Propuestas metodológicas para el rellenado de datos ausentes en series de tiempo geofísicas.

Guía Práctica de uso.

Pablo Ureña[1], Eric J. Alfaro[1, 2, 3] y F. Javier Soley[1] †

1. Centro de Investigaciones Geofísicas
2. Escuela de Física
3. Centro de Investigaciones en Ciencias del Mar y Limnología

Universidad de Costa Rica

Correo electrónico: juan.urenamora@ucr.ac.cr, erick.alfaro@ucr.ac.cr

Agosto de 2016

Universidad de Costa Rica

Disponible a través de los siguientes repositorios:
http://kerwa.ucr.ac.cr/
http://kimuk.conare.ac.cr/

Índice

Resumen 3

Abstract 3

1. Introducción 3

2. Análisis en componentes principales 4

2.1 Definición de la matriz de covarianza y de correlación 5

2.2 Definición de los componentes principales 6

2.3 Idea fundamental del método 6

2.4 Determinación del número de componentes principales a utilizar en el rellenado 7

2.5 La subrutina rellena.sci 8

2.6 Ejemplo de rellenado de datos ausentes con rellena.sci 11

3. Filtro predictivo AR(p) 17

3.1 Modelos autoregresivos y filtros predictivos AR(p) 17

3.2 Otros detalles de los métodos 18

3.3 Filtrado de los datos 18

3.4 La subrutina llenaar.sci 18

3.5 Ejemplo de rellenado de datos ausentes con llenaar.sci 20

4. Método Integrado rellenafull.sci 24

4.1 Metodología y forma de uso 25

4.1.1 Descripción del programa de rellenado de datos integrado 29

4.1.2 Ejemplo de una corrida del programa 30

4.2 Rutinas para la preparación y evaluación de los datos 31

4.3 Descripción del rellenado por filtro autoregresivo AR(m) en la parte integrada 32

4.4 Rellenado por el método de componentes principales en la parte integrada 33

4.5 Ejemplo de comparación del método integrado con el de Componentes Principales previo 33

5. Agradecimientos 40

6. Referencias 40

Resumen

En un trabajo reciente, Alfaro y Soley (2009), presentaron dos metodologías para el rellenado de datos ausentes, enfocadas hacia su uso en series de tiempo geofísicas. Una de ellas se basó en la descomposición en componentes principales de la matriz de correlación de datos de una misma variable. El segundo método consistió en ajustar un modelo autoregresivo a la serie de tiempo y utilizar ese modelo como estimador de los datos ausentes. En el presente trabajo se propone además, una combinación de las metodologías anteriores, al utilizar el método autoregresivo como una primera aproximación de rellenado que luego es utilizada en forma iterativa por el método de componentes principales. A modo de ejemplo, esta nueva aproximación se compara luego con la propuesta original del método de rellenado por componentes principales utilizado por Alfaro y Soley (2009), usando los datos de los puntos de rejilla del Caribe costarricense de Johnson et al. (2003). El programa utilizado para la elaboración de estar rutinas es SCILAB, el cual tiene la ventaja de que es código abierto.

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

Abstract

In a recent paper, Alfaro and Soley (2009) presented two methods for filling missing data in geophysical time series. One of them was based on principal component decomposition of the correlation matrix build with common time series records of the same variable. The other method adjusts an autoregressive model to the time series which is then used to estimate the missing data. In this work we add an additional option because as a first approximation of the missing data, we fill the data with the autoregressive method which is then used iteratively by the principal components method. As an example, for this new approach, it is then compared with the original principal components methodology used by Alfaro and Soley (2009), using data from the grid points of the Costa Rican Caribbean slope from Johnson et al. (2003). The program used for the production of these routines is SCILAB, which has the advantage of being open source.

Keywords: missing data, auto-regressive filters, principal component analysis, free software applications.

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 esta nota se describen dos métodos de rellenado de datos ausentes, que posteriormente se propondrán para su uso en forma conjunta.

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 describe en esta nota 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 un modelo predictivo autoregresivo conocidos como AR(p) que 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 y Bishop, 1975; Ulrych y Clayton, 1976).

Como una tercera opción, se propone incluir el método de rellenado por modelo autoregresivo como una primera aproximación para las iteraciones del método de componente principales, a su vez se actualizan las rutinas a la versión más reciente de SCILAB y se agregan los criterios de selección *Akaike Infomation Criterion* o AIC y *Bayesian Information Criterion* o BIC para escoger automáticamente el orden del filtro autoregresivo (Wilks, 2011). Se lleva a cabo un experimento con varias series completas de datos de puntos de rejilla del Caribe costarricense de Johnson et al. (2003), con el propósito de comparar gráficamente el error estadístico de la rutina original de Componentes Principales y la rutina modificada para esta nueva aproximación.

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.

Esta nota hace uso extenso de la teoría desarrollada en dos notas anteriores (Soley, 2003; 2005):

- Sistemas Lineales ARMM(p,q) con $p+q \leq 4$. Primera Parte: Sistemas Lineales AR($p \leq 4$).(de aquí en adelante Nota ARP4)
- Análisis en Componentes Principales. (de aquí en adelante Nota ACP)

Por último, para poder ejecutar las subrutinas el usuario debe tener instalados en su computadora el Programa SCILAB. 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.

2. Análisis en componentes principales.

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 más 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 primordialmente 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. Segundo, si la escogencia de las estaciones se basa en el coeficiente de correlación entre las mismas, es conveniente utilizar algún criterio de significancia que tome en cuenta la autocorrelación de las series de tiempo (eg. Ebisuzaki, 1997; Sciremammano, 1979).

La segunda etapa consiste en descomponerla matemáticamente en tres matrices muy particulares como se explica a continuación.

2.1 Definición de la matriz de covarianza y de correlación.

Representemos la serie de tiempo de una estación *i* y a la que a cada valor se la ha restado el valor medio por el vector columna $si = [s_{1i} s_{2i} \dots s_{nti}]^T$ donde nt es la longitud total de la serie. Es decir **si** es la secuencia de desviaciones respecto a la media de la estación i. Las ns estaciones escogidas deben tener un periodo común y de igual longitud nt. Los vectores columna ***si*** forman una matriz **S** de tamaño nt x ns.

La autocovarianza y la covarianza cruzada entre estaciones se definen

$$cov(si, si) = var(si) = \frac{1}{nt}\sum_{k=1}^{nt} s_{ki}^2 = \frac{1}{nt} si^T si,$$

$$cov(si, sj) = cov(sj, si) = \frac{1}{nt}\sum_{k=1}^{nt} s_{ki}\, s_{kj} = \frac{1}{nt} si^T sj.$$

Los elementos a lo largo de la diagonal corresponden a la autocovarianza (o de forma más sencilla, a la varianza) de las estaciones. Luego, la traza, que es la suma de los elementos diagonales, equivale a la varianza total.

La auto correlación y la correlación cruzada se definen similarmente,

$$corr(si, si) = \frac{var(si)}{\sigma_i^2} = \frac{1}{nt}\sum_{k=1}^{nt} \frac{s_{ki}^2}{\sigma_i^2} = \frac{1}{nt} si'^T si',$$

$$corr(si, sj) = corr(sj, si) = \frac{1}{nt}\sum_{k=1}^{nt} \frac{s_{ki} s_{kj}}{\sigma_i \sigma_j} = \frac{1}{nt} si'^T sj',$$

donde $si' = si/\sigma_i$ son las desviaciones normalizadas.

Las matrices de covarianza y de correlación se forman tomando como elementos la covarianza o correlación cruzada entre estaciones. Las dos matrices son cuadradas de dimensiones ns x ns y simétricas.

Una matriz simétrica **M** se puede descomponer como el triple producto matricial de dos matrices **M = E^T L E**. La matriz cuadrada **E** se conoce como la matriz de vectores propios (autovectores, eigen vectores) en la cual los ns vectores propios son sus columnas. **L** es la matriz de valores propios (autovalores, eigen valores) y es diagonal (ver la Nota ACP para más detalles). Usualmente se ordenan los vectores propios y valores propios de acuerdo al valor descendente de los autovalores. Los autovectores y auto valores tienen las siguientes propiedades:

1. Dos vectores propios correspondientes a valores propios diferentes son ortogonales,

   E^T E = **E E^T** = **I** ,donde **I** es la matriz identidad .

2. Todos los ns valores propios son reales.

3. A m ≤ ns valores propios iguales les corresponde m vectores propios ortogonales.

4. La traza (**M**) = traza (**L**) = $\lambda_1 + \lambda_2 + ... + \lambda_{ns}$. Es decir, la suma de los autovalores es igual a la varianza total de las estaciones.

La última propiedad es muy importante en la práctica porque cuantifica el aporte a la varianza total de cada uno de los vectores propios (Lorenz, 1956 y Wilks, 2011).

2.2 Definición de los componentes principales.

La matriz de componentes principales **A** se define **A** = **S E** $\mathbf{L}^{-1/2}$ en la cual cada una de las columnas es un componente principal. **A** tiene las mismas dimensiones que la matriz de datos **S** pero con la propiedad de que la columnas, los componentes principales, son ortogonales entre sí: **A** $\mathbf{A}^T$ = **I**. La matriz **A** tiene la misma "información" que **S,** sólo que reordenada de tal forma que cada columna tiene "información" independientes de las otras. De la matriz **A** se recobran los datos originales mediante la relación **S** = **A** $\mathbf{L}^{1/2}$ $\mathbf{E}^T$ (Breteherton et al., 1992). Esta última relación es la base para el rellenado de datos faltantes y se pueden hacer dos observaciones sobre ella. Primero nos indica que los datos originales son una combinación lineal de las componentes principales en los que los factores de peso se calculan de los vectores y valores propios. En segundo lugar los componentes que más contribuyen a "explicar" la varianza total son los primeros mientras que los últimos sólo "explican" una fracción menor, y por lo general estos se asocian con ruido no correlacionado.

2.3 Idea fundamental del método.

La idea fundamental del método es la siguiente y tiene las siguientes etapas:

1. Calcular la matriz de covarianza o correlación y obtener los vectores (**E)** y los valores propios (**L)**.

2. Calcular los componentes principales **A**= **S E** $\mathbf{L}^{-1/2}$.

3. Estimar los valores ausentes con la expresión **S'** = **A L'** $^{1/2}$ $\mathbf{E}^T$ utilizando los primeros componentes principales únicamente. Se puede visualizar esto como equivalente a truncar la matriz de valores propios a un tamaño np x np (np < ns) o a igualar a cero los últimos valores propios (los de valor propio menor y que contribuyen poco a la varianza total).

En la primera iteración la matriz de covarianza o 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 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 es menor que un valor fijado por el usuario,
2. la diferencia máxima entre dos iteraciones sucesivas aumenta,
3. el número de iteraciones excede un número especificado por el usuario.

2.4 Determinación del número de componentes principales a utilizar en el rellenado.

Parte fundamental del método es determinar cuántos componentes principales utilizar. En la Nota ACP se toma este tema y se describen 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 utiliza el gráfico de "scree" (*scree graph*, ver Wilks, 2011) que muestra los autovalores vs el índice del componente principal. Este gráfico se ha modificado añadiéndole las barras de errores calculadas siguiendo a 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 Figura 1 muestra un gráfico de scree típico y las barras de error. En este caso los cinco componentes principales son no degenerados y cada uno contiene varianza que es ortogonal a la varianza de los otros. Del gráfico se aprecia que los dos primeros componentes son los que más contribuyen a la varianza total mientras que los tres últimos contribuyen en menor grado. Del gráfico se concluye que el rellenado se puede llevar a cabo con los dos primeros componentes principales.

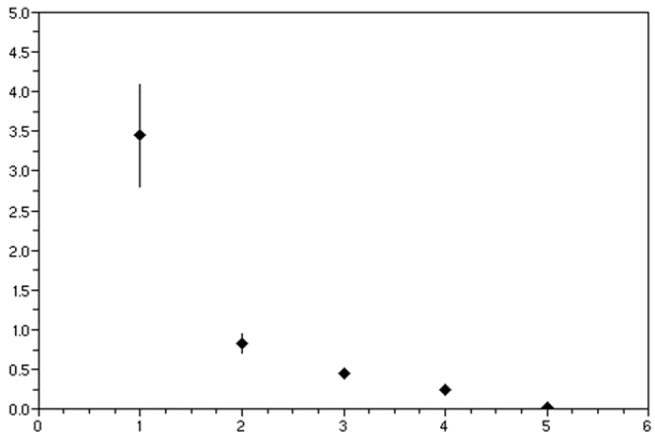


Fig. 1: Gráfico "scree" para determinar el número de componentes a usar.

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.

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. En Alfaro y Soley (2009) se utiliza este procedimiento para estudiar la convergencia del rellenado de series de precipitación diaria con el número de componentes principales y el porcentaje de datos faltantes.

2.5 La subrutina rellena.sci.

La subrutina rellena.sci realiza las iteraciones en las tres etapas del método. Originalmente esta subrutina fue escrita en MATLAB por Eric Alfaro, Victor Jara y Pamela Sobarzo en 1996 en la Universidad de Concepción. La transcripción con ligeras modificaciones a SCILAB la realizó Javier Soley. En esta manual las subrutinas cuyo nombre incluyen la extensión sci son escritas por alguno de los autores mientras aquellas sin extensión son subrutinas propias de SCILAB.

Después de leer los datos se calculan la media y varianza de cada estación utilizando nanmean y nanstdev. En la primera iteración la subrutina corcru.sci hace el primer cálculo (grosero) de la matriz de correlación en el que para cada par de estaciones se reduce el número de pares de valores a aquellos en los que los dos valores son válidos. Luego se calcula la correlación cruzada con la subrutina mvcorrel. Los valores y autovectores propios se calculan con bdiag. Los valores ausentes se aproximan por el valor medio de los valores presentes de la estación correspondiente que se calculan usando nanmean. Los valores calculados con $\mathbf{S'} = \mathbf{A}\,\mathbf{L'}^{1/2}\,\mathbf{E}^{\mathbf{T}}$ se "recolorean" para que tengan la misma media y varianza que los datos originales (subrutinas normaliza.sci y colorea.sci). En las siguientes iteraciones la matriz de correlación se calcula con mvcorrel usando los valores ausentes rellenados por la iteración anterior. El resto del procedimiento sigue igual hasta que se cumplan alguna de las tres condiciones que terminan el proceso, descritos anteriormente en la sección 2.3.

