochastic processes and the second for deterministic series. The two methodologies described in this work are recursive: a first estimation of the missing data is done running the algorithms but ignoring or using a crude approximation of them. Then, the algorithm runs again with the new estimated data, replacing the previous run missing data estimations. The run stops when the maximum difference between two successive estimations is smaller than the value fixed by the user. Filled data conserves the mean and standard deviation of the original time series. These algorithms have been adapted and modified for its use in SCILAB using also Graphic User Interfaces. Scilab is an open source platform, similar to MATLAB, and runs indistinctively in Windows and Linux. They were elaborated as an extension activity of the University of Costa Rica.

Keywords: missing data, quality control, autoregressive filters, principal component analysis, free software applications.

Mathematics Subject Classification: 62-04, 62-06, 62-07.

1 Introducción

Todo aquel que trabaja con series de tiempo meteorológicas se encuentra con el problema que en muchos casos las series están incompletas. Algunos métodos de análisis pueden adecuarse a esta situación pero otros requieren que las series estén completas. En este trabajo se describen dos métodos de rellenado de datos ausentes.

Un sensor de una estación meteorológica capta señales al mismo tiempo de varios fenómenos con escalas espaciales y temporales diferentes. Aquellos de mayor extensión espacial serán detectados por varias estaciones mientras que los de menor extensión no. Los métodos multivariados permiten separar las señales mediante criterios estadísticos de tal forma que los componentes encontrados explican la variabilidad total de la señal. Es decir, no han perdido "información". En la mayoría de los casos se encuentra que esos

componentes de la señal están correlacionados con fenómenos meteorológicos identificables. El primer método que se estudia rellena los datos ausentes utilizando la información de estaciones climatológicamente cercanas utilizando la técnica multivariada de componentes principales (Tabony, 1983).

El segundo método es de utilidad en las situaciones, desgraciadamente muy comunes, donde no hay estaciones cercanas y el rellenado se debe hacer con la información de la estación misma. Este método puede recuperar la señal estacional y aquellas señales cuya persistencia en tiempo sean compatibles con el tiempo de muestreo. Este método utiliza modelos predictivos autoregresivos conocidos como $AR(p)$, el cual es un modelo lineal que utiliza los valores de p tiempos de muestreo anteriores y posteriores para estimar el valor en un tiempo dado (Ulrych & Bishop, 1975; Ulrych & Clayton, 1976).

Debe quedar claro que estos métodos son incapaces de reproducir los datos perdidos. Lo que verdaderamente sucedió se perdió irremediablemente. Estos métodos permiten rellenar las series con valores "razonables" que son consistentes con la estadística y la física de algunas de las señales captadas. Debido a lo anterior, el objetivo principal de este trabajo es el de desarrollar herramientas para el rellenado de datos ausentes de registros geofísicos utilizando la información de estaciones cercanas y/o el registro histórico de la misma estación. Estas herramientas son de utilidad tanto en las labores de investigadores y docentes en el campo de la climatología, así como para el trabajo operativo del personal de los servicios meteorológicos e hidrológicos, debido a que algunas veces se requieren conjuntos de datos completos en ciertas labores como por ejemplo cuando se realiza el análisis espectral.

2 Metodología

Para la programación de ambos métodos de rellenado de datos se usó el programa SCILAB y los lenguajes de comandos interpretados TCL/TK. SCILAB es un entorno numérico, de programación y gráfico desarrollado por el Institut Nationale de Recherche en Informatique et en Automatique (INRIA). Las fuentes, los ejecutables y manuales se pueden obtener gratis de http://www.scilab.org. TCL es una interfaz de usuario que interpreta comandos al igual que las interfaces bash, sh, korn, etc. , mientras que TK es un juego de herramientas de TCL que permite crear interfaces gráficas al usuario. Los interpretadores y manuales de la combinación TCL/TK se encuentran en http://www.tc.tk y http://www.etsimo.uniovi.es/pub/tcl. En el apéndice se incluyen los algoritmos (en seudocódigo) programados, junto con las fórmulas matemáticas utilizadas y los criterios de parada de los procesos iterativos.

2.1 Análisis en componentes principales

Los detalles sobre la metodología de las componentes principales han sido detalladas en Soley (2003) y Soley & Alfaro (1999)[1]. Este último trabajo discute su aplicación para estudios climáticos en Centroamérica.

[1]Una primera versión de las rutinas que utilizan esta metodología fue escrita en MATLAB por Eric Alfaro, Victor Jara y Pamela Sobarzo en 1996 en la Universidad de Concepción, Chile.

Se mencionó en la Introducción que el rellenado utilizando este método se realiza con estaciones climatológicamente cercanas. El concepto de cercanía se explicita tradicionalmente utilizando la matriz de covarianza o de correlación que cuantifican el grado de información común compartido entre estaciones. La matriz de covarianza se utiliza cuando en el análisis es importante conservar la diferencia en la amplitud o varianza de las estaciones; mientras que la matriz de correlación se utiliza cuando se desea un análisis más basado en la forma de las curvas de las estaciones estudiadas que en su amplitud. Entre más altos los valores de covarianza o correlación mas afines son las estaciones.

