

# Estadística climatológica con R

## 5. Series temporales

### 5.1 Homogeneización y relleno de lagunas

### 5.2 Paquete Climatol

## 5. Series temporales

Una **serie temporal** o **cronológica** es una secuencia de datos, observaciones o valores, medidos en determinados momentos y ordenados cronológicamente.

Para el análisis de las series temporales se usan métodos que ayudan a interpretarlas y que permiten extraer información representativa sobre las relaciones subyacentes entre los datos de la serie o de diversas series y que permiten en diferente medida y con distinta confianza *extrapolar o interpolar los datos* y así *predecir el comportamiento de la serie* en momentos no observados.

Uno de los usos más habituales de las series de datos temporales es su análisis para predicción y pronóstico (así se hace por ejemplo con los datos climáticos).

Las series temporales se usan también para estudiar la relación causal entre diversas variables que cambian con el tiempo y se influyen entre sí.

La característica fundamental de las series temporales es que las observaciones sucesivas *no son independientes entre sí*, y el análisis debe llevarse a cabo teniendo en cuenta el *orden temporal* de las observaciones. Los métodos estadísticos basados en la independencia de las observaciones no son válidos para el análisis de series temporales porque las observaciones en un instante de tiempo dependen de los valores de la serie en el pasado.

Una serie es *estacionaria* si la media y la variabilidad se mantienen constantes a lo largo del tiempo. Esto se refleja gráficamente en que los valores de la serie tienden a oscilar alrededor de una media constante y la variabilidad con respecto a esa media también permanece constante en el tiempo. Para este tipo de series tiene sentido conceptos como la media y la varianza. Esta es la base cuando hablamos de las series climatológicas en periodos de 30 años, suponemos que son estacionarias en ese periodo y las variables son independientes durante estos años.

Un ejemplo de una serie estacionaria:

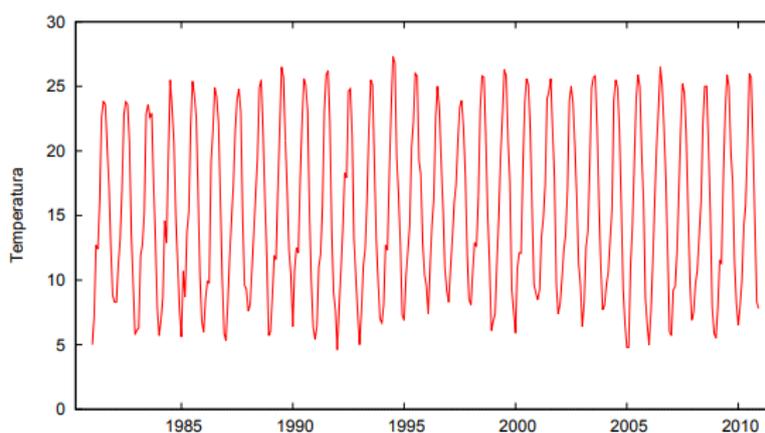


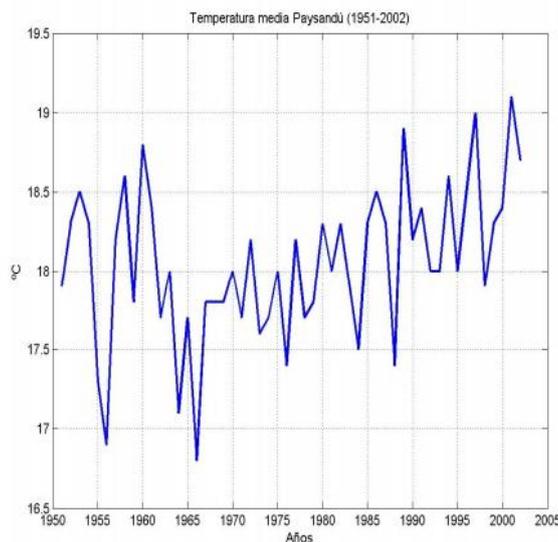
Figura 2: Representación gráfica de las temperaturas en la ciudad de Granada desde enero de 1981 a diciembre de 2010

La serie es estable, a simple vista, alrededor de un valor central (podría tener una pequeña tendencia creciente). Si representamos un histograma de esta serie, podemos describir adecuadamente la información.

Una serie es *no estacionaria* si la media y/o la variabilidad cambian a lo largo del tiempo. Los cambios en la media determinan una tendencia a crecer o decrecer a largo plazo, por lo que la serie no oscila alrededor de un valor constante.

- Series no estacionarias pueden mostrar *cambios de varianza*.
- Series no estacionarias pueden mostrar *una tendencia*, es decir que la media crece o baja a lo largo del tiempo.

Ejemplo de serie no estacionaria:



Las series temporales se suelen analizar suponiendo que son suma de varias componentes:

$$X_t = T_t + S_t + I_t$$

*Valor observado = Tendencia + Estacionalidad + Irregular (Aleatoria)*

**Tendencia:** Se puede definir como un cambio a largo plazo que se produce en relación al nivel medio, o el cambio a largo plazo de la media. La tendencia se identifica con un movimiento suave de la serie a largo plazo.