Nótese que en el párrafo anterior sólo se menciona la matriz de correlación aunque el método teórico también admite la matriz de covarianza.

Se debe tener el siguiente cuidado en el proceso de "recolorización" porque algunas variables tienen cotas. Por ejemplo, 0 es el valor mínimo de precipitación o la rapidez del viento por ejemplo, mientras que humedad relativa está limitada entre 0 y 100%. El programa permite una opción (lcolor) que después de colorear revisa los valores que excedan las cotas máximas o mínimas, y si los hay cambia los valores

que las exceden al valor de las cotas.

La subrutina se llama así:

nuevos = rellena(datos, nmodos, difmax, itmax, informe, lcolor) .

Los argumentos son:

1. **datos**: es la matriz de datos en la cual las columnas son las ns diferentes estaciones de la misma variable y todas de igual longitud nt. Los datos ausentes deben estar codificados con Nan.
2. **nmodos**: el número de autovalores que se usarán

   Opciones:

   a) 0 si quiere usar el gráfico scree para determinarlo

   b) el número de autovalores (1:ns)
3. **difmax**: diferencia máxima permitida entre dos iteraciones sucesivas
4. **itmax**: número de iteraciones máximas permitidas
5. **informe**: lista ['archivo_salida', [it1 it2 ...])

   a) 'archivo_salida': tira de caracteres con nombre de archivo donde guardar el informe

   b) [it1 it2 ...]: arreglo con iteraciones para las cuales guardar el informe , p.e. [0 10 20]. 0: denota el inicio de la iteración

   c) [] si no se quiere obtener el informe
6. **lcolor**: lista con opciones para colorear los datos

   Opciones:

   [] : no se activa la opción

   ('min', mínimo) ó ('max', máximo) ó ('ambos',mínino y máximo). Sustituye por mínimo y/o máximo los valores menores y/o mayores.
7. **nuevos**: matriz nt x ns con los valores ausentes rellenados.

El formato de entrada de los datos es un archivo texto con las estaciones en columnas separadas por espacios o tabulación y con datos ausentes codificados por Nan. Las primeras filas de un archivo de cinco estaciones se muestran a continuación:

```
-->tmp10f
 tmp10f  =


   Nan     12.9   Nan     19.9    25.9
   21.2    12.6    19.6   Nan     Nan
   23.8    14.5    21.9    22.1    27.3
   24.3    14.6    21.8    22.4    26.9
```

24.6 15.7 21.9 Nan 27.

24.4 16.2 21.7 21.9 26.8

23.5 Nan 20.9 21.3 26.1

La opción informe permite obtener copia de las cantidades calculadas en las iteraciones especificadas en el vector y que se guardan en el archivo con el nombre especificado por el usuario. Así se puede utilizar la matriz de correlación, componentes principales, autovalores y autovectores para su graficación o posterior análisis.

Una salida típica de rellena.sci se muestra abajo para una corrida en la que se utilizan dos componentes principales para rellenar. Estos dos componentes principales explican casi el 92% de la varianza total. La diferencia máxima permitida es 0.05 y el número de iteraciones máximo 50. Las opciones de informe y colorear están deshabilitadas. En la iteración 11 la diferencia máxima aumentó por lo que se terminó el proceso. En este caso rellena.sci reporta entonces los resultados de la iteración 10 que corresponden a una diferencia de 0.124089.

->tmp10f2=rellena(tmpdes10f,2,0.05,50,[],[]);

iter= 1 delta_max = 0.330467

iter= 2 delta_max = 0.301684

iter= 3 delta_max = 0.270480

iter= 4 delta_max = 0.240562

iter= 5 delta_max = 0.213289

iter= 6 delta_max = 0.189019

iter= 7 delta_max = 0.167672

iter= 8 delta_max = 0.148990

iter= 9 delta_max = 0.132660

iter= 10 delta_max = 0.124089

Diferencia no disminuyó

iter= 11 delta_max = 0.126190

Varianza acumulada por autovalor

1 75.435605

2 91.873580

3 96.768095

4 99.441402

5 100.000000

Varianza acumulada por autovalor

1 75.322211

2 91.882081

3 96.771810

4 99.441199

5 100.000000

2.6 Ejemplo de rellenado de datos ausentes con rellena.sci.

Para el ejemplo se utilizan datos de precipitación acumulada mensual de las estaciones Juan Santamaría, San José, Argentina de Grecia, Fabio Baudrit, Pavas y Embalse la Garita. Las seis están situadas en el sector oeste del Valle Intermontano Central de Costa Rica y del registro de información se extrajeron 34 años (408meses). El total de datos ausentes es 375 y en la Tabla 1 se da el número de datos ausentes de cada estación.

	Juan Santamaría	***San José***	***Argentina de Grecia***	***Fabio Baudrit***	***Pavas***	***Embalse La Garita***
Datos ausentes	8 (1.960 %)	4 (0.980%)	14 (3.43%)	5 (1.225%)	271 (66.4%)	73 (17.89%)

Tabla 1. Desglose del número de datos ausentes de las estaciones.

El registro completo de las estaciones Argentina de Grecia y Fabio Baudrit se muestra en la Figura 2 y en la 3 un segmento de este registro. Las seis estaciones muestran el ciclo anual con las épocas seca y lluviosa claramente marcadas. La época lluviosa tiene un comportamiento bimodal debido al veranillo o canícula (Mid Summer Drought, MSD, en inglés de acuerdo a Magaña et al. (1999)).

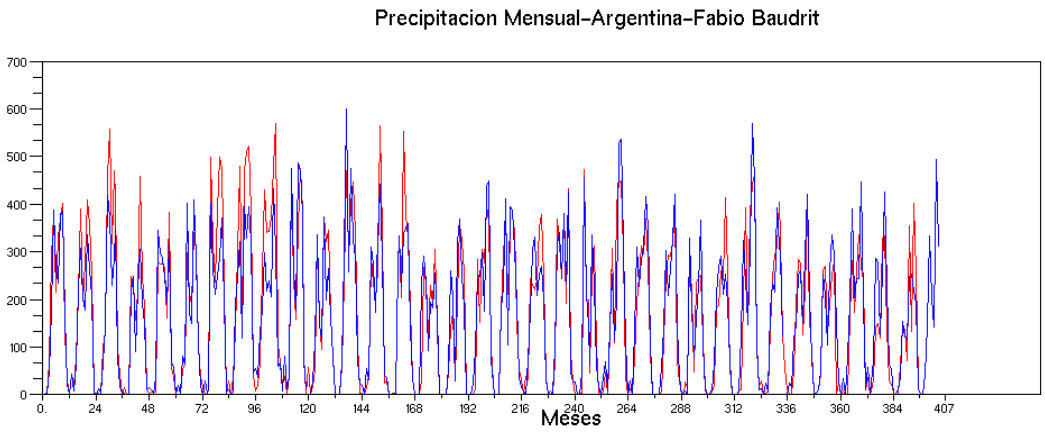


Fig. 2: Precipitación acumulada mensual de las estaciones Argentina de Grecia (azul) y Fabio Baudrit (rojo).

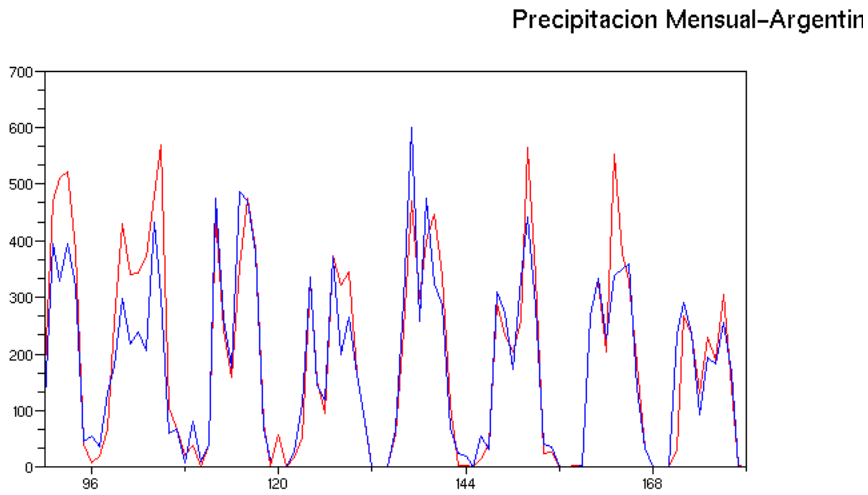


Fig. 3. Segmento de la Figura 2 que muestra dos brechas de datos ausentes en la estación Fabio Baudrit (rojo).

La matriz de autocorrelación de la Tabla 2 se calcula con los pares de datos válidos de cada una de las parejas posibles descartando la parejas en que haya al menos un dato ausente. La tabla muestra valores de correlaciones cruzadas cercanos a 0.9 lo que indica que comparten señales climatológicas comunes.

	Juan Santamaría	***San José***	***Argentina de Grecia***	***Fabio Baudrit***	***Pavas***	***Embalse La Garita***
Juan Santamaria	1.	0.88769	0.88519	0.93575	0.91999	0.90750
San José	0.88769	1	0.89071	0.88265	0.89265	0.86064
Argentina	0.88519	0.89071	1	0.91995	0.91240	0.90334
Fab. Baudrit	0.93575	0.88265	0.91995	1	0.89253	0.91520
Pavas	0.91999	0.89265	0.91240	0.89253	1	0.86631
La Garita	0.90750	0.86064	0.90334	0.91520	0.86631	0.86631

Tabla 2. La matriz de correlaciones cruzadas entre las seis estaciones muestra que comparten señales climatológicas comunes.

En las corridas es necesario especificar una cota mínima de 0 por ser los datos de precipitación acumulada. En la primera corrida también se puso a 0 el número de modos para así obtener el gráfico "scree" que se muestra en la Figura 4. El primer modo claramente domina sobre los demás por lo que se decidió utilizar un modo para el rellenado de datos. Los resultados de la corrida indicaron que la varianza acumulada en el primer modo es superior a 90% reforzando la bondad de la escogencia.

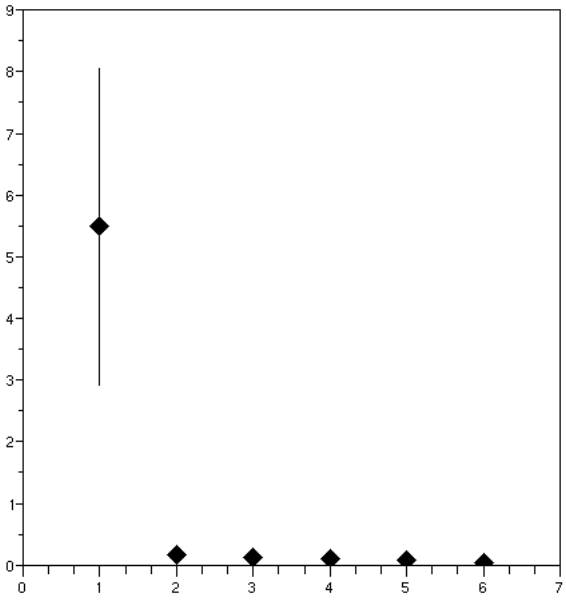


Fig. 4. El gráfico "scree muestra un predominio del primer modo.

La corrida final se realizó con los siguientes parámetros: número de modos, 1; número de iteraciones, 30; diferencia máxima, 1; cota mínima 0. El archivo texto con los datos es estacionesch.txt y el código de dato ausente utilizado -9999.

La salida en el Marco 1 apunta que la diferencia máxima inicial de 11.2 mm se redujo en 15 iteraciones a la diferencia máxima especificada. El primer modo recoge el 93.07% de la varianza total y el segundo modo sólo aporta cerca de un 2% adicional. La Figura 5 muestra un segmento del registro de la estación Juan Santamaría con los valores rellenados en rojo que son consistentes con el comportamiento estacional y la variabilidad propia de la serie de tiempo. De las seis estaciones, Pavas tiene la menor longitud de registro durante el periodo de interés. En la Figura 6 se muestra la serie rellenada en base a la información de las otras estaciones.

iter= 1 delta_max = 11.222092
iter= 2 delta_max = 6.278696
iter= 3 delta_max = 5.301103
iter= 4 delta_max = 4.589270
iter= 5 delta_max = 3.990977
iter= 6 delta_max = 3.474592
iter= 7 delta_max = 3.025945
iter= 8 delta_max = 2.635296
iter= 9 delta_max = 2.294890
iter= 10 delta_max = 1.998211
iter= 11 delta_max = 1.739659
iter= 12 delta_max = 1.514371
iter= 13 delta_max = 1.318109
iter= 14 delta_max = 1.147168
iter= 15 delta_max = 0.998311
Varianza acumulada por autovalor
1 93.073574
2 95.359178
3 97.167828
4 98.532166
5 99.443361
6 100.000000

Marco 1. Salida de la última corrida.

