

AUTOMATIZACIÓN DE LA HOMOGENEIZACIÓN DE SERIES CLIMÁTICAS: NUEVAS FUNCIONES DEL PAQUETE CLIMATOL 3.0

José Antonio GUIJARRO

Agencia Estatal de Meteorología.

jguijarrop@aemet.es

RESUMEN

La homogeneización de las series climáticas, para eliminar las perturbaciones que frecuentemente contienen por cambios en las condiciones de observación, resulta obligatoria para aumentar la fiabilidad de los estudios de variabilidad climática derivados de su análisis. Pero esta es una tarea muy laboriosa si no se realiza con la ayuda de paquetes de software que liberen al investigador de la repetitiva manipulación y comparación de largas series de datos. El paquete *Climatol*, programado en *R*, es una de las herramientas disponibles para facilitar estas tareas, y en su nueva versión 3.0, a la función de homogeneización automática *homogen* ya disponible en la anterior versión 2.2, añade nuevas funcionalidades para facilitar tanto el proceso de la homogeneización como la explotación de las series homogeneizadas. En esta comunicación se detalla la aplicabilidad de estas funciones y se presentan ejemplos sobre su uso.

Palabras clave: software, homogeneización de series climáticas, Climatol.

ABSTRACT

Homogenization of climatic series to eliminate disturbances often contained in them due to changes in the conditions of observation is a mandatory process to increase the reliability of studies of climate variability derived from their analysis. But this is a very laborious task if not done with the help of software packages that relieve the researcher from repetitive processing and comparison of long series of data. The *Climatol* package, programmed in *R*, is one of the available tools that facilitate these tasks, and in its new version 3.0, to the automatic homogenization function *homogen* already available in the previous version 2.2, new features are added to facilitate both the homogenization process as the exploitation of the homogenized series. In this communication the applicability of these functions is detailed, with examples of their use.

Key words: comparison, methodologies, homogenization of climatic series.

1. INTRODUCCIÓN

Como es sabido, cambios de emplazamiento de los observatorios meteorológicos (aunque sean pequeños), cambios en la instrumentación o en los abrigos destinados a proteger los sensores de la radiación solar, o incluso cambios en el entorno (usos del suelo, nuevas construcciones, etc) introducen en las series climatológicas alteraciones que no cabe achacar a cambios en el clima. Por tanto es necesario proceder a eliminar

estas alteraciones antes de caracterizar la variabilidad climática a partir de esas series, proceso denominado *homogeneización*.

Hay una extensa literatura científica dedicada a los distintos métodos que pueden usarse para ese fin, como recopilan Peterson *et al.* (1998) y Aguilar *et al.* (2003). En cualquier caso, la homogeneización de series se convierte en un procedimiento laborioso y tedioso cuando su número supera una o dos docenas, por lo que resulta muy conveniente usar algunas de las implementaciones existentes en forma de paquetes de software públicamente disponibles, cuya lista, junto con unas tablas comparativas que resumen sus características, se pueden encontrar en la página web <http://www.climatol.eu/tt-hom/>

Uno de ellos, desarrollado por este autor en lenguaje R (R Development Core Team, 2013), es el paquete Climatol (Guijarro, 2016), cuya anterior versión (2.2) ya contaba con la función *homogen* para automatizar la homogeneización de un conjunto de series limitado únicamente por la capacidad del ordenador, así como con la función *dahstat* para obtener diversos estadísticos (medias, tendencias, percentiles, etc) de las series mensuales homogeneizadas.

La nueva versión 3.0 añade nuevas funcionalidades, que se detallan en los siguientes apartados, finalizando con algunos ejemplos de aplicación.

2. NUEVAS FUNCIONALIDADES

2.1. db2dat

Una parte importante del tiempo necesario para realizar una aplicación de Climatol u otros paquetes de homogeneización a un caso real hay que dedicarlo a preparar los datos de entrada en el formato requerido por el programa a utilizar. La función *db2dat* viene a facilitar este proceso, pues puede acceder a bases de datos a través de un controlador ODBC, y generar los ficheros de entrada para Climatol, *.dat y *.est

2.2. dd2m y m2dd

Una de las estrategias más seguidas para la homogeneización de datos diarios es la de ajustarlos mediante las correcciones calculadas mediante una homogeneización a nivel mensual. Para ello hay que empezar con agregar los datos diarios para formar las series mensuales, lo que se lleva a cabo fácilmente mediante la función *dd2m*.