La primera etapa del proceso es la inspección cuidadosa de estas matrices para escoger el conjunto de estaciones idóneas: no sólo deben ser las estaciones climatológicamente cercanas, sino que también las secciones de datos ausentes no se deben traslapar. Si bien es cierto, esta parte del proceso es subjetiva y se basa grandemente en la experiencia de la persona que realiza el análisis, vale la pena resaltar dos puntos. Primero, el concepto de estaciones climatológicamente cercanas, por lo general lo que sugiere implícitamente es que la variabilidad del grupo de estaciones escogido este influenciada por los mismos fenómenos de gran escala (común a todas las estaciones). Segundo, si la escogencia de las estaciones se basa en el coeficiente de correlación entre las mismas, es conveniente utilizar algún criterio de significación que tome en cuenta la autocorrelación de las series de tiempo (eg. Ebisuzaki, 1997; Sciremammano, 1979).

La idea fundamental del método consiste en iterar sucesivamente por las siguientes tres etapas para ir obteniendo en cada iteración mejores estimados de las cantidades intermedias involucradas y de los valores estimados de los datos ausentes. Se siguieron las siguientes etapas:

1. Calcular la matriz de covarianza o correlación, $\mathbf{R}$ y obtener los vectores $\mathbf{E}$ y los valores $\mathbf{L}$ propios.

2. Calcular los componentes principales $\mathbf{Y} = \mathbf{X_o E L}^{-1/2}$. Donde la matriz $\mathbf{X_o}$ contiene los datos originales de tamaño nt (longitud de la serie de tiempo) x ns (número de estaciones usadas). Cabe destacar que los datos ausentes en primera instancia se sustituyen por promedios.

3. Tomando en cuenta que los valores originales se pueden recobrar mediante la ecuación $\mathbf{X_o} = \mathbf{Y L}^{1/2}\mathbf{E^T}$, estimar los valores ausentes con la expresión $\mathbf{X_a} = \mathbf{Y L'}^{1/2}\mathbf{E^T}$ utilizando los primeros k componentes principales únicamente. Se puede visualizar esto como equivalente a truncar la matriz de valores propios $\mathbf{L}$ a un tamaño k x k, con $k < ns$, o dejar esta matriz de iguales dimensiones e igualar a cero los $ns - k$ últimos valores propios que son los que menos contribuyen a la varianza total.

Recordemos aquí que los datos originales son una combinación lineal de los componentes principales en los que los factores de peso se calculan de los vectores y valores propios. Además, los componentes que más contribuyen a "explicar" la varianza total son los primeros mientras que los últimos sólo "explican" una fracción menor, ya que por lo general estas se asocian con ruido no correlacionado. En la primera iteración la matriz de correlación se calcula sólo con los pares de datos en los cuales no hay datos ausentes y

los componentes principales aproximando los valores ausentes con el promedio de la serie. Después de realizar las primeras tres etapas del método se obtiene una aproximación mejorada de los valores ausentes. Entonces se usan todos los valores para calcular la matriz de correlación y los componentes principales. Se repite entonces el procedimiento, esta vez con una matriz de correlación mejorada. Las iteraciones se continúan hasta que una de las siguientes condiciones se cumpla:

1. las diferencias entre los valores calculados entre dos iteraciones sucesivas, i e $i+1$, es menor que un valor e fijado por el usuario o $\|\mathbf{X}_a(i) - \mathbf{X}_a(i+1)\| < e$,
2. la diferencia máxima entre dos iteraciones sucesivas aumenta o $\max(\|\mathbf{X}_a(i) - \mathbf{X}_a(i+1)\|) > 0$,
3. el número de iteraciones excede un número especificado por el usuario.

Parte fundamental del método es determinar cuantos componentes principales utilizar. Soley (2003) discute este tema y describe varias maneras de escoger el número a utilizar. Estos métodos dependen del conocimiento previo del número de señales en los datos o de argumentos estadísticos. El programa descrito aquí utiliza el gráfico de "scree" (scree graph, ver Wilks, 1995) que muestra los autovalores vs. el número de componente principal. Este gráfico se ha modificado añadiéndole las barras de errores calculadas siguiendo el procedimiento sugerido por North *et al.* (1982). Estos autores determinaron que cuando las barras de error de dos autovalores se traslapan existe una degeneración efectiva de los modos traslapados. En otras palabras, los autovalores son iguales (degenerados) en la práctica. Cuando existe degeneración, los modos degenerados contienen la misma "información" y entonces no se puede incluir sólo uno de ellos porque entonces no se estaría tomando la señal completa. La Fig. 1 muestra un gráfico de scree típico y las barras de error. En este caso los dos primeros componentes principales son degenerados y del gráfico se aprecia que los dos primeros componentes son los que más contribuyen a la varianza total (27% aproximadamente) mientras que los ocho últimos contribuyen en menor grado. Del gráfico se concluye que el rellenado se puede llevar a cabo con los dos primeros componentes principales.

Para algunos conjuntos de datos se puede determinar el número de componentes a usar de una manera más confiable siguiendo el siguiente procedimiento:

1. Se escoge una sección de los datos sin valores ausentes.
2. Al azar se marcan como valores ausentes un número igual al porcentaje de datos ausentes en los datos completos. Los valores extraídos se reservan.
3. Se calculan los valores ausentes con 1, 2, 3, ... p componentes principales.
4. Utilizando una métrica apropiada (potencia del error, desviación absoluta, etc.) se calcula el error cometido con cada uno de los posibles números de componentes principales del punto anterior.
5. Se utiliza el número de componentes principales que minimiza el error.