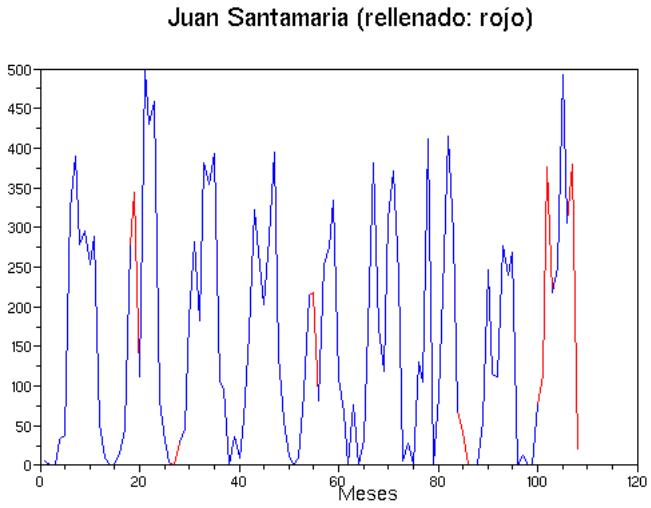


Fig. 5. Segmento de la serie de la estación Juan Santamaría con los valores rellenados en rojo.

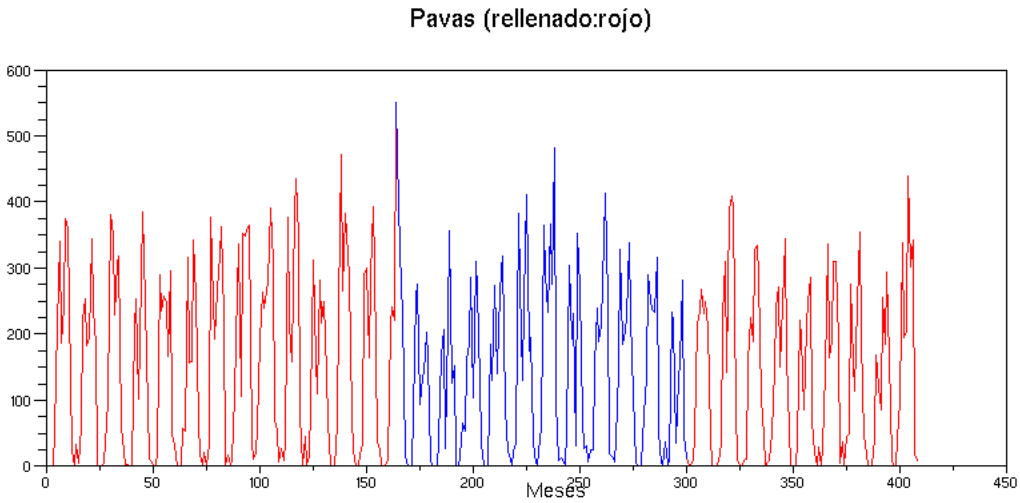


Fig. 6. Serie de la estación Pavas con los valores rellenados en rojo.

La Opción Informe es muy útil en la depuración del método de rellenado y para obtener un listado de los cálculos intermedios. La matriz de correlación de la Tabla 1 se obtuvo digitando 0 en Iter y el nombre del archivo de salida en Def. La iteración 0 se refiere a la primera aproximación del apartado 2.3. A continuación se muestra la salida de esta opción cortada después de los primeros valores del primer componente (Marco 2).

Primera aprox. a matriz de autocorrelacion 6 x 6

1. 0.887690615 0.885190426 0.935759363 0.919998315 0.907504422

0.887690615 1. 0.890719703 0.882658733 0.892650033 0.860643122

0.885190426 0.890719703 1. 0.919959022 0.912407252 0.903347984

0.935759363 0.882658733 0.919959022 1. 0.892533945 0.915206152

0.919998315 0.892650033 0.912407252 0.892533945 1. 0.86631623

0.907504422 0.860643122 0.903347984 0.915206152 0.86631623 1.

Primera aprox. autovalores

5.49126014 0.158555409 0.118662961 0.110055774 0.0809898221 0.0404758896

Primera aprox. autovectores

-0.411662755 -0.11833186 0.621519517 0.237657087 -0.018813369 -0.611072829

-0.402407764 0.602988648 -0.343444113 0.595927863 0.00726139887

0.0365523858

-0.409788368 0.0211559461 -0.512055136 -0.562484517 -0.200281387 -0.461436878

-0.412427009 -0.311163167 0.0890810166 0.0670329851 -0.702632998

0.476402631

-0.407712434 0.41116801 0.396323795 -0.491307087 0.329721171 0.39691332

-0.405400782 -0.596718559 -0.261822824 0.161629775 0.597554612 0.166824093

1 modos ortogonales

-2.18547947

-13.4057151

-16.7800712

-156.525224

-577.118087

-902.908971

-507.050908

-649.51123

.....

.....

Marco 2. Salida parcial de la Opción Informe con los resultados intermedios de la primera aproximación.

3. Filtro predictivo AR(p).

Cuando no hay estaciones cercanas el rellenado de los datos ausentes debe hacerse con la información de la estación misma. Recordemos 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 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.

3.1 Modelos autoregresivos y filtros predictivos AR(p).

El modelo autoregresivo de orden *p, AR(p),* obedece la ecuación de diferencias finitas

$$y[k] = \phi_1 y[k-1] + \phi_2 y[k-2] + \phi_3 y[k-3] + \cdots + \phi_p y[k-p] + x[k].$$

Nos dice que la salida *y* en tiempo *k* 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 (Ver la nota ARP4). El filtro predictivo correspondiente es,

$$\hat{y}[k] = \phi_1 y[k-1] + \phi_2 y[k-2] + \phi_3 y[k-3] + \cdots + \phi_p y[k-p].$$

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

$$\overleftarrow{y}[k] = \phi_1 y[k+1] + \phi_2 y[k+2] + \phi_3 y[k+3] + \cdots + \phi_p y[k+p].$$

Ahora el valor de la señal en tiempo *k* se pronostica con los *p* valores futuros de la señal.

El error total de pronóstico es

$$error = 1/nt \sum_{k=p+1}^{k=nt} (y[k] - \hat{y}[k])^2 + 1/nt \sum_{k=1}^{k=nt-p} (y[k] - \overleftarrow{y}[k])^2$$

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 el error total se minimice. El programa llenaar.sci utiliza dos algoritmos para calcular los coeficientes autoregresivos minimizando el error total: el estimador de Burg (Ulrych y Bishop, 1975) y el propuesto por Ulrych y Clayton (1976). El primero es desarrollado para procesos estocásticos estacionarios y el segundo para series determinísticas. En el estimador de Burg se incorpora la relación de recurrencia propia de procesos autoregresivos estacionarios conocida como relación de recurrencia de Levinson.

$$\phi_k^p = \phi_k^{p-1} + \phi_p^p \phi_{p-k}^{p-1}, k = 1 \ldots p.$$

El segundo subíndice se refiere a la iteración del proceso recursivo. Esta relación garantiza que los polos de la función de transferencia se encuentren dentro del círculo unitario del plano complejo. Por otro lado, el estimador de Ulrych y Clayton corresponde a un ajuste de mínimos cuadrados clásico.

3.2 Otros detalles de los métodos.

El programa es desarrollado en Linux, para el algoritmo de Burg este se transcribe a SCILAB y se llama de la siguiente manera:

[autoreg, parcor, erro] = pred_error(entran, nar)

Como el método de Ulrych y Clayton corresponde a míninos cuadrados se usan las funciones propias de SCILAB. La subrutina que realiza el método es aruc.sci.

3.3 Filtrado de los datos.

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. La subrutina filpred.sci realiza estos pasos.

3.4 La subrutina llenaar.sci.

El proceso iterativo para rellenar los valores ausentes los lleva a cabo la subrutina llenaar.sci que hace las siguientes tareas:

1. Calcula la media de los datos utlizando únicamente los datos válidos con nanmean.
2. Sustituye los valores ausentes con la media como una primera aproximación grosera.
3. Calcula los coeficientes del filtro predictivo con el método y para el número de coeficientes especificados por el usuario.
4. Filtra los datos con filpred.sci.
5. Calcula la diferencia máxima entre los valores filtrados y con los que se calcularon los coeficientes autoregresivos. Se la diferencia máxima es mayor que la tolerada y no se ha excedido el número de iteraciones vuelve al paso 3.

La subrutina llenaar.sci se ejecuta con el comando

```
salen = llenaar(entran, nar, tol, itmax, opcion) .
```

Sus parámetros de entrada son

1. **entran**: es un vector columna con los datos. Los datos ausentes deben estar codificados como

Nan.

2. **nar**: orden del filtro predictivo con el que se va a rellenar los datos ausentes.
3. **tol**: diferencia máxima tolerada entre valores predichos en iteraciones sucesivas
4. **itmax**: número máximo de iteraciones permitidas
5. **opcion**: BURG|ULCL se calculan los coeficientes *AR* con el método de BURG o de Ulrych-Clayton

y devuelve los datos rellenados en el vector columna salen.

En *Linux* llenaar.sci busca memcofsci.o en su mismo directorio y lo liga automáticamente. Si no lo encuentra se produce un error.

El archivo de entrada es una columna de valores.

Nan

-3.258696

-1.697674

-2.244444

-1.634783

-1.206522

-1.875556

-1.893750

Nan

-1.215217

-2.218182

-2.411111

-1.567442

-1.858696

La escogencia del número de coeficientes del modelo autoregresivo en la etapa 5 es crucial. Hemos utilizado tres métodos:

1. conceptos de modelado de series autoregresivas,
2. el criterio de persistencia, y
3. cálculo del error de rellenado.

En el primero se inspeccionan funciones o criterios de diagnóstico como la función de autocorrelación, función de correlación parcial y los criterios de Akahike y Parzen. Para aplicar el criterio de persistencia es crucial el conocimiento experto de los fenómenos importantes que generan la serie de interés. El ejemplo de la sección siguiente usa estos dos métodos.

El tercero es similar al descrito la sección 2.4. En Alfaro y Soley (2009) se utiliza para estudiar la convergencia del rellenado de series de temperatura mensual con el número de coeficientes autoregresivos y el porcentaje de datos faltantes.

3.5 Ejemplo de rellenado de datos ausentes con llenaar.sci.

Como ejemplo se escogió la serie de temperatura máxima diaria de la estación Fabio Baudrit. Se extrajeron los datos del 1-01-1976 al 31-05-2001 que comprende 9283 valores diarios. Este período contiene una mayoría de brechas de pocos días y una brecha de 31 días. Esta combinación de brechas sirve para ilustrar el comportamiento aceptable del método para las brechas cortas y el comportamiento menos aceptable para las brechas largas. Dos años, 1979 y 1980 se muestran en la Figura 7 y es evidente que los datos tienen una componente anual. La Figura 8 es el espectro de amplitud y los picos anual y semianual están localizados en los valores de índice de frecuencia 26 y 52. Para obtener el espectro se aproximaron los valores ausentes con la media del día juliano correspondiente.

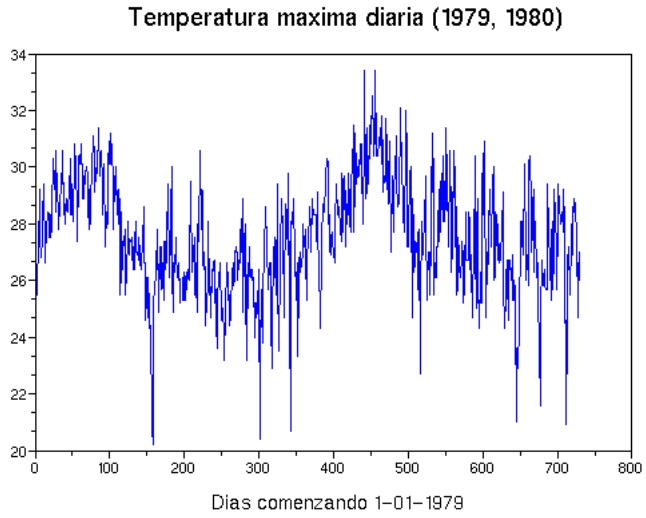


Fig. 7. Dos años (1979 y 1980) del registro de temperatura máxima diaria de la estación Fabio Baudrit. La componente estacional anual es evidente.

La serie se desestacionalizó mediante el método de las anomalías y el resultado de los años 1979 y 1980 se muestran en la Figura 9. El espectro de amplitud de esta serie en la Figura 8 (en rojo) no muestra los picos anual ni semianual. Para el cálculo del espectro se aproximaron los datos ausentes a cero.

Como el método de rellenado se base en los coeficientes autoregresivos es conveniente verificar si la serie desestacionalizada se puede modelar con un modelo autoregresivo ya que el orden de este modelo definirá el número de coeficientes a usar. Se observa que la función de autocorrelacion en la Figura 10 no decae a cero sino que se estabiliza a partir del retardo 6 o 7. Ese comportamiento no es típico de modelos autoregresivos puros. Una inspección de la correlación parcial y los criterios de Akahike y de Parzen en la Tabla 3 revela que ninguno de los tres presenta un mínimo. Todo esto indica que para modelar la serie se debe usar un modelo ARMM por lo que se debe definir el número de coeficientes de otra forma. Recordemos que es posible deducir un modelo AR infinito equivalente a un modelo ARMM. Por tanto hay justificación en utilizar los coeficientes autoregresivos aunque no sean un modelo que cumpla con el principio de parsimonia en este caso. Como en la práctica no se pueden utilizar un número infinito de coeficientes, se debe escoger un subconjunto finito teniendo como guía la persistencia de la serie. Si el número de coeficientes es menor a la persistencia medida en tiempos de muestreo, se está perdiendo información valiosa y aumenta el error. Si se excede el número de coeficientes, los valores afuera del tiempo de persistencia introducen ruido al cálculo de los valores ausentes y también aumenta

el error. En este caso, las señales meteorológicas importantes tienen persistencias de 3 a 6 días, los cuales corresponderían principalmente a fenómenos meteorológicos de escala sinóptica por lo que se hizo el rellenado con 6 coeficientes.