A continuación se pasaría a homogeneizar las series mensuales, y posteriormente los datos diarios se ajustarían por interpolación de las correcciones mensuales mediante la segunda función, *m2dd*.

2.3. homogen

Esta función ya existía en la versión anterior, pero se han añadido nuevos parámetros, como poder especificar independientemente los umbrales superior e inferior de rechazo de datos anómalos, así como los valores máximo y mínimo admisibles (100 en humedad relativa, 0 en muchas variables, etc). Hay que tener en cuenta que algunos parámetros de esta función han sufrido cambios de nombre para que resulten más fáciles de recordar, como los antiguos umbrales SNHT para cortar las

series en las fases 1 (por ventanas solapadas) y 2 (sobre las series completas), que antes se llamaban *tVt* y *snhtt*, y ahora pasan a denominarse *snht1* y *snht2* respectivamente.

Además, cuando la distribución de frecuencias de la variable a homogeneizar es muy sesgada, a las transformaciones raíz que se ofrecían antes se añade también la transformación logarítmica $\log(x+1)$.

2.4. homogsplitt

Cuando el número de series a tratar es muy elevado (varios miles), puede suceder que lleguemos a superar las capacidades de memoria de nuestro ordenador, o bien el tiempo de proceso se alargará durante muchos días. Esto se puede solventar mediante esta nueva función, que subdivide el dominio espacial a tratar en áreas rectangulares solapadas, de modo que el número de series se reduzca a unos cientos al tiempo que el solapamiento actúe como zona tampón para minimizar el riesgo de que aparezcan fronteras espurias en las divisorias de cada área.

2.5. dahstat

Esta función de postproceso añade a las funcionalidades anteriores el cómputo de tendencias por regresión lineal con el tiempo (OLS) y sus p-valores, así como la posibilidad de extraer todas las series homogeneizadas a ficheros CSV individuales.

2.6. dahgrid

Esta es otra función de postproceso pensada para generar datos homogeneizados y normalizados interpolados sobre una rejilla definida por el usuario, que se graban en un fichero con formato NetCDF. La normalización es la que se haya elegido al realizar la homogeneización mediante el parámetro *std*, que puede adoptar los valores 1 (restar a cada término de cada serie su valor medio), 2 (dividir por su valor medio) o 3 (restar su valor medio y dividir por su desviación típica, que es la opción por defecto).

El fichero generado contiene también las interpolaciones de los valores medios (y de las desviaciones típicas si *std*=3), de modo que pasar de capas de valores normalizados a valores absolutos es inmediato. No obstante, si se desea generar mapas de calidad a partir de este fichero, lo más conveniente es suministrar rejillas de alta resolución de las medias (y desviaciones típicas en su caso) obtenidas con algún método geoestadístico que incluya el efecto de factores externos como la orografía, puesto que la interpolación que realiza esta función no puede tener en cuenta esos factores.

3. EJEMPLO DE APLICACIÓN

Este ejemplo se va a basar en un supuesto práctico en el que vamos a partir de un conjunto de series de datos diarios observados, para homogeneizarlos y obtener distintos índices y mapas.

3.1. Datos de partida

Supongamos que queremos hacer un estudio de las características de la precipitación diaria de una determinada zona, y nuestros datos están en una base de datos

accesible mediante el protocolo ODBC. Para preparar los datos de entrada mediante la función *db2dat* necesitaremos saber tanto el nombre de la base de datos como los de las tablas y campos en los que están almacenados los datos diarios de precipitación y sus fechas correspondientes, así como los de las coordenadas, códigos y nombres de las estaciones pluviométricas, que también vamos a necesitar. Estos nombres podrían ser:

- *mcheng*: Nombre de la base de datos (con usuario *USER* y clave *PASS*).
- *Precipitation*: Nombre de la tabla que almacena las precipitaciones diarias.
- *Date*: Nombre del campo que contiene las fechas.
- *Value*: Nombre del campo que contiene las precipitaciones diarias.
- *Stations*: Nombre de la tabla que almacena los datos de las estaciones.
- *Code*: Nombre del campo con los códigos de las estaciones. (Supondremos que este campo también está en la tabla de las precipitaciones).
- *Name*: Nombre del campo con los nombres de las estaciones.
- *Latitude*: Nombre del campo con las latitudes (en grados, con decimales).
- *Longitude*: Nombre del campo con las longitudes (en grados, con decimales).
- *Elevation*: Nombre del campo con las altitudes (en metros).

Conociendo estos datos, elegimos *Prec* como nombre de nuestra variable a estudiar, y el periodo 2001-2010, y procedemos del siguiente modo (*minny*=1 es para rechazar series con menos de 1 año de datos, *dformat* especifica el formato de las fechas, y todo lo que sigue a # son comentarios):

```
R #arrancamos R (versión 3 o mayor)
library(RODBC)
library(climatology) #o bien, "source('depurdat.R')
ch <- odbcConnect("mcheng",uid="USER",pwd="PASS")
db2dat('Prec',2001,2010,minny=1,ch,dformat='%Y-%m-%d',
'Precipitation','Code','Date','Value','Stations',
'Code','Name','Longitude','Latitude','Elevation')
odbcClose(ch) #cerrar la conexión con la base de datos
```

Esto generaría los ficheros *Prec-d_2001-2010.est* y *Prec-d_2001-2010.dat* conteniendo el primero las coordenadas, códigos y nombres de las estaciones, y el segundo los datos diarios de precipitación de los diez años en la primera de las estaciones, seguidos por los demás, estación por estación. En los días en los que no se disponga de datos figurará NA (Not Available), que es el código de ausencia de dato estándar en R.

3.2. Homogeneización de los datos diarios.

Con los ficheros obtenidos en el paso anterior ya podríamos proceder a su homogeneización, pero dada la elevada variabilidad de los datos diarios, conviene hacer una homogeneización exploratoria (especificando *expl*=TRUE) para ajustar los umbrales de detección. En la siguiente orden, *nm*=0 indica que son datos diarios, *std*=2 que se hará una normalización por proporciones, e *ini*='2001-01-01' indica la fecha inicial (únicamente a efectos de rotular el eje temporal en los gráficos):

```
homogen('Prec',2001,2010,nm=0,std=2,expl=TRUE,
ini='2001-01-01')
```

La salida gráfica de este proceso la encontramos en el documento `Prec-d_2001-2010.pdf`, donde podemos ver (Figura 1) que podemos mantener el valor por defecto de `snht1=25` para el umbral de corrección de los saltos en las series, y adoptar un umbral de rechazo de datos anómalos de 25 desviaciones típicas. Por tanto ahora ejecutaremos:

```
homogen('Prec', 2001, 2010, nm=0, std=2, dz.max=25,
        ini='2001-01-01')
```

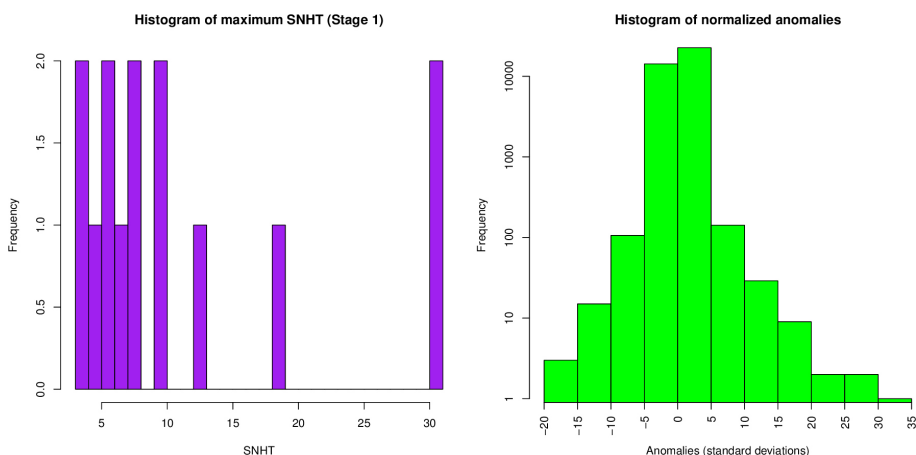


Fig. 1: Valores SNHT de las series (izquierda) y anomalías de los datos (derecha).