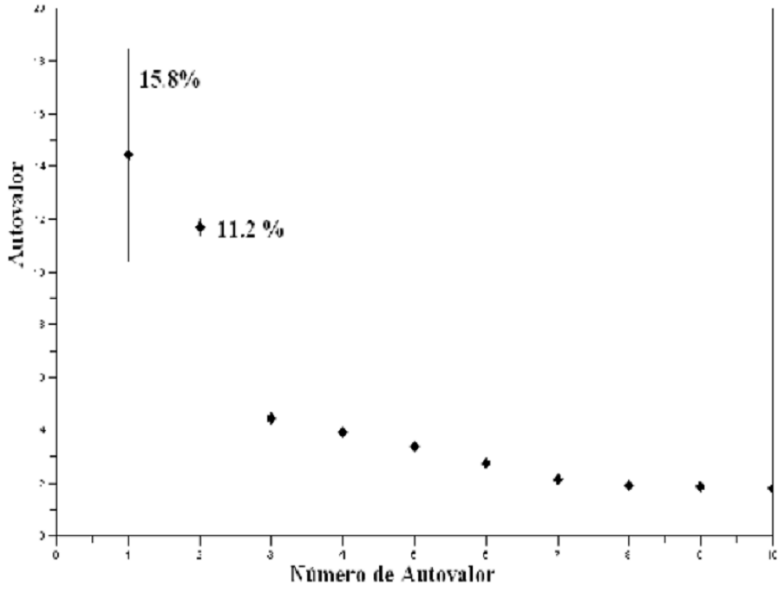

Figura 1: Gráfico "scree" para determinar el número de componentes a usar. En esta figura se muestran los primeros diez autovalores de 146 estaciones de datos diarios localizadas en Mesoamérica ($1971 - 2000$). Nótese que al haber un traslape de los dos primeros modos se deberían usar ambos para el rellenado de los datos.

En casos de duda se puede repetir el proceso varias veces para verificar si se obtiene el mismo número de componentes principales para diferentes conjuntos de datos "ausentes". Este método práctico de obtener el número óptimo toma en cuenta las particularidades de los diferentes conjuntos de datos mientras que los métodos más generales y teóricos no lo pueden hacer. Es importante resaltar que la práctica nos indica que no se puede aplicar un criterio único para determinar el número de componentes principales. Los autores no conocen de ningún estudio teórico que garantice la convergencia de este método de rellenado de datos ausentes. Como se verá más adelante, si el porcentaje de datos ausentes en series de temperatura mensual es alto (más de 25 %) el método en general no converge. Este artículo expone un método que empíricamente determina la zona de convergencia y la calidad del rellenado para combinaciones de porcentaje de datos ausentes y número de componentes principales que puede ser utilizado para otras variables y tiempos de muestreo.

2.2 Filtro predictivo $AR(p)$

Cuando no hay estaciones cercanas el rellenado de las datos ausentes debe hacerse con la información de la estación misma. Se debe recordar que el sensor capta señales de escalas temporales diferentes y aquellas señales cuya escala temporal es menor que el tiempo de

muestreo no pueden ser resueltas y actúan como ruido. Por ejemplo, si el muestreo es diario, se espera que los fenómenos cuya escala de tiempo es de unos días (p.e. ondas de los estes) pueden ser detectados y resueltos en la práctica. Esa señal viene acompañada que fenómenos de escala temporal de horas que la oscurecen y dificultan su detección. Una propiedad de los filtros predictivos $AR(p)$ es que pueden recoger señales cuya persistencia es comparable a la longitud del filtro. Además, estos filtros tienen la propiedad de que por el principio de Máxima Entropía los valores calculados son consistentes con las propiedades estadísticas de la serie sin incluir suposiciones externas a los datos. Es decir, aunque la información ausente se perdió, los valores rellenados son consistentes estadísticamente con el resto de la serie.

2.2.1 Algunos detalles de los modelos autoregresivos y filtros predictivos $AR(p)$

El modelo autoregresivo de orden $p, AR(p)$, obedece la ecuación

$$y[t] = \phi_1 y[t-1] + \phi_2 y[t-2] + \phi_3 y[t-3] + \ldots + \phi_p y[t-p] + x[t],$$

ésta nos dice que la salida y en tiempo t depende de los p valores anteriores de ella misma más el valor presente de la innovación x. Cuando se modelan señales con este modelo, los coeficientes se ajustan de tal manera que la innovación corresponda a ruido blanco con varianza mínima (Soley, 2005). El filtro predictivo correspondiente es

$$\widehat{y}[t] = \phi_1 y[t-1] + \phi_2 y[t-2] + \phi_3 y[t-3] + \ldots + \phi_p y[t-p].$$

El valor de la señal en tiempo t se pronostica con los p valores de la señal anteriores. El error que se comete es $x[t]$. Los dos métodos que utilizamos para calcular los coeficientes también corren el filtro de pronóstico en tiempo reverso, ahora el valor de la señal en tiempo t se pronostica con los p valores futuros de la señal,

$$\widehat{y}[t] = \phi_1 y[t+1] + \phi_2 y[t+2] + \phi_3 y[t+3] + \ldots + \phi_p y[t+p].$$

El error total de pronóstico es

$$error = 1/nt \sum_{t=p+1}^{t=nt} (y[t] - \widehat{y}[t])^2 + 1/nt \sum_{t=1}^{t=nt-p} (y[t] - \widehat{y}[t])^2.$$

Nótese que ambos filtros predictivos se corren dentro de los datos sin salirse de los extremos. Los coeficientes ϕ_i se calculan de tal forma que el error total se minimice. El programa elaborado en este trabajo utiliza dos algoritmos para calcular los coeficientes autoregresivos minimizando el error total: el estimador de Burg (Ulrych & Bishop, 1975) y el propuesto por Ulrych & Clayton (1976). El primero es desarrollado para procesos estocásticos estacionarios y el segundo para series determinísticas. El estimador de Burg es muy utilizado en la práctica y ha sido estudiado extensamente. Un resumen comprensivo de sus propiedades y en particular, de las condiciones que garantizan su estabilidad, se encuentran en Kay & Marple (1981) para las diversas implementaciones que existen de

este algoritmo. Por otro lado, el estimador de Ulrych & Clayton corresponde a un ajuste de mínimos cuadrados clásico.