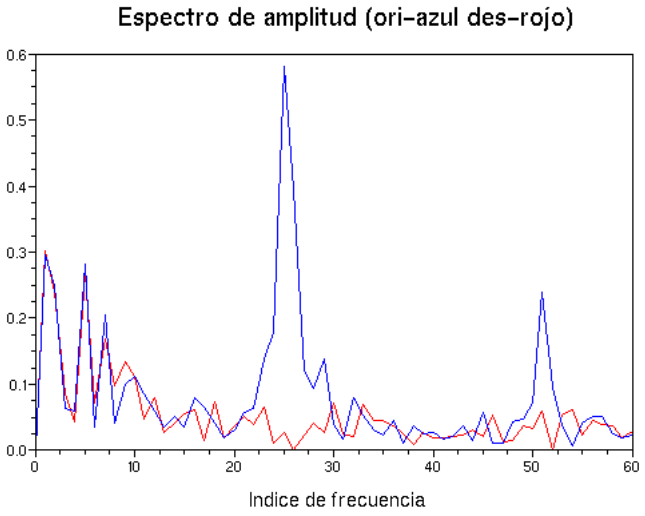


Fig. 8. Espectro de amplitud de la serie original (azul) y desestacionalizada (rojo) de la serie de temperatura máxima diaria de la estación Fabio Baudrit.

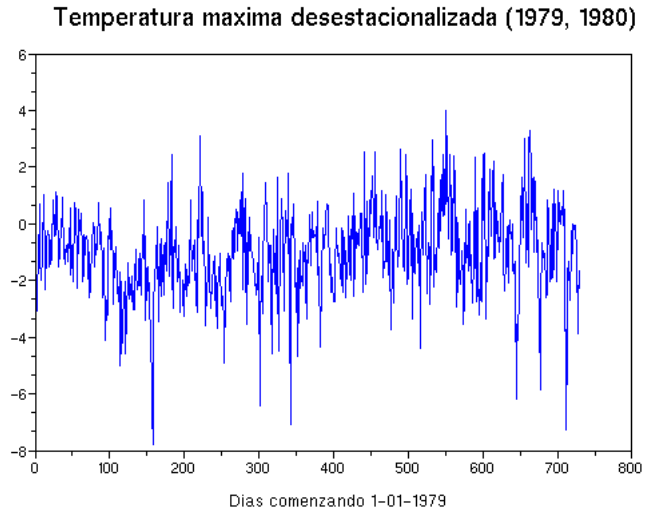


Fig. 9. Los años 1979 y 1980 del registro de temperatura máxima diaria de la estación Fabio Baudrit desestacionalizados.

Orden	Coef. AR	Corr. Parcial	Akahike	Parzen
1	3.94626e-01	5.23407e-01	7.47492e+03	-4.54907e-01
2	7.30349e-02	1.48035e-01	7.26664e+03	-4.65002e-01
3	4.52808e-02	1.05732e-01	7.16195e+03	-4.70162e-01
4	3.63562e-02	8.93124e-02	7.08794e+03	-4.73843e-01
5	4.29071e-02	8.71217e-02	7.01764e+03	-4.77368e-01
6	4.71614e-02	8.05046e-02	6.95793e+03	-4.80382e-01
7	4.57225e-02	6.52290e-02	6.91947e+03	-4.82335e-01
8	7.04135e-03	3.33734e-02	6.91089e+03	-4.82771e-01
9	4.69210e-02	5.99947e-02	6.87867e+03	-4.84414e-01
10	3.29640e-02	3.29640e-02	6.87036e+03	-4.84839e-01

Tabla 3. Coeficiente autorregresivo, correlación parcial, criterio de Akahike y criterio de Parzen de la serie desestacionalizada de temperatura máxima diaria.

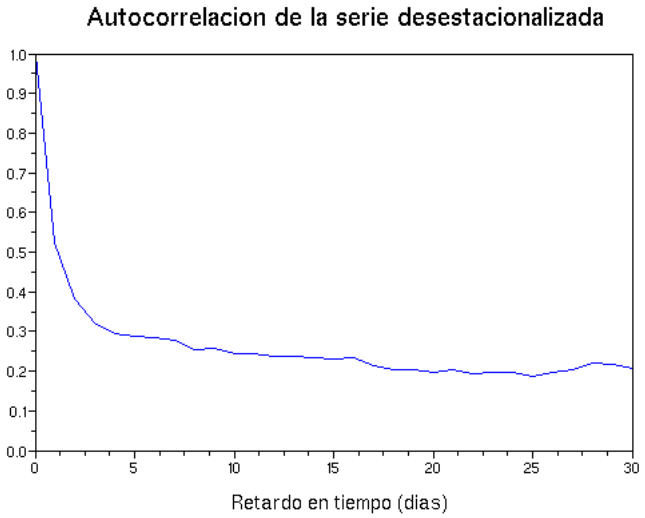


Fig. 10. Autocorrelación de la serie desestacionalizada de temperatura máxima diaria.

Los parámetros finales utilizados fueron: 6 coeficientes, una diferencia máxima de 0.01 °C y un tope de 20 iteraciones. La diferencia máxima se alcanzó después de 5 iteraciones. Las Figuras 11 y 12 muestran los valores rellenados en rojo para dos brechas de 4 y 5 tiempos de muestreo. La variabilidad de los datos rellenados es un tanto menor pero si no estuvieran diferenciados por el color podrían pasar por datos reales. Los valores rellenados para la brecha de 31 tiempos de muestreo de la Figura 13 siguen la

tendencia a la baja pero la menor variabilidad los distingue fácilmente de los datos reales.

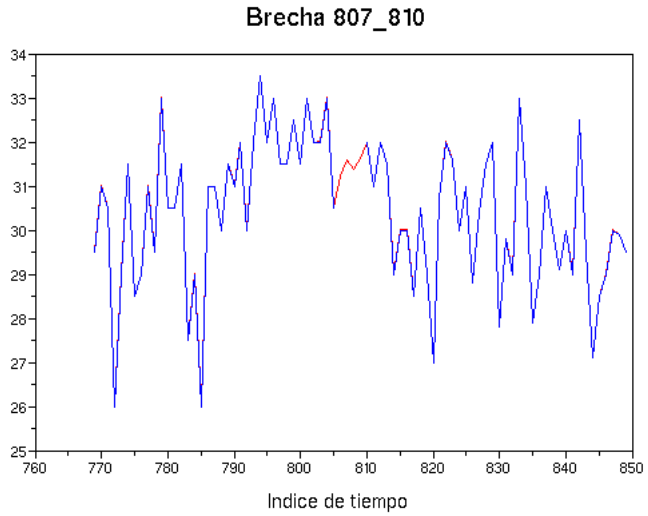


Fig. 11: Valores rellenados (en rojo) en la brecha con índices de tiempo 807-810.

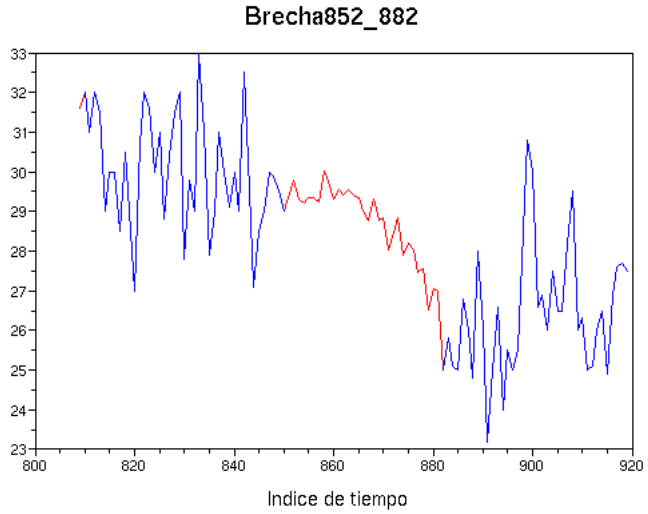


Fig. 12: Valores rellenados (en rojo) en la brecha con índices de tiempo 183-187.

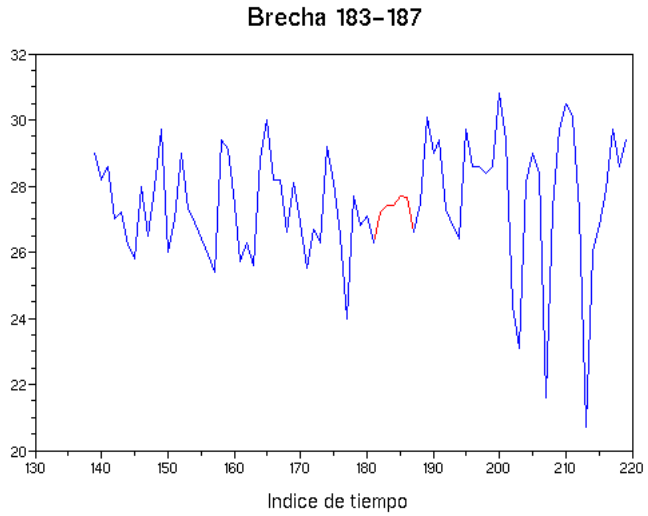


Fig. 13: Valores rellenados (en rojo) en la brecha con índices de tiempo 852-882. Esta brecha excede los tiempos de persistencia y la poca variabilidad de los valores rellenados los diferencia de los reales.

Por curiosidad se repitió el rellenado con el método de Burg para el cálculo de los coeficientes y los resultados coinciden en más de cuatro cifras significativas. Este resultado es notable ya que se había establecido que un modelo AR no era adecuado para la serie en estudio.

4. Método Integrado rellenafull.sci

Como se detalla en las secciones anteriores, cuando se necesita rellenar datos faltantes de una misma variable en series de tiempo geofísicas se puede optar por usar los registros de la misma serie de tiempo utilizando algún método estadístico, por otro lado si se tienen datos de puntos geográficos climatológicamente cercanos se puede completar la información faltante en un periodo de tiempo de una o varias series con la vecindad donde hayan datos en ese periodo. Alfaro y Soley (2009) proponen dos metodologías iterativas para el rellenado de datos geofísicos de una serie de tiempo o de varias series cercanas: El primer método consiste en ajustar un modelo autoregresivo a la serie de tiempo, y el segundo se basó en la descomposición en componentes principales de la matriz de correlación de datos de una misma variable. Estas metodologías han sido exitosamente usadas en proyectos de investigación, en tesis de grado y como material didáctico. El método de componentes principales (c.p.) como primera aproximación de los datos rellenados utiliza el promedio de los datos de las series de tiempo y luego utiliza el algoritmo de componentes principales (Alfaro y Soley, 2009) para hacer iteraciones sobre esta.

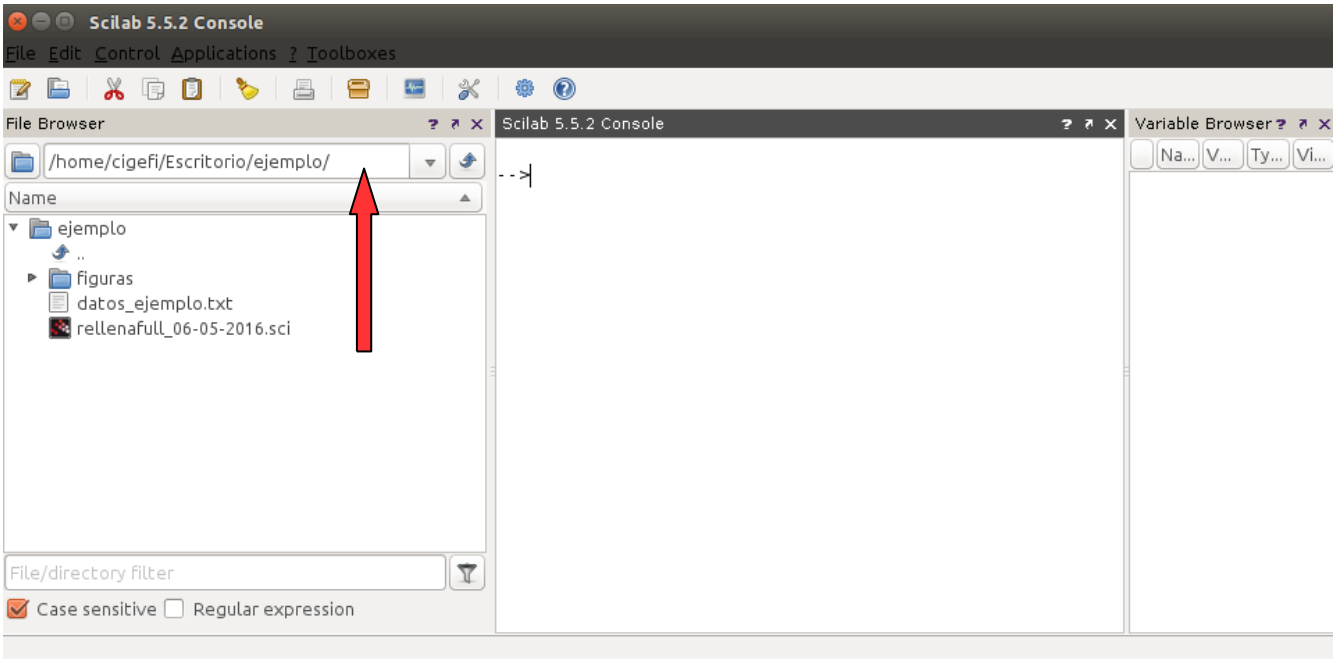
En esta sección se propone el uso conjunto de ambas metodologías, está organizada en dos partes. La Metodología donde se muestra cómo cargar y usar la rutina, también se describe y se justifica los cambios hechos a las rutinas, el software utilizado, las subrutinas editadas y desarrolladas que complementan el experimento, la descripción de los datos utilizados y la preparación del experimento, y los Resultados, donde se muestran gráficos y se discute la comparación entre lo hecho en este artículo y las rutinas originales de Alfaro y Soley (2009).