Este proceso ha corregido 4 saltos en la media, y ha rechazado un dato anómalo. Pero, dada la baja relación señal/ruido de las series diarias, sobre todo de precipitación y en clima mediterráneo, la detección de saltos es mucho más difícil en ellas que en series mensuales, de modo que en el siguiente apartado vamos a seguir el procedimiento recomendado, consistente en realizar una homogeneización a escala mensual y luego ajustar los datos diarios mediante interpolación de las correcciones mensuales.

3.3. Homogeneización a nivel mensual y ajuste diario

Lo primero que tenemos que hacer es agregar los datos diarios en valores mensuales, para lo que usaremos la función `dd2m` (`valm=1` indica que deben acumularse los datos de cada mes, `hom=FALSE` que no queremos agregar datos previamente homogeneizados, y `namax=15` hará que el dato mensual se considere ausente si faltan más de 15 datos diarios):

```
dd2m('Prec', 2001, 2010, valm=1, hom=FALSE, namax=15)
```

Obtenemos así el fichero de precipitaciones mensuales `Prec-m_2001-2010.dat` y su correspondiente de estaciones `Prec-m_2001-2010.est` (que será una copia del diario), de modo que ya podemos proceder a efectuar la homogeneización mensual. Una primera pasada exploratoria con:

```
homogen('Prec-m', 2001, 2010, std=2, gp=4, expl=TRUE)
```

nos permite ajustar los umbrales de la homogeneización, que efectuamos con:

```
homogen('Prec-m', 2001, 2010, std=2, dz.max=7, snhtl=10, gp=4)
```

que corrige 3 saltos en la media y no elimina ningún dato anómalo.

Para ajustar los datos diarios por interpolación de las correcciones mensuales usaremos la función *m2dd*:

```
m2dd('Prec', 2001, 2010)
```

Los datos diarios ajustados se han grabado en *Prec_AJ-d_2001-2010.dat*, pero contienen las mismas lagunas de datos de las series originales. Para rellenar esas lagunas, aplicaremos de nuevo la función de homogeneización (especificando ahora *snhtl=0*), con lo que habremos finalizado el proceso de ajuste:

```
homogen('Prec_AJ', 2001, 2010, nm=0, std=2, snhtl=0)
```

La función *homogen* devuelve los resultados en archivos binarios de R, con extensión '*rda*', que serán leídos por las funciones *dahstat* y *dahgrid* para generar los productos deseados (y también pueden ser cargados en memoria por el usuario mediante la función *load*).

3.4. Obtención de productos con *dahstat*

Esta función permite obtener rápidamente tablas mensuales de distintos estadísticos a partir de las series homogeneizadas. Ejemplos a partir de las series mensuales:

```
dahstat('Prec-m', 2001, 2010) #medias mensuales
dahstat('Prec-m', 2001, 2010, stat='std') #desv. Típicas
dahstat('Prec-m', 2001, 2010, stat='q', prob='.2') #quintil 1
dahstat('Prec-m', 2001, 2010, stat='tnd') #tendencias
```

Esta última orden genera dos ficheros: el *Prec-m_2001-2010_tnd.csv* que contiene las tendencias (calculadas por regresión lineal con el tiempo), y otro con los correspondientes p-valores (*Prec-m_2001-2010_pval.csv*).

También pueden generarse ficheros conteniendo las series homogeneizadas:

```
dahstat('Prec-m', 2001, 2010, stat='series')
```

Se obtendrán así dos ficheros CSV por cada serie reconstruida a partir de cada fragmento homogéneo, uno con los datos homogeneizados, y otro indicando si cada dato es el observado originalmente (0), ha sido rellenado (1), o se ha corregido (2).