Teóricamente ambos métodos deben converger con las series de tiempo meteorológicas usuales. En la práctica ambos métodos pueden divergir por motivos externos como acumulación de los errores de truncamiento de la representación de los números en punto flotante, por posibles errores en la adquisición de los datos que introducen inestabilidades, por utilizar un número de coeficientes autoregresivos incongruente con la persistencia de la serie o porque el porcentaje de valores ausentes es alto.

Dependiendo del tipo de sistema operativo, *Linux* o *Windows*, el algoritmo de Burg se realiza de dos maneras distintas. En *Linux* se utiliza la subrutina memcof.c de *Numerical Recipes in C* (Press et al., 1992). En *Windows* no fue posible utilizar el mismo método debido a dificultades en adquirir un compilador adecuado. Por lo tanto se programó el algoritmo de Burg directamente en SCILAB siguiendo las ecuaciones en Kay & Marple (1981). En la práctica se encontró que los dos métodos producen resultados iguales dentro de un 1%. Como el método de Ulrych y Clayton corresponde a mínimos cuadrados en los dos sistemas operativos se usan las funciones propias de SCILAB.

Como vimos anteriormente, los datos se filtran hacia adelante y atrás en tiempo. Como los filtros se aplican dentro de los datos el filtro hacia adelante no produce salida para los primeros p valores y el filtro hacia atrás no produce salida para los últimos p valores. Los valores intermedios se suman y se dividen por dos. Para los valores de los extremos se toma la única salida disponible.

3 Resultados

Se utilizaron sendos archivos de comandos TCL/TK para crear las interfaces gráficas al usuario de la subrutinas que realizan el análisis llamadas rellena.sci y llenaar.sci y mostradas en las Figs. 2 y 3, las cuales se activan desde la línea de comandos de SCILAB por el usuario. Ambas le ofrecen al usuario en términos generales las siguientes opciones: *Inicializar* las rutinas, elegir el código de datos ausentes que por defecto es *Nan*. Se puede utilizar otro código siempre que sea numérico (-9999, por ejemplo) y el valor utilizado se debe digitar en la casilla de texto correspondiente. La interfaz cambia el código numérico a *Nan*. En *Windows* sólo se puede utilizar el código numérico. A la fecha, los autores no saben la razón de esto.

El usuario elige el nombre del archivo con los datos de entrada que se puede digitar en la casilla de texto *Archivo texto con datos* o al pulsar *Buscar* se abre una ventana que permite navegar por el sistema de directorios hasta encontrar el archivo deseado y escogerlo, una vez hecha alguna de estas dos cosas el usuario pulsa *Cargar* para leerlos.

Para el caso de la Fig. 2 del análisis de componentes principales el usuario debe especificar también el número de modos a usar, el número de iteraciones máximo que ejecutara la rutina y la diferencia máxima de los valores estimados entre dos iteraciones sucesivas. Para el caso del método autoregresivo, Fig. 3, el usuario debe escoger el número de coeficientes y también el método con el que se van a calcular los coeficientes: el de Burg o el de Ulrych y Clayton.