4.1 Metodología y forma de uso

Las rutinas se programaron en el software de computación numérica de código abierto SCILAB, versión 5.5.0, que se puede obtener junto con la documentación en www.scilab.org (NOTA PARA INSTALAR SCILAB: Ubuntu se encuentra en el Ubuntu Software Center buscando con la palabra SCILAB. Para otras distribuciones de Linux se descarga el paquete, se descomprime y desde la terminal en la carpeta '/scilab-version/bin' se ejecuta el programa ./scilab). Lo que a continuación se describe, es aplicable también para los programas descritos anteriormente.

¿CÓMO CARGAR EL PROGRAMA?

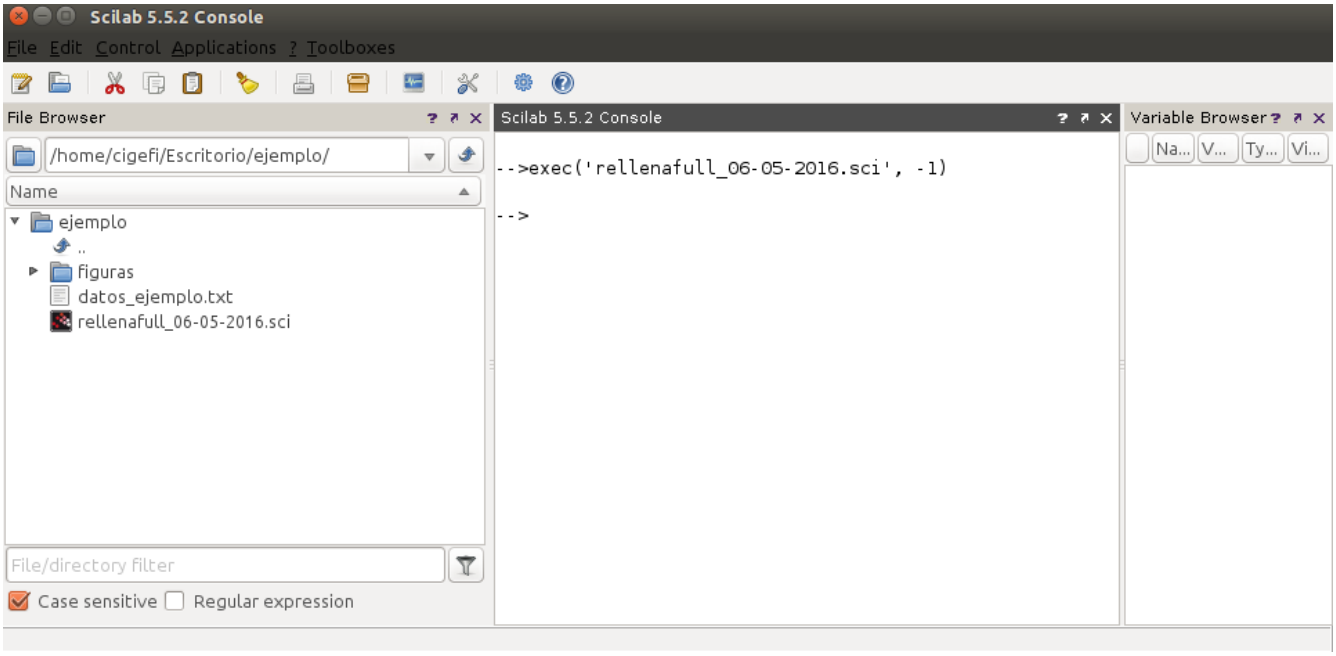
La demostración se hará por medio de capturas de pantalla para ilustrar el proceso. En la ventana de directorios (a la izquierda) hay que ubicar la carpeta que contiene los datos y el programa de rellenado.



Para cargar el programa se escribe en la línea de comandos:

->exec('rellenafull.sci', -1),

se presiona Enter y listo, la función para rellenar los datos "rellenaf( )" que se va a usar más adelante está cargada.



¿Cómo importar los datos a SCILAB?

Hay varias formas de importar los datos. Por línea de comando:

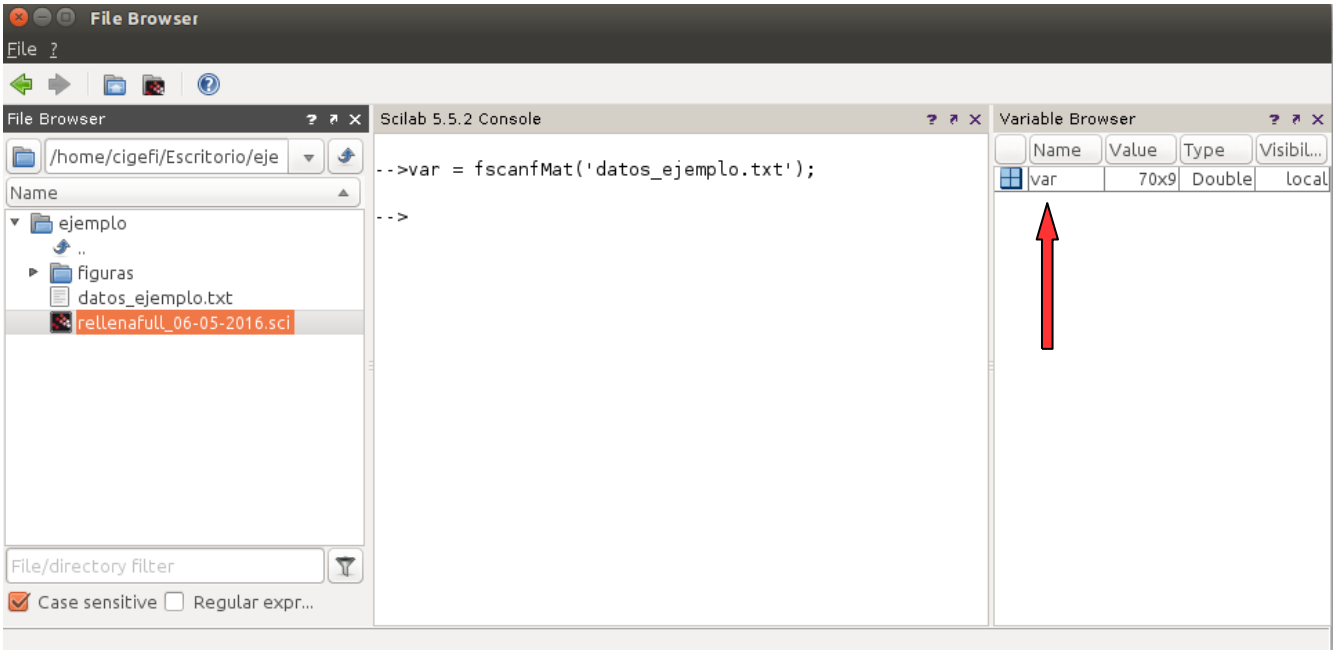
Los datos deben de estar en *.dat, *.txt, ó *.bin, deben estar separados con tabulación, sin fechas o nombres de las columnas, la matriz debe estar completa esto quiere decir que en todas las filas y columnas debe de haber un valor para cada entrada, los datos ausentes o nulos deben estar codificados como Nan,
ejemplo:

10	12	45	Nan
11	13.2	21	78
Nan	85	Nan	5
78	93	Nan	16

Para cargar los datos se usa la función fscanfMat( ), y se escribe el nombre de la variable que se le quiere dar a los datos, en este ejemplo se le va a asignar la variable preci:

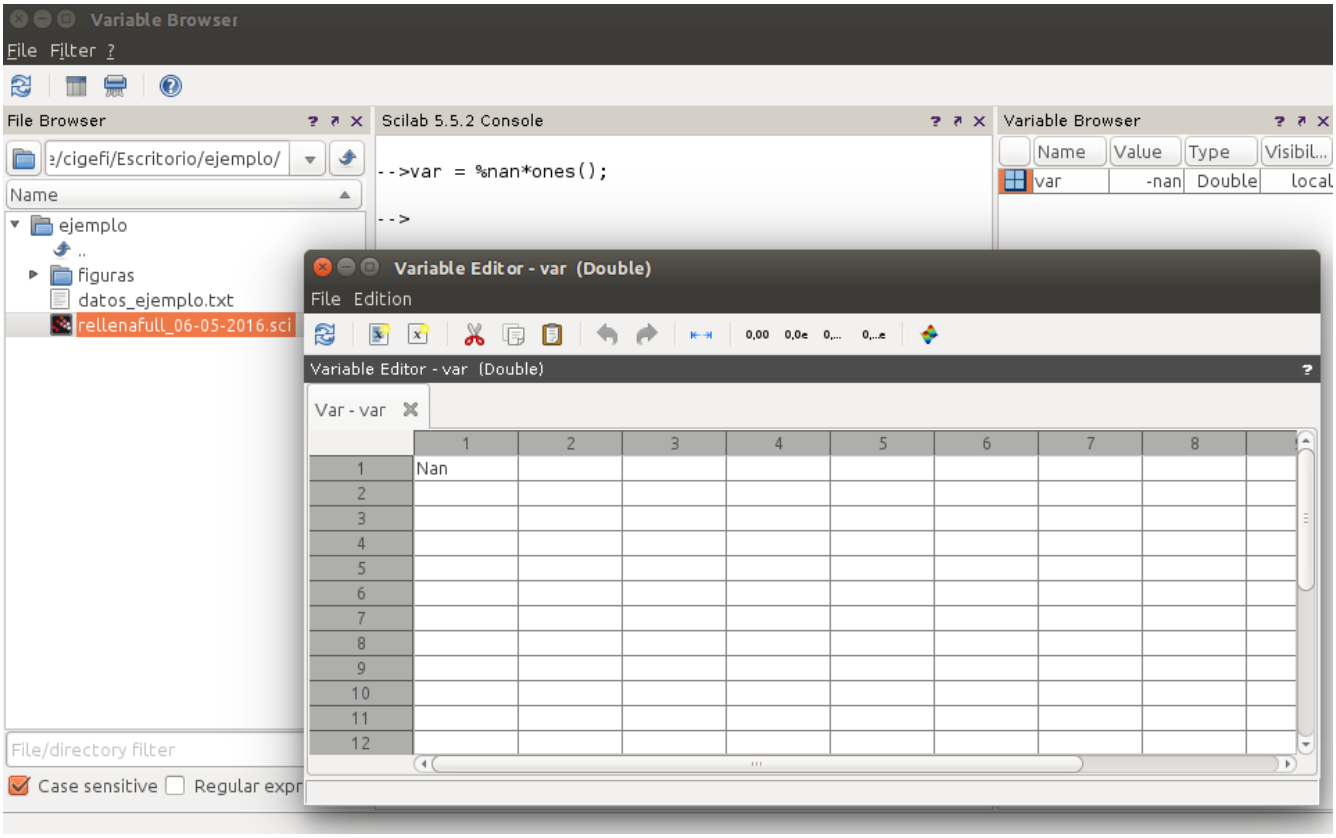
->var = fscanfMat('datos_ejemplo.txt');

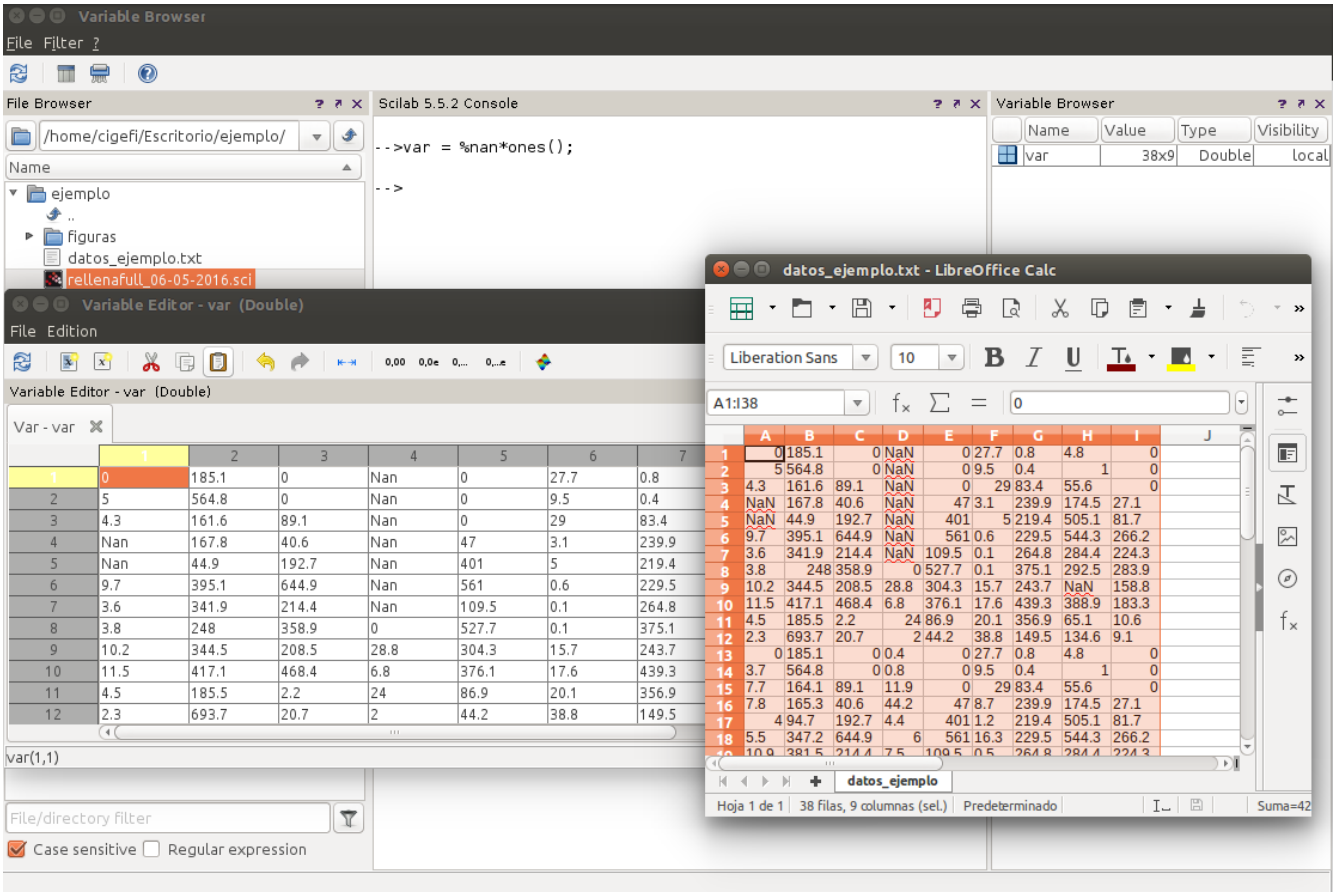
se presiona Enter y deberá aparecer la variable a la derecha en el Variable Browser, donde se indica el nombre, las dimensiones, el tipo de variable, etc, como se observa a continuación:



Por hoja de cálculo:

Se crea una variable ‘dummi’ en scilab con Nan, se abre con el gestor de datos de SCILAB y de la hoja de cálculo se copia y se pegan los valores en la variable:


Variable Browser
File Filter ?
File Browser
:/cigefi/Escritorio/ejemplo/
Name
ejemplo
..
figuras
datos_ejemplo.txt
rellenafull_06-05-2016.sci
File/directory filter
Case sensitive
Regular expr
Scilab 5.5.2 Console
-->var = %nan*ones();
-->
Variable Browser
Name
Value
Type
Visibil...
var
-nan
Double
local
Variable Editor - var (Double)
File Edition
Variable Editor - var (Double)
Var - var
1
2
3
4
5
6
7
8
Nan
9
10
11
12



4.1.1 Descripción del programa de rellenado de datos integrado

Este programa rellena los datos utilizando al inicio como primera aproximación un rellenado por filtro predictivo autoregresivo AR(p) en cada estación (columna) ajustando automáticamente para cada caso el orden del modelo que se va a usar, y luego con la técnica de componentes principales rellena los datos utilizando la información de las demás estaciones.