Por defecto estos productos se generan para todas las series reconstruidas, pero se pueden especificar los códigos de las estaciones que nos interesen, o seleccionarlas de acuerdo a diferentes criterios (funcionamiento al final del periodo estudiado, fragmento más largo, SNHT superior a un umbral, etc).

Cuando se aplica a datos diarios, esta función solo genera un valor por serie, pero en futuras versiones se procurará que también produzca tablas mensuales.

3.5. Series homogeneizadas interpoladas en una rejilla

Muchas bases de datos climato lógicas suministran la información en forma de rejilla, lo que implica una estructura espacio-temporal regular que facilita el estudio de la variabilidad climática. La función *dahgrid* se ha programado para poder obtener automáticamente datos en rejilla homogeneizados, procedentes del fichero *.rda que genera la función *homogen* vista anteriormente.

Para ello el usuario debe definir previamente la rejilla espacial sobre la que desee que se haga la interpolación, como por ejemplo:

```
rejilla=expand.grid(x=seq(1,2,.02),y=seq(41,41.6,.02))
```

(Si el área contiene zonas marítimas o de países vecinos para los que no se dispone de datos, puede ser conveniente eliminar de la rejilla los puntos correspondientes).

Y ahora ya se pueden interpolar los datos homogeneizados sobre la rejilla definida, para cada paso de tiempo, con:

```
dahgrid('Prec-m',2001,2010,grid=rejilla)
```

quedando grabados en el fichero *Prec-m_2001-2010.nc* en formato NetCDF.

3.6. Homogeneización de miles de series mediante áreas solapadas

Si el número de series excede la capacidad del ordenador, o alarga el tiempo de proceso demasiado (muchos días), se puede utilizar la función *homogsplit* para realizar la homogeneización dividiendo el dominio espacial en áreas rectangulares. Estas áreas han de solaparse en sus fronteras, para que se puedan utilizar series próximas de las áreas vecinas, pues de lo contrario aparecerían discontinuidades al analizar los resultados globalmente.

Como ejemplo vamos a suponer que queremos analizar varios miles de series de temperatura media de la Península Ibérica, que hemos recopilado en el fichero de datos *Tm_1951-2015.dat* (y *Tm_1951-2015.est* sería su correspondiente fichero de estaciones). Entonces podríamos realizar la homogeneización subdividiendo la península por el meridiano -3,5° y los paralelos 38,5° y 41°, con un solapamiento de 0,6° en longitud y 0,4° en latitud, mediante la orden:

```
homogsplit('Tm',1951,2015,xc=3.5,yc=c(38.5,41),
xo=.6,yo=.4)
```

Se realizarían así 6 procesos independientes de homogeneización (una para cada subárea), cuyos gráficos aparecerían también en sendos documentos en PDF, pero los datos homogeneizados quedarían finalmente agrupados en un solo fichero llamado *Tm_1951-2015.rda*

4. CONCLUSIONES

Climatol permite automatizar de una manera cómoda todo el proceso de un estudio climatológico, desde la recopilación de datos de observación hasta la obtención de productos a partir de las series homogeneizadas y con control de calidad.

Como el trabajo mediante órdenes en una terminal de texto puede resultar un poco árido para el usuario, en el futuro se tratará de implementar una interfaz gráfica que facilite su uso.

REFERENCIAS

- Aguilar, E., I. Auer, M. Brunet, T. C. Peterson, J. Wieringa (2003). Guidelines on climate metadata and homogenization. WCDMP-No. 53, WMO-TD No. 1186. World Meteorological Organization, Geneva.
- Guijarro, J. A. (2016). R contributed package ‘Climatol’, available at the web site <http://www.climatol.eu/index.html>
- Peterson, T., D. Easterling, T. Karl, P. Groisman, N. Nicholls, N. Plummer, S. Torok, I. Auer, R. Böhm, D. Gullett, L. Vincent, R. Heino, H. Tuomenvirta, O. Mestre, T. Szentimrey, J. Salinger, E. Førland, I. Hanssen-Bauer, H. Alexandersson, P. Jones, D. Parker (1998). Homogeneity Adjustments of ‘In Situ’ Atmospheric Climate Data: A Review. *Int. J. Climatol.*, 18, pp. 1493-1518.
- R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.