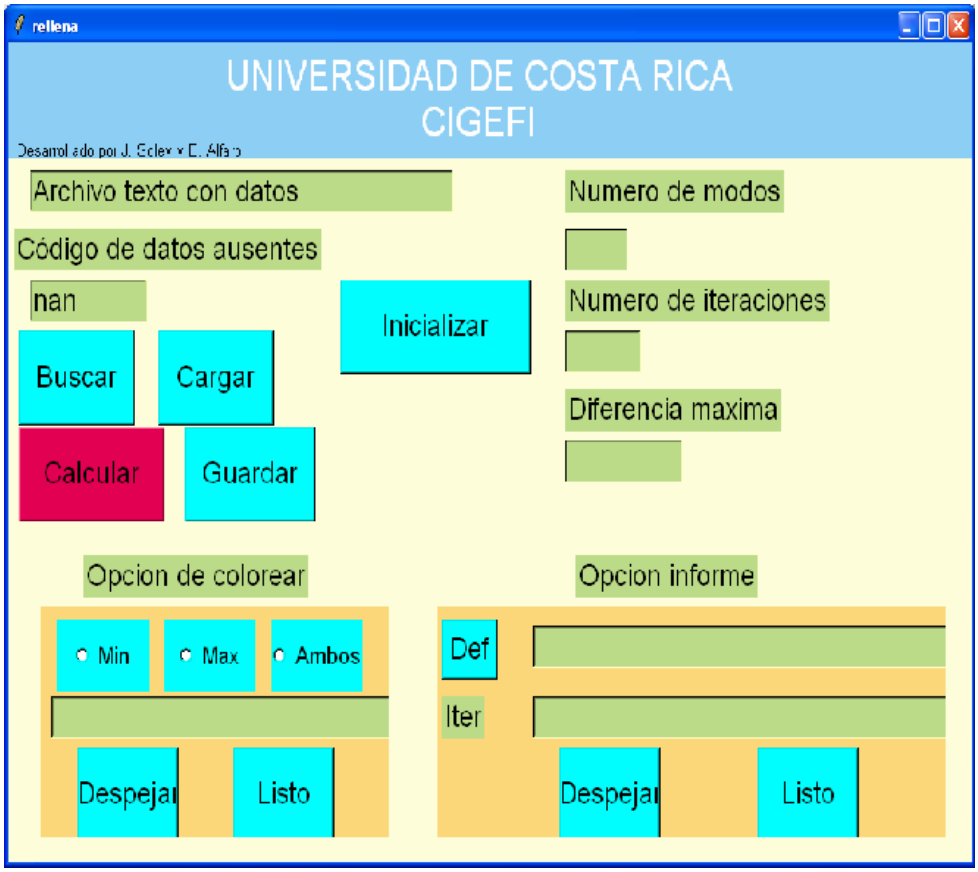

Figura 2: Interfaz gráfica al usuario de la subrutina rellena.sci elaborada en este trabajo. Las opciones de la interfaz se describen en el texto.

El análisis se realiza una vez que el usuario pulse *Calcular*. Una vez realizada la estimación de los datos faltantes, los nuevos datos rellenados se pueden acceder desde la línea de comandos de SCILAB en el arreglo *nuevos*. Si se quiere hacer un registro permanente de los datos rellenados el usuario pulsa *Guardar*. Esto activa una ventana que permite especificar el directorio y el nombre del archivo en el que se guardan. El archivo es creado por fprintfMat en formato CSV y puede ser incorporado fácilmente a una hoja electrónica.

El usuario puede activar en la Fig. 2 la opción *Min, Max* ó *Ambos* si los datos tienen una cota mínima, una cota máxima o cota mínima y máxima, en otras palabras, si se conoce que físicamente la variable no puede presentar esos valores, por ejemplo valores de lluvia negativos o valores de humedad relativa negativos o mayores a cien. El usuario también tiene la opción de guardar un informe del proceso ya sea al especificar el nombre del archivo en la casilla de texto o pulsando *Def* para activar la ventana de definición del archivo de salida.

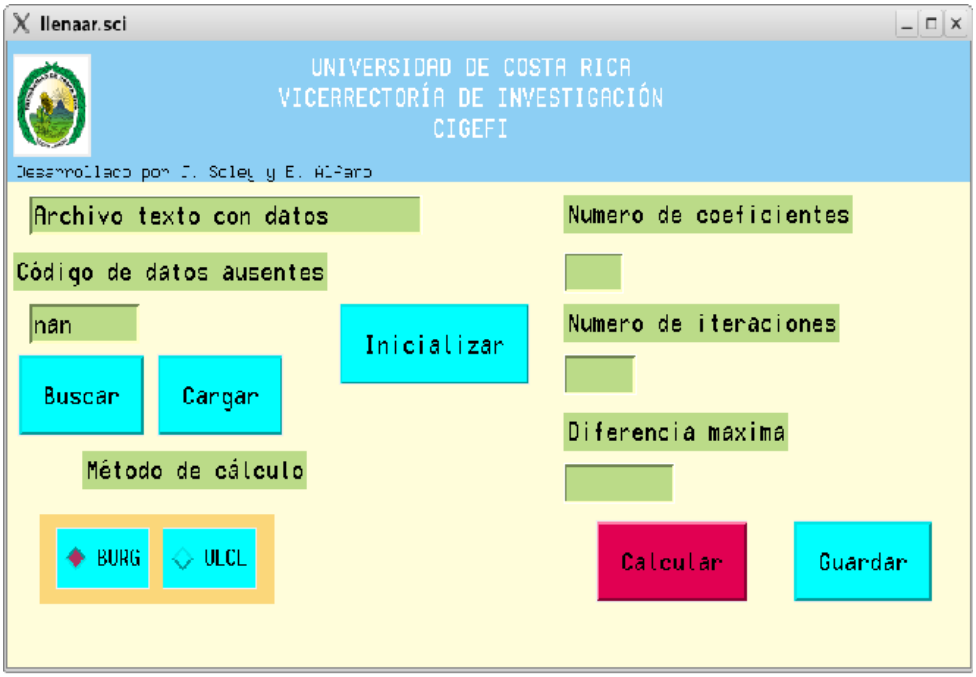

Figura 3: Interfaz gráfica al usuario de la subrutina llenaar.sci elaborada en este trabajo. Las opciones de la interfaz se describen en el texto.

Adicionalmente se desarrollaron funciones en Scilab con el objetivo de calcular el error cuadrático medio y el error absoluto medio del ajuste de datos ausentes en función del porcentaje de datos ausentes y del número de modos (caso de componentes principales) o coeficientes (caso del modelo autoregresivo) usados en el ajuste a partir de conjuntos de datos completos y así poder tener una estimación de la bondad del ajuste. Los pasos que se siguen para calcular ambos errores para un porcentaje de datos ausentes dado y un número de modos o coeficientes dado es:

1. introducir aleatoriamente el porcentaje de datos ausentes en la(s) serie(s) de tiempo de prueba,

2. estimar los valores ausentes y el número de modos o coeficientes,

3. calcular el error cuadrático medio y el error absoluto medio entre los valores estimados y los reales,

4. almacenar los errores calculados en una hipermatriz.