La función se llama de la siguiente manera

->nuevos = rellenaf(datos, k, nmodos, opcion, difmax, itmax, informe, umbral)

donde:

nuevos : variable donde se van a encontrar los datos rellenados.
datos: matriz de datos (mt,ne) en la cual las columnas son las ne diferentes estaciones de la misma variable y todas de igual longitud mt. Los datos ausentes deben estar codificados con Nan.
k: número de coeficientes de autocorrelación a utilizar para el criterio de selección del orden del modelo AR(p) (BIC se usa por default, se puede cambiar a AIC desde el script, ver líneas 284-285 de la rutina).

- BIC (Bayesian Information Criterion) y AIC (Akaike Information Criterion), son criterios para la selección del órden del modelo AR(p).

nmodos: número de autovalores a incluir en el método de componentes principales. Introducir ‘0’ si se desea observar el gráfico de Scree para determinar el número de autovalores a usar.

opcion: 'BURG' o 'ULCL' se calculan los coeficientes AR con el método de BURG o de Ulrych-Clayton

difmax: diferencia máxima permitida entre iteraciones sucesivas.

itmax: número de iteraciones máximas permitidas

informe: lista ['archivo_salida',[it1 it2 ..])

1) tira de caracteres con nombre de archivo donde guardar el informe
2) arreglo con iteraciones para las cuales guardar el informe p.e. [0 10 20]

0: denota iteración de rellena1

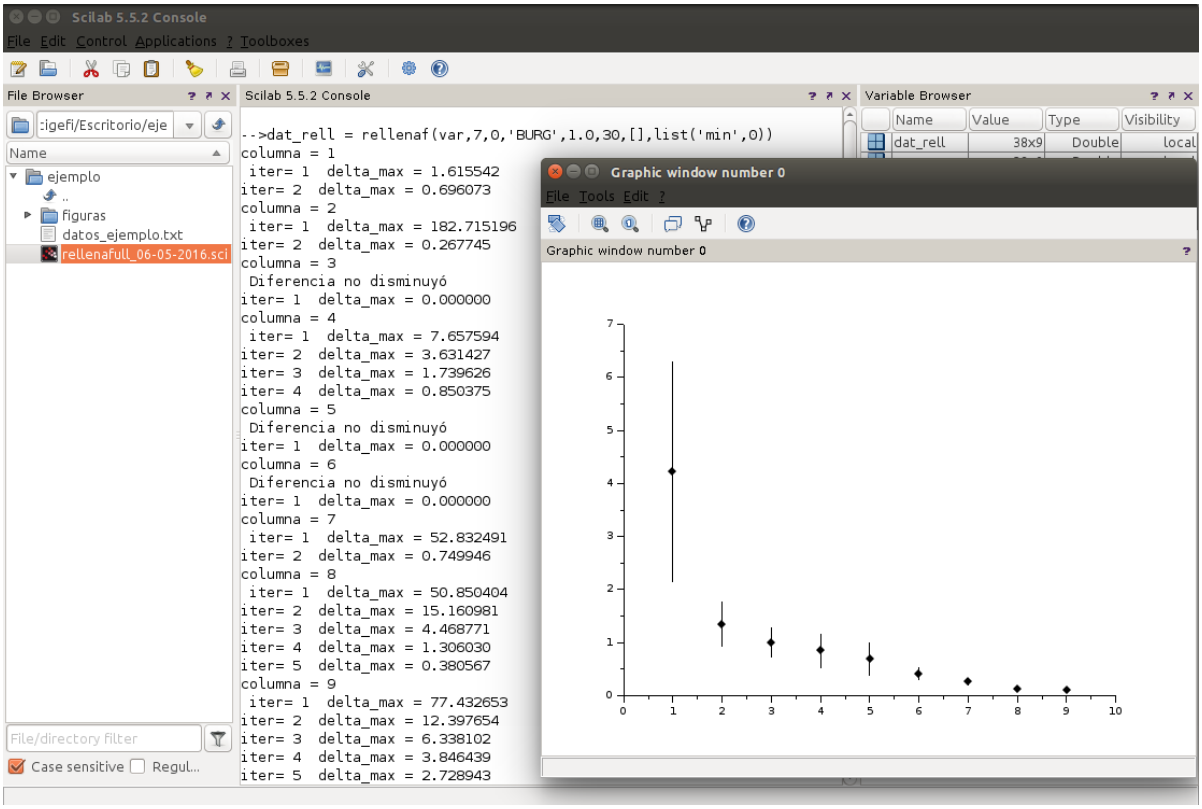
o [ ] si no se quiere guardar (recomendado).

umbral: para delimitar los posibles valores de la variable a rellenar (e.g. para que la precipitación no tome valores menores a cero)’.

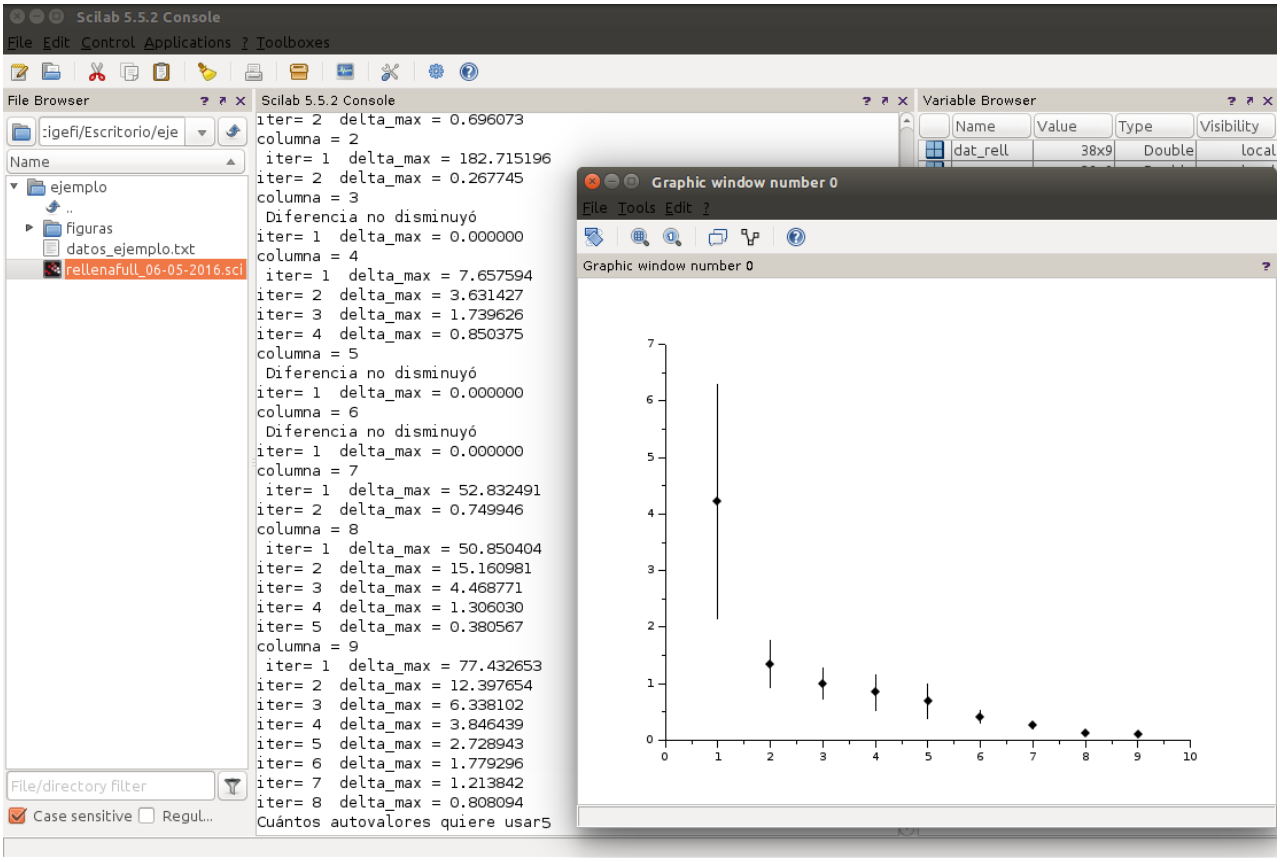
e.g. list('min', 0) --- para que el mínimo sea cero.

4.1.2 Ejemplo de una corrida del programa

->dat_rell = rellenaf(var,7,0,'BURG',1.0,30,[],list('min',0))



En este caso se utiliza nmodos = 0 para que se despliegue el gráfico de scree y con el criterio apropiado seleccionar el número de modos.
Luego el programa pregunta cuantos autovalores se quieren utilizar para las componentes principales, a modo de ejemplo se utilizaran 5.



Al final se puede abrir la variable de los datos ya rellenados dat_rell desde el Variable Browser y se puede exportar en diferentes formatos.

Nota: En la carpeta de trabajo se encontrará un archivo llamado “salida.sod” que es la variable de datos rellenados, esta se puede abrir desde el File Browser de SCILAB.

Recordatorio: El criterio sobre las estaciones que se usan y los parámetros de la función para rellenar la información ausente es subjetivo y se debe sustentar en algún juicio de experto.

4.2 Rutinas para la preparación y evaluación de los datos

Basado en rutinas utilizadas por Alfaro y Soley (2009), se reeditaron éstas para que sea compatible ya no con una serie de datos, si no con varias series de datos de la misma variable, al final estas rutinas de evaluación quedan de la siguiente manera:

genhuecos_hyp: Esta rutina genera los datos faltantes de manera aleatoria para cada entrada de un vector correspondiente al porcentaje de datos faltantes. Se agregan y se editan comandos para generar una

hipermatriz boleana donde las entradas "true" son los datos que deben tomarse como nulos, se utiliza un algoritmo para generar números aleatorios uniformemente en toda la matriz, esta rutina conserva el lugar de las entradas que van a ser codificadas como Nan cuando se incrementa el valor porcentual de datos faltantes para la misma matriz.

contajuar_ar_cp: Esta rutina se edita para que genere hipermatrices de error cuadrático medio y absoluto medio comparando los datos completos con los rellenados con ambos métodos por aparte, para cada porcentaje de datos faltante y para cada componente principal, utiliza la salida de genhuecos_hyp, los datos originales, la rutina de componentes principales propuesta por Alfaro y Soley (2009), y la rutina de componentes principales propuesta en este manual.

correcontar_ar_cp: Esta es la rutina "madre" que controla las dos anteriores y repite las realizaciones del proceso n veces, se edita para que guarde cada realización en un índice de una hipermatriz.

esthiper: de las hiérmatrices con n realizaciones, calcula el promedio y las variancias del error absoluto medio y cuadrático medio y los guarda en matrices de la forma (porcentaje de datos ausentes x componentes principales).

grid_mat: Esta rutina se crea para preparar la salida de esthiper y poder graficar las estadísticas con las funciones de contorno de SCILAB de la forma (x, y, f(xy))

contourerr: Se crea para hacer un gráficos de contorno donde el eje de las abscisas es porcentaje de dato faltante, el eje de las ordenadas corresponde a cada componente principal, y una paleta que muestra el error del rellenado respecto a la serie de datos completa, esta paleta numérica se conserva con los mismos valores cuando se tiene el mismo error para el rellenado de c.p. utilizado por Alfaro y Soley (2009), y la versión propuesta en este artículo, para un mismo estadístico se mantiene la misma paleta.