Se llama una realización del experimento a la repetición de los pasos anteriores para todos los valores de porcentaje de datos ausentes y con todos los números de coeficientes AR o de modos de interés. Se deben hacer suficientes realizaciones para que el promedio y varianza de los errores cuadrático medio y absoluto medio de cada ajuste tengan estabilidad estadística, sin embargo se debe tomar en cuenta que la corrida de estas realizaciones consume mucho tiempo computacional. Para el caso específico del cálculo de los coeficientes autoregresivos, se utilizará el método de UC para calcular los coeficientes AR por dos razones teóricas: es un ajuste de mínimos cuadrados que se puede aplicar a series estacionarias y no estacionarias. Por otro lado, siendo las series no estacionarias, el uso del método de Burg no tiene justificación teórica a pesar que en la práctica da resultados parecidos y es más rápido que UC.

Cabe destacar aquí que los resultados del error calculados dependerán de cada experimento en sí, si bien es cierto se han descrito anteriormente algunos considerándos que el usuario debe tener en cuenta, la decisión final de conservar o no los datos estimados dependerá del juicio experto del usuario, es decir, al analizar sus resultados no sólo desde el punto de vista estadístico sino también al tomar en cuenta consideraciones físicas.

Los datos usados para ejemplificar las rutinas que evalúan el error fueron compilados por el Centro de Ciencias de la Atmósfera en la UNAM, México. Este conjunto de datos se produjo a partir de los registros de estaciones terrenas, datos de satélite y salidas de modelos numéricos, combinados en una rejilla con una resolución espacial de 0.5° de longitud x 0.5° de latitud (Magaña *et al.*, 1999).

En la Fig. 4 se muestra el promedio del Error Cuadrático Medio calculado al utilizar una serie de tiempo de 494 valores de temperatura mensual del punto de rejilla 87.0° W y 5.5° N. La serie se inicia en enero de 1958 y termina en febrero de 1999, con media de 28.03 °C y varianza de 1.05 °C^2. Se hicieron 100 realizaciones y se estableció un porcentaje máximo de datos faltantes de 30% y un número máximo de coeficientes autoregresivos de 13.

En la Fig. 5 se muestra el promedio del Error Cuadrático Medio calculado al utilizar series de tiempo de 15944 valores de precipitación diaria de los puntos de rejilla: 87.5-5.0, 87.0-6.0, 87.5-5.5, 87.5-6.0, 86.5-5.0, 86.5-6.0, 87.0-5.5, 86.5-5.5 y 87.0-5.0 (°W-°N). Las series inician el 01/01/1958 y terminan el 26/08/2001. Se hicieron 100 realizaciones y se estableció un porcentaje máximo de datos faltantes de 30% y un número máximo de modos de 9.

Vale la pena hacer notar de las Figs. 4 y 5, que el uso de muchos coeficientes autoregresivos (Fig. 4) o modos (Fig. 5), no necesariamente garantiza la convergencia del método hacia algún valor estable de los datos que se requieran rellenar.

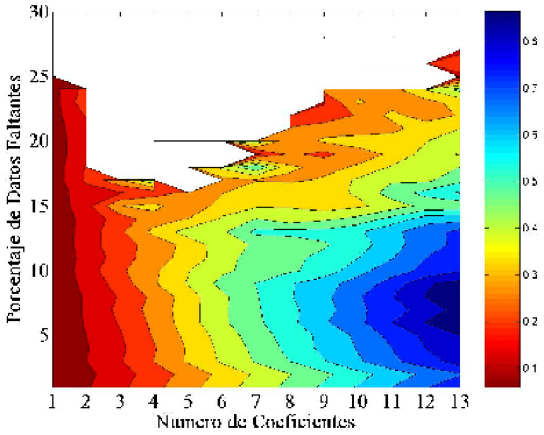

Figura 4: Promedio del Error Cuadrático Medio, en °C, de 100 realizaciones (de la rutina de evaluación elaborada en este trabajo) para la serie de temperatura superficial del aire. Los valores en blanco indican que la metodología no convergió a ningún valor estimado para ese dato faltante en ninguna de las realizaciones.

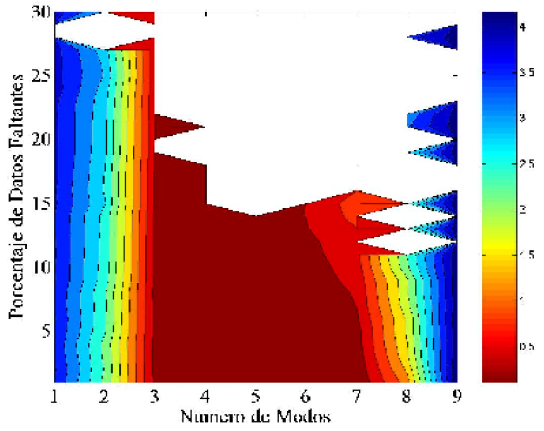

Figura 5: Promedio del Error Cuadrático Medio, en *mm*, de 100 realizaciones (de la rutina de evaluación elaborada en este trabajo) para las series de precipitación. Los valores en blanco indican que la metodología no convergió a ningún valor estimado para ese dato faltante en ninguna de las realizaciones.

4 Discusión

Por medio de la programación de los algoritmos matemáticos utilizando principalmente software de licencia libre, como el programa SCILAB, en ambientes *Windows* y *Linux*, y la elaboración de interfaces gráficas al usuario (GUI) por medio del uso del TCL/TK, se pudo desarrollar una herramienta para el rellenado de datos ausentes utilizando información de estaciones cercanas y otra para el rellenado de datos ausentes utilizando el registro histórico de la misma estación. Cabe destacar que el uso de una GUI, facilitó el empleo de las rutinas para el usuario no familiarizado con los programas antes descritos.

Para el desarrollo de la herramienta de rellenado de datos ausentes utilizando información de estaciones correlacionadas entre si, se utilizó la metodología propuesta por Tabony (1983), basada en el análisis de componentes principales. Este método se basa en que si se tiene ns variables con nt observaciones cada una, se puede tener un conjunto nuevo de k variables mediante la relación: $\mathbf{Y} = \mathbf{X_o E}$, donde $\mathbf{X_o}$ es la matriz de datos originales $nt \times ns$, $\mathbf{E}$ es una matriz de $ns \times k$ cuyas k columnas son los autovectores de la matriz de correlación $\mathbf{R}$. En general, los k vectores propios asociados a los k valores propios grandes representan las variaciones de gran escala (variabilidad que se quiere conservar) y explican la mayor parte de la variabilidad del conjunto de datos en $\mathbf{X_o}$, mientras que los $ns - k$ vectores propios asociados a los $ns - k$ valores propios más pequeños representan variaciones de pequeña escala (ruido no correlacionado en general). El conjunto de datos de las variables originales en las ns estaciones puede ser recuperado en forma aproximada por la relación $\mathbf{X_a} = \mathbf{Y E^T}$, en donde si $k = ns$ entonces $\mathbf{X_a} = \mathbf{X_o}$. En resumen, el análisis de componentes principales utiliza la información de estaciones correlacionadas al descomponer la información en "patrones" que capturan y separan la variabilidad de las series, donde se espera que recupere los efectos de escala mayor y que se pierdan los de escala menor, además, podría utilizarse para rellenar brechas amplias. Una aplicación al campo de la precipitación en Centroamérica se encuentra en Alfaro & Cid (1999).

Por otra parte, para el desarrollo de un método de rellenado de datos ausentes utilizando el registro histórico de la misma estación se utilizaron las metodologías propuestas por Ulrych & Bishop (1975) y Ulrych & Clayton (1976), basadas en el uso de un filtro predictivo autorregresivo de orden p o $AR(p)$. Este método utiliza la información de la misma serie en el principio de máxima entropía, por lo que los valores predichos siguen la estadística de la serie sin alterarla. El mismo podría incorporar efectos de pequeña escala y es útil sólo para brechas cortas. Para el cálculo de los coeficientes del proceso $AR(p)$ se utilizaron dos aproximaciones: 1) Método recursivo de Burg (Ulrych & Bishop, 1975) y 2) Método de mínimos cuadrados (Ulrych & Clayton, 1976). Estos dos métodos suponen procesos estacionarios, por lo que hay que desestacionalizar las series de tiempo. Adicionalmente, existen criterios objetivos para determinar el orden de los procesos que representan las series, aunque el orden óptimo para modelar y el orden óptimo para predicción no coinciden necesariamente.

En ambas metodologías, los procesos se puede hacer en forma iterativa hasta que $\|\mathbf{X}_a(i) - \mathbf{X}_a(i+1)\| < e$, donde, e es un valor definido por el usuario.

A pesar de que el usuario debe tener siempre presente que cualquier dato faltante estimado utilizando estas metodologías es FALSO y que en el mejor de los casos se obtiene

una aproximación grosera a la realidad, en ambos métodos la estimación de los datos faltantes no alteró las propiedades estadísticas de las series de tiempo.

Estas herramientas podrían ser de utilidad tanto en las labores de investigadores y docentes en el campo de la climatología, así como para el trabajo operativo del personal de los servicios meteorológicos e hidrológicos. Sin embargo, se debe tener en cuenta que la decisión final de conservar o no los datos estimados dependerá del juicio experto del usuario, es decir, al analizar sus resultados no sólo desde el punto de vista estadístico sino también al tomar en cuenta consideraciones físicas.

5 Apéndice

Algoritmo para el rellenado de datos con componentes principales.

Inicio:

S: matriz con los datos de entrada de dimensiones número de estaciones $(ns) \times longitud$ de las series de tiempo (nt) y con los datos ausentes codificados.

N: la matriz **S** con cada columna normalizada (media cero y desviación estándar unitaria).

np: número de componentes principales utilizados para el rellenado $(np < ns)$

1. Calcula la matriz de correlación $\mathbf{R_n}$, $\mathbf{R_n} = \frac{1}{nt}\mathbf{N_n}^T\mathbf{N_n}$.

2. Calcula los vectores propios $\mathbf{E_n}$ y valores propios $\mathbf{L_n}$, $\mathbf{R_n} = \mathbf{E_n}^T\mathbf{L_n}\mathbf{E_n}$.

3. Calcula los np modos ortogonales $\mathbf{N_{n+1}}$ usando la matriz de autovalores $\mathbf{L_n}$ truncada a dimensiones $np \times np$ que llamaremos $\mathbf{T_n}$, $\mathbf{N_{n+1}} = \mathbf{N_n}\mathbf{E_n}\mathbf{T_n^{-1/2}}$.

4. Ajusta la media y la varianza de cada columna de $\mathbf{N_{n+1}}$ para que correspondan a las de **S**.

5. Sustituye en $\mathbf{S_n}$ los valores ausentes por los valores correspondientes calculados en $\mathbf{N_{n+1}}$ y obtiene la matriz $\mathbf{S_{n+1}}$.

6. Calcula la diferencia absoluta máxima entre dos iteraciones, `máxima diferencia actual` = `max(abs(` $\mathbf{S_{n+1}} - \mathbf{S_n}$`))`.

7. Si `máxima diferencia actual` $\geq$ `máxima diferencia previa` entonces termina en caso contrario continua con el bucle.

8. `máxima diferencia actual := máxima diferencia previa`.