4.3 Descripción del rellenado por filtro autoregresivo AR(m) en la parte integrada.

En la rutina propuesta por Alfaro y Soley (2009) el orden m del filtro autoregresivo es un parámetro dado por el usuario, esto puede llevar a la divergencia del método si se utiliza un número de coeficientes autoregresivos incongruente con la persistencia de la serie (Wilks, 2011). El orden de selección de la autoregresión de esta rutina debe ser correctamente escogido ya que agregar más complejidad al modelo autoregresivo no mejora su representación de los datos, es decir puede haber un sobreajuste del modelo para la estimación de los parámetros si se escoge un orden mayor al que se debería. Para este trabajo se utilizan dos criterios de selección para la escogencia del orden del modelo, los estadísticos Bayesian Information Criterion (BIC) y Akaike Information Criterion (AIC) que envuelven una función de verosimilitud logarítmica (log-likehood) más una penalización por el número de parámetros, ambos difieren de la forma de la función de penalización. Estos estadísticos se calculan para cada candidato orden m:

$$BIC(m) = nln\left[\frac{n}{n-m-1}s_\varepsilon^2(m)\right] + (m+1)ln(n) \qquad (1)$$

$$AIC(m) = nln\left[\frac{n}{n-m-1}s_\varepsilon^2(m)\right] + 2(m+1) \qquad (2)$$

El $s^2_\varepsilon(m)$ es la varianza de ruido blanco (white-noise variance). El estadístico BIC generalmente es

preferible para series de tiempo suficientemente largas (Wilks, 2011).

El orden m que se escoge como apropiado es el que minimiza cualquiera de las ecuaciones (1) y (2) para m coeficientes de autocorrelación calculados. La cantidad de coeficientes de autocorrelación a calcular es el parámetro que ahora el usuario de la rutina debe de ingresar. El algoritmo para este cálculo se agregó a la rutina que rellena los datos por medio del filtro autoregresivo AR(m).

4.4 Rellenado por el método de componentes principales en la parte integrada.

En la rutina propuesta por Alfaro y Soley (2009) en la primera aproximación para calcular los componentes principales, los datos ausentes se sustituyen por el promedio de cada serie, esta primera aproximación es grosera y no toma en cuenta la tendencia de la serie, se propone en este artículo que la primera aproximación para los componentes principales sea un rellenado de cada serie de tiempo con el método de filtro autoregresivo, así la primera aproximación para empezar a iterar con los componentes principales se hace sobre datos rellenados que toman en cuenta la información de la misma serie y esto capta una señal más aproximada a la realidad de los datos rellenados.

En esta primera aproximación se rellenan los datos por cada estación o serie que conforme la totalidad los datos, para cada estación la rutina hace los cálculos del orden del filtro AR(m).

4.5 Ejemplo de comparación del método integrado con el de Componentes Principales previo.

A modo de ejemplo para obtener una comparación con la rutina original propuesta con Alfaro y Soley (2009) se calcula el error cuadrático medio, el error absoluto medio y la varianza de los datos rellenados con una serie completa de datos de 9 puntos de rejilla, del promedio mensual de precipitación para 1950 - 1999, rejilla de 0.5° por 0.5° de la vertiente Caribe costarricense del Tropical Land-Surface Precipitation: Gridded Monthly and Annual Climatologies de Johnson et al. (2003) (Figura 14).

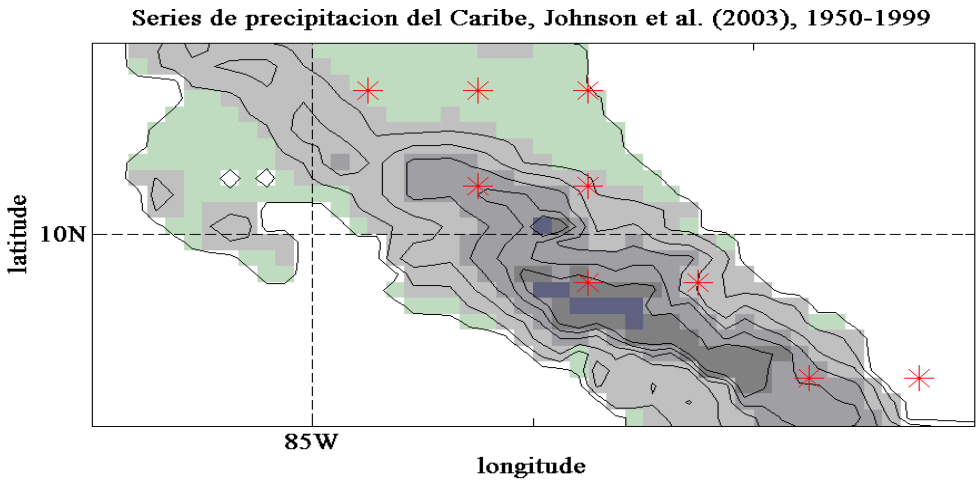


Fig. 14. Los asteriscos rojos muestran la ubicación de los 9 puntos de rejilla del Caribe costarricense de Johnson et al. (2003).

Se generaron valores de datos ausentes (Nan) aleatoriamente en la serie completa de 600x9 datos de precipitación mensual descrita anteriormente, con un rango de 0.5% al 30% con intervalos de 0.5%, para cada porcentaje se rellenó la serie con la rutina original y la nueva utilizando de 1 a 9 componentes principales, para el filtro autoregresivo la rutina escogió el orden del filtro automáticamente para el rango del porcentaje de datos faltantes, luego de cada iteración se calculó el error cuadrático medio, el error absoluto medio y la varianza del promedio de ambos errores, este experimento se repitió 100 veces, y luego se promediaron los estadísticos de todas las repeticiones. Cuando finalizaron todas las iteraciones y los cálculos de los estadísticos se hizo un gráfico de contornos para observar el comportamiento de ambas rutinas de rellenado.

El promedio de las 100 realizaciones del error absoluto medio para esta serie de datos muestra un ensanchamiento donde el error disminuye entre el primer y el cuarto componente principal en la rutina que incorpora el rellenado AR con C.P. El error es mínimo en ambos para el segundo componente principal como se muestra en el gráfico de scree de la Figura 15. Para ambas rutinas de rellenado de datos el error mínimo ronda al rededor del 68% (Figura 16).

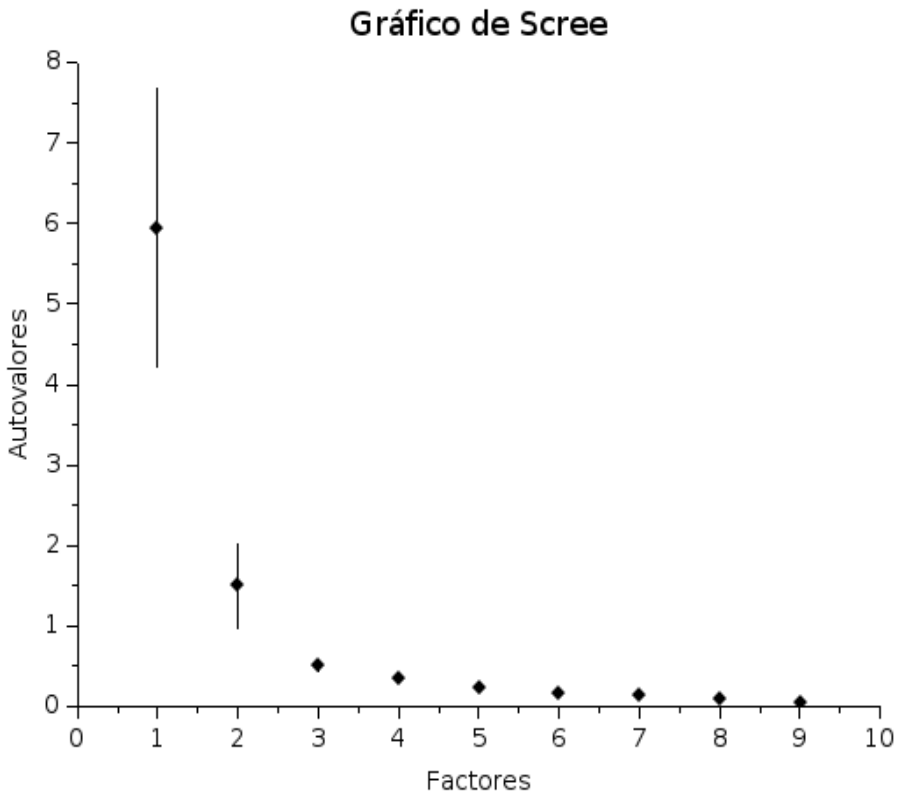
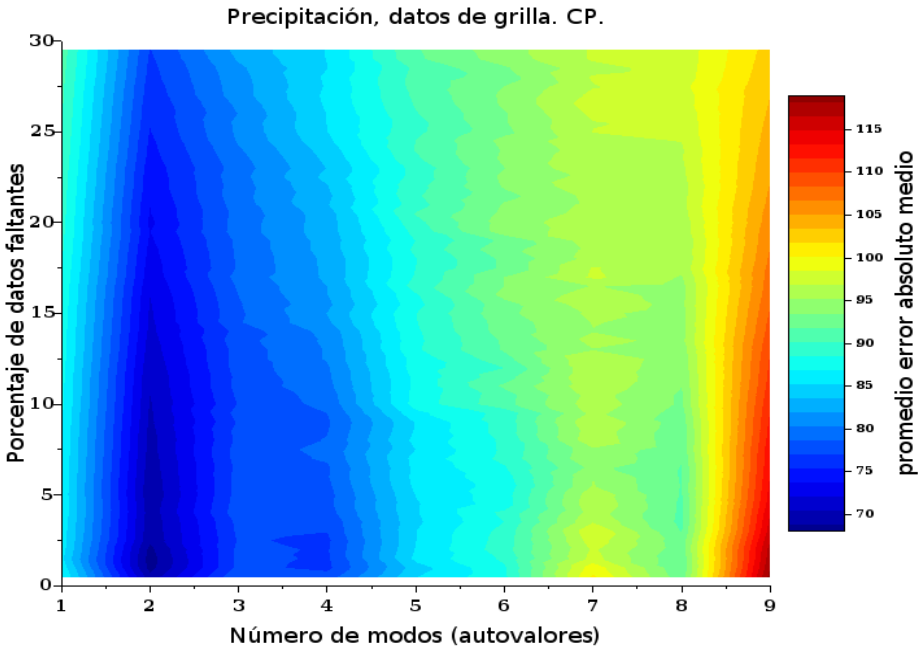


Fig. 15. Gráfico de Scree de las series de tiempo utilizadas. Observe que los primeros dos componentes son los que contribuyen más a la varianza total.

a)



b)

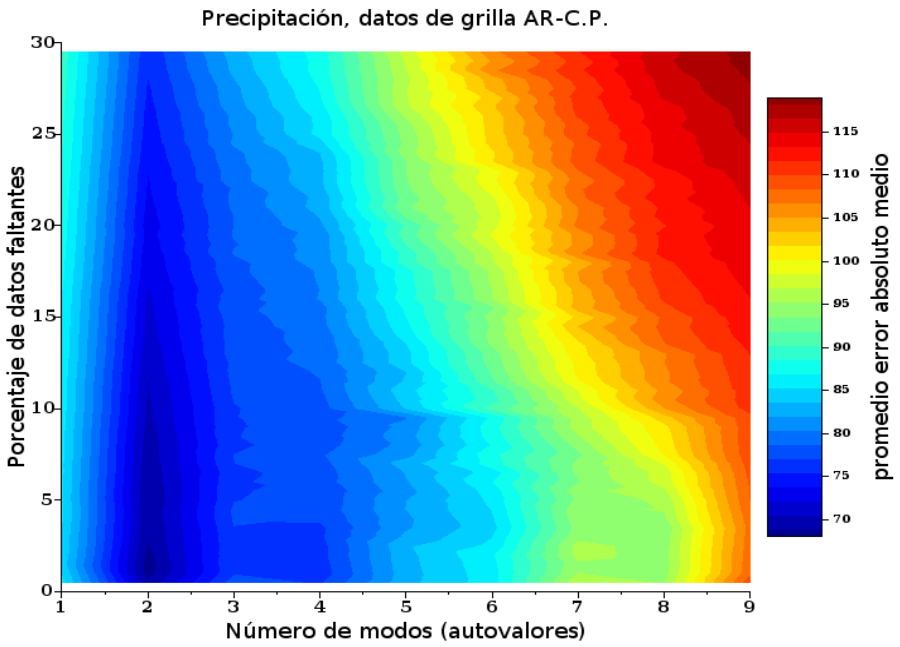
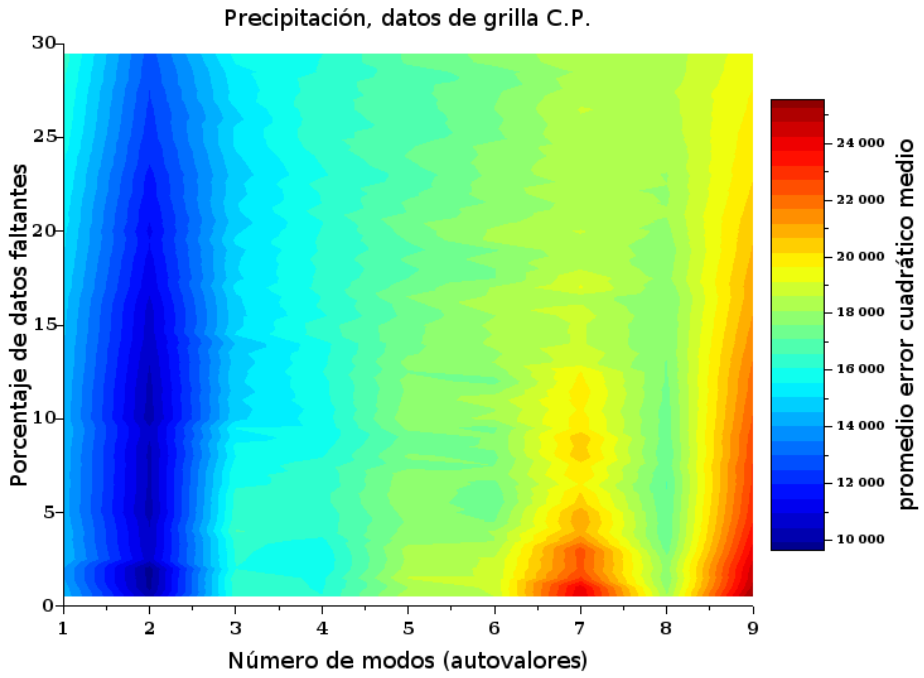


Fig. 16. Promedio del error absoluto medio para 100 realizaciones del rellenado de datos por componentes principales (a), y el rellenado por componentes principales con el filtro autoregresivo AR(m) (b).