9. `n := n+1`

Fin del bucle:

Fin:

Algoritmo para el rellenado de datos con un filtro predictivo $AR(p)$.

Inicio:

D : vector columna $(1 \dots N)$ con los datos con los ausentes codificados.
p: orden del filtro predictivo

1. Sustituye los valores ausentes en **D** por el valor medio para obtener $\mathbf{D_0}$.

   `n := 0`

   Inicio del bucle:

   Ejecuta hasta `Fin del bucle` mientras $n \leq$ `número máximo de iteraciones` y `máxima diferencia` $>$ `máxima diferencia especificada`

2. Calcula los coeficientes autoregresivos $\phi_1, \phi_2, \dots, \phi_p$ que modelan $\mathbf{D_n}$ mediante el algoritmo de Burg.

3. Filtra $\mathbf{D_n}$ hacia adelante para obtener el vector columna $\mathbf{F_n}$.

   $$f_n[k] = \phi_1 d_n[k-1] + \phi_2 d_n[k-2] + \phi_3 d_n[k-3] + \ldots + \phi_p d_n[k-p] \qquad p+1 \leq k \leq N.$$

4. Filtra $\mathbf{D_n}$ hacia atrás para obtener el vector columna $\mathbf{B_n}$.

   $$b_n[k] = \phi_1 d_n[k+1] + \phi_2 d_n[k+2] + \phi_3 d_n[k+3] + \ldots + \phi_p d_n[k+p] \qquad 1 \leq k \leq N-p.$$

5. Construye $\mathbf{D_{n+1}}$ como sigue:

   (a) $\mathbf{D_{n+1}}[1..p] = \mathbf{B_n}[1..p]$,
   (b) $\mathbf{D_{n+1}}[N-p \dots N] = \mathbf{F_n}[N-p \dots N]$,
   (c) $\mathbf{D_{n+1}}[p+1 \dots N-p-1] = \frac{1}{2}(\mathbf{F_n}[p+1 \dots N-p-1] + \mathbf{B_n}[p+1 \dots N-p-1])$.

6. Calcula la diferencia absoluta máxima entre dos iteraciones,

   `máxima diferencia` $=$ `max(abs(`$\mathbf{D_{n+1}} - \mathbf{D_n}$`))`.

7. `n := n+1`

   `Fin del bucle:`

6 Agradecimientos

Este trabajo se realizó como parte del proyecto de extensión $ED-1977$ de la Universidad de Costa Rica (UCR). Se agradece también a los proyectos: $805-A7-002, 808-A7-520, 805-98-506$, UCR y CRN-2050-IAI.

Referencias

[1] Alfaro, E.; Cid, L. (1999) "Ajuste de un modelo VARMA para los campos de anomalías de precipitación en Centroamérica y los índices de los océanos Pacífico y Atlántico Tropical", *Atmósfera*, **12**(4): 205–222.

[2] Ebisuzaki, W. (1997) "A method to estimate the statistical significance of a correlation when the data are serially correlated", *J. Climate* **10**: 2147–2153.

[3] Kay, S.M.; Marple, S.L. (1981) "Spectrum analysis – A modern perspective", *Proc. IEEE,* **69**: 1380–1419.

[4] Magaña, V.; Amador, J.; Medina, S. (1999) "The midsummer drougth over Mexico and Central America", *Journal of Climate* **12**: 1577-1588.

[5] North, G.R.; Bell, T.L.; Cahalan, R.F.; Moeng, F.J. (1982) "Sampling errors in the estimation of empirical orthogonal functions", *Mon. Wea. Rev.* **110**: 699–706.

[6] Press, W.H.; Teukolsky, S.A.; Vetterling, W.A.; Flannery, B.P. (1992) *Numerical Recipes in C: the Art of Scientific Computing.* Cambridge University Press, Cambridge.

[7] Sciremammano, F. (1979) "A suggestion for the presentation of correlations and their significance levels", *J. Phys. Oceanogr.* **9**: 1273–1276.

[8] Soley, F.J. (2003) *Análisis en Componentes Principales.* Notas de clase del curso SP-5906, Métodos Digitales de Análisis de Secuencias Temporales. Programa de Posgrado en Ciencias de la Atmósfera. Sistema de Estudios de Posgrado Universidad de Costa Rica (Accesible en `http://fjsoley.com`).

[9] Soley, F.J. (2005) *Sistemas lineales ARMM(p,q) con* $p + q \leq 4$. *Primera Parte: Sistemas lineales AR* ($p \leq 4$). Notas de clase del curso SP-5906, Métodos Digitales de Análisis de Secuencias Temporales. Programa de Posgrado en Ciencias de la Atmósfera. Sistema de Estudios de Posgrado. Universidad de Costa Rica (Accesible en `http://fjsoley.com`).

[10] Soley, F.J.; Alfaro, E. (1999) "Aplicación de análisis multivariado al campo de anomalías de precipitación en Centroamérica", *Tóp. Meteor. Oceanog.* **6**(2): 71–93.

[11] Tabony, R.C. (1983) "The Estimation of Missing Climatological Data", *Journal of Climatology* **3**: 297–314.

[12] Ulrych T.J.; Bishop, T.N. (1975) "Maximum Spectral Analysis and Autoregressive Decomposition", *Reviews of Geophysics and Space Physics* **13**(1): 183–200.

[13] Ulrych T.J.; Clayton, R.W. (1976) "Time Series Modeling and Maximum Entropy", *Physics of the Earth and Planetary Interiors* **12**: 188–200.

[14] Wilks, D. (1995) *Statistical Methods in the Atmospheric Sciences.* Academic Press, New York.