Para un porcentaje entre el 25% y el 30% y con dos componentes principales de datos faltantes se observa que la rutina que incorpora el rellenado AR con C.P. generó ligeramente menor promedio del error absoluto medio para este ejemplo en particular, sin embargo se observa que esta rutina genera un error ligeramente mayor en los últimos componentes principales con un porcentaje de datos faltante mayor al 10% para esta serie de datos.

El cálculo del error cuadrático medio (Figura 17) muestra comportamiento similar al error absoluto medio donde el segundo componente principal minimiza el error, además se destaca que para un porcentaje de datos faltantes entre 0 y 15% los resultados son similares en ambas rutinas, sin embargo para porcentajes mayores al 15 % la rutina que incorpora el rellenado AR con C.P. genera menor error respecto a la serie de datos originales.

a)



b)

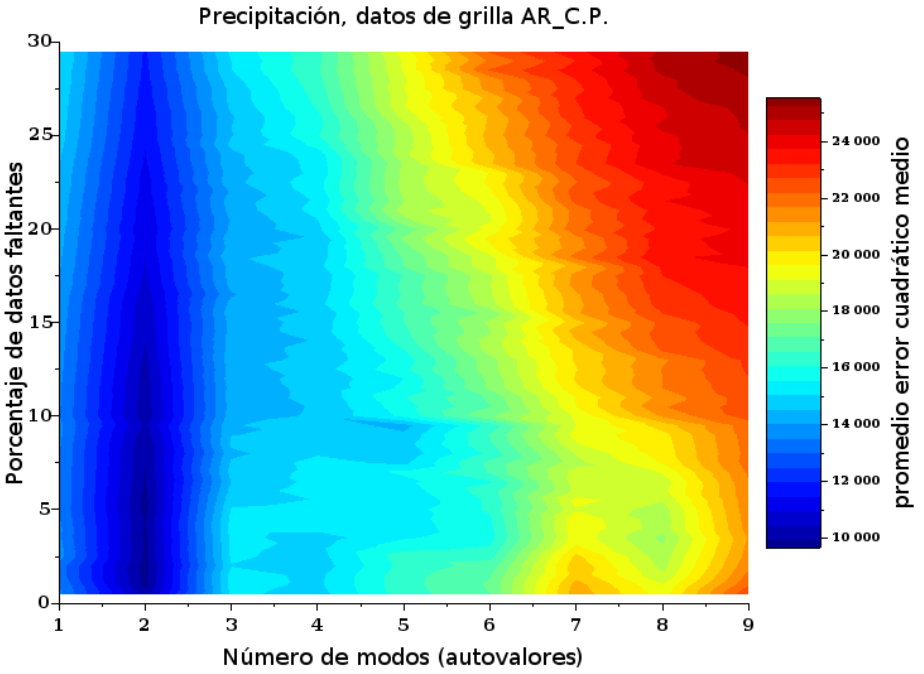
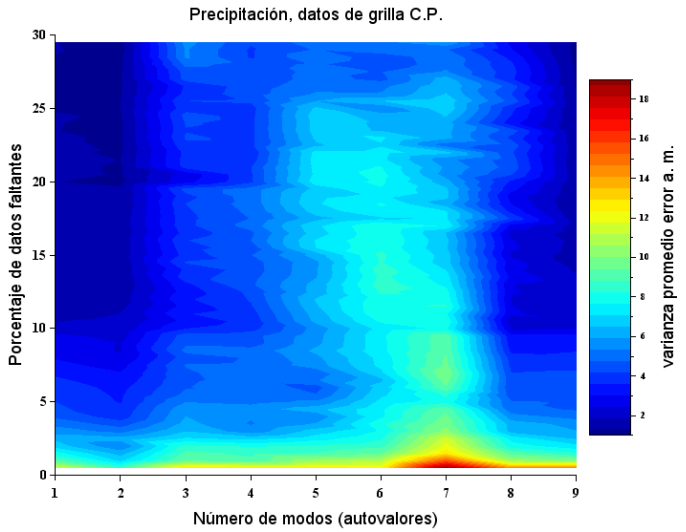


Fig. 17. Promedio del error cuadrático medio para 100 realizaciones del rellenado de datos por componentes principales (a), y el rellenado de datos por componentes principales con el filtro autoregresivo AR(m) (b).

El cálculo de la varianza para los dos casos muestra que para el error absoluto medio en el segundo componente principal ambas rutinas se comportan similar, utilizando el séptimo componente principal la rutina que incorpora el rellenado AR con C.P. se observa un incremento en la varianza respecto a la rutina original (Figura 18).

a)



b)

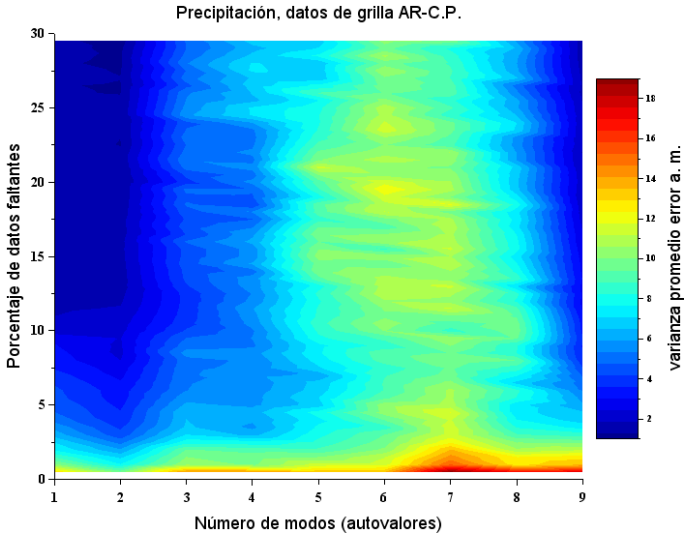


Fig. 18. Varianza del promedio del error cuadrático medio para 100 realizaciones del rellenado de datos por componentes principales (a), y del rellenado de datos por componentes principales con el filtro autoregresivo AR(m) (b).

Para el error cuadrático medio (Figura 19) la varianza es mucho menor en la rutina que incorpora el rellenado AR con C.P. para el segundo componente principal y datos faltantes menores al 10%, después del 10% de datos faltantes la varianza es similar en ambas rutinas, también se observa un aumento en la varianza alrededor del séptimo componente principal comparado con la rutina original.

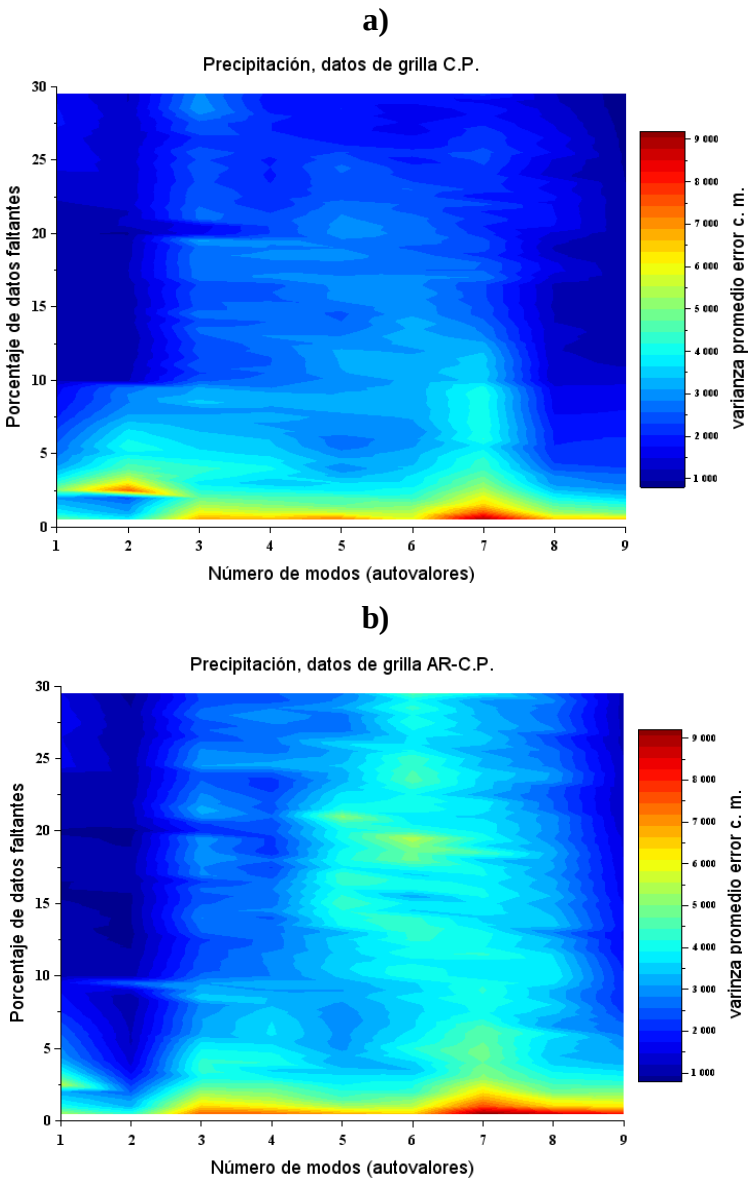


Fig. 19. Varianza del promedio del error cuadrático medio para 100 realizaciones del rellenado de datos por componentes principales (a), y del rellenado de datos por componentes principales con el filtro autoregresivo AR(m) (b).

El análisis de los resultados anteriores comprueba que para este experimento en el componente principal n = 2 la rutina integrada generó menos error. Se están dando sugerencias estadísticas para la evaluación, pero cada experimento con una serie de datos en particular arrojaría resultados distintos. El criterio sobre

las estaciones que se usan para rellenar la información ausente es subjetivo y se debe sustentar en algún juicio de experto. No sólo se lograron usar en forma conjunta las rutinas, lo que ahorra tiempo al usuario y facilita su uso, sino que también se actualizaron a la última versión de SCILAB.

El problema de datos faltantes es persistente, por lo que esto se puede usar en los foros de predicción climática de América Central del Sistema de Integración Centroamericana (SICA) (Donoso y Ramírez, 2000; García-Solera y Ramírez, 2012).

5. Agradecimientos

Se agradece a los proyectos VI y VAS-UCR 805-B6-143, 808-B5-298, 805-B4-227, 805-B0-065, 805-A9-532, 805-A1-715, 808-B5-298 y ED-1977.

6. Referencias

- Alfaro, E. J. y F. J. Soley, 2009: Descripción de Dos Métodos de Rellenado de Datos Ausentes en Series de Tiempo Meteorológicas. *Revista de Matemática: Teoría y Aplicaciones*, **16**(1); 60-75.

- Bretherton C. S., C. Smith y J. M. Wallace, 1992: An Intercomparison of Methods of finding Coupled Patterns in Climate Data. *J. Climate*, **5**, 541-560.

- Donoso, M., & Ramírez, P. (2000, October). Latin America and the Caribbean. Report on the Climate Outlook Forums for Mesoamerica. In Coping with the climate: A step Forward, Workshop Report. A multi-stakeholder review of Regional Climate Outlook Forums, Pretoria, South Africa. (pp. 16-20).

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

- García-Solera, I., & Ramírez, P. (2012). Central American Seasonal Climate Outlook Forum. The Climate Services Partnership, 8 pp. http://www.climate-services.org/case-studies/central-american-climate-outlook-forum/ (visited 06/09/2016).

- Johnson, M., K. Matsuura, C. Willmott and P. Zimmermann, 2003. Tropical Land-Surface Precipitation: Gridded Monthly and Annual Climatologies. Disponible en http://climate.geog.udel.edu/~climate/html_pages/Tropics_files/README.tropic_precip_clim.html

- Lorenz, E. N., 1956: Empirical Orthogonal Functions and Statistical Weather Prediction. Sci. Rep.1. Statistical Forecasting Project. Department of Metorology, MIT (NTIS AD 110268), 49 pp.

- Magaña, V., J. A. Amador & S. Medina. 1999. The Mid-Summer Drought over Mexico and Central America. *J. Climate*. **12,** 1577-1588.

- North, G. R., T. L. Bell, R. F. Cahalan, y F. J. Moeng, 1982: Sampling Errors in the Estimation of Empirical Orthogonal Functions. *Mon. Wea. Rev*., **110**, 699-706.

- Sciremammano, F., 1979: A suggestion for the presentation of correlations and their significance levels. *J. Phys. Oceanogr*., **9**, 1273-1276.

- 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.

- Soley, F. J., 2005: Sistemas lineales ARMM(p,q) con p+q ≤ 4. Primera Parte: Sistemas lineales AR(p ≤ 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.

- Tabony, R. C., 1983. The Estimation of Missing Climatological Data. *Journal of Climatology*, **3**, 297-314 .

- Ulrych T. J. y T. N. Bishop, 1975. Maximum Spectral Analysis and Autoregressive Decomposition. *Reviews of Geophysics and Space Physics*, **13(**1), 183-200.

- Ulrych T. J. y R. W. Clayton, 1976. Time Series Modeling and Maximum Entropy. *Physics of the Earth and Planetary Interiors*, **12**, 188-200.

- Wilks, D., 2011: Statistical Methods in the Atmospheric Sciences. 3ra. ed. Academic Press. 676pp.

Nota de advertencia: Debido a que este software es distribuido libre de cargo, no se ofrece garantía de ningún tipo, ni explicita ni implícitamente. Los autores no se hacen responsables por el uso del mismo. Sin embargo están en la mejor disposición de contestar cualquier consulta que tenga el usuario sobre el material presentado, así como el considerar las sugerencias que se deseen hacer sobre el mismo, por lo que agradecemos su contacto con nosotros.