**Estacionalidad:** Muchas series temporales presentan cierta periodicidad o dicho de otro modo, variación de cierto periodo (anual, mensual, diario, ...). Por ejemplo, la temperatura media aumenta en general en verano y disminuye en invierno. Estos tipos de efectos son fáciles de entender y se pueden medir explícitamente o incluso se pueden eliminar del conjunto de los datos, *desestacionalizando* la serie original.

**Componente aleatoria o irregular o ruido:** Una vez identificados los componentes anteriores y después de haberlos eliminado, persisten unos valores que son aleatorios. Se pretende estudiar qué tipo de comportamiento aleatorio presentan estos residuos, utilizando algún tipo de modelo probabilístico que los describa. Son variaciones aleatorias alrededor de los componentes anteriores.

De las tres componentes reseñadas, las dos primeras son componentes determinísticas, mientras que la última es aleatoria. Es interesante obtener o “aislar” los distintos componentes.

### **Análisis de la tendencia**

Para estimar la tendencia supondremos que tenemos una serie no estacionaria sin componente estacional, es decir, que la serie se puede descomponer en

$$X_t = T_t + I_t$$

Para estimar  $T_t$  debemos realizar alguna hipótesis sobre su forma. Vamos a analizar varios casos.

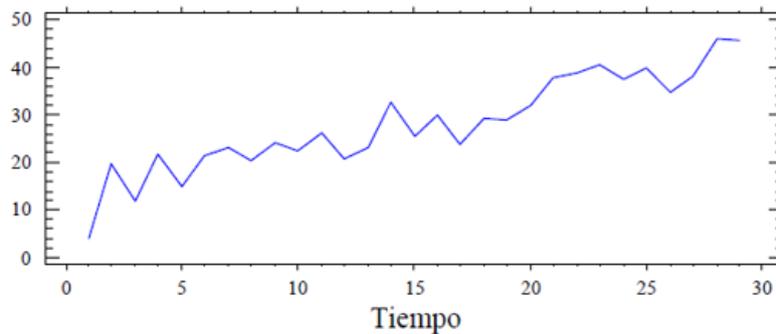
#### **Tendencia determinista**

En este caso supondremos que la tendencia es una función determinística. La función más sencilla posible es una recta, es decir, una **tendencia lineal**

$$T_t = a + bt$$

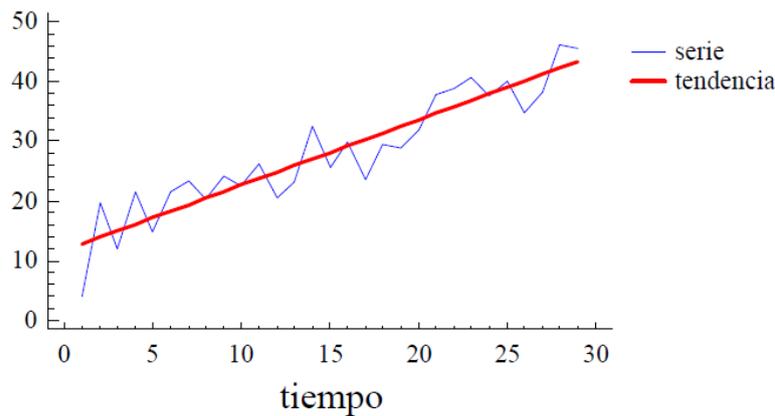
que se estima mediante un modelo de regresión lineal entre las variables  $X_t$  y el tiempo  $t = 1, 2, 3, \dots$  (método de mínimos cuadrados).

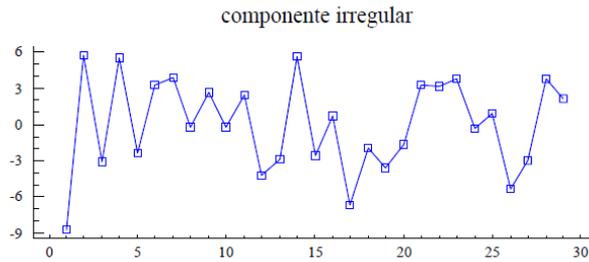
Este ejemplo:



presenta una serie sin estacionalidad que presenta una tendencia que se podría expresar de forma lineal. La tendencia de la serie y la componente irregular serían, en este ejemplo,

#### **serie y tendencia lineal**





En otros casos, no es posible ajustar la tendencia mediante una recta. En estos casos, lo mejor sería ajustar la tendencia a un polinomio o a la curva que mejor se pueda ajustar. Para ello tendríamos que ajustar una regresión no lineal. Otra opción es describir la tendencia *de manera evolutiva (por medio de medias móviles)* o *diferenciar la serie*.

### **Tendencia evolutiva (medias móviles)**

A menudo, la tendencia de la serie no sigue una recta y evoluciona a lo largo del tiempo. En ese caso, un método general de estimar  $T_t$  es suponer que evoluciona lentamente en el tiempo, y que se puede aproximar con una función sencilla para intervalos cortos del tiempo. Por ejemplo, si una recta es una representación válida para tres periodos consecutivos:

$$T_{t-1} = T_t - \Delta T, \quad T_t = T_t \text{ y } T_{t+1} = T_t + \Delta T$$

Si hacemos la media de las tres observaciones consecutivas,  $m_t = (x_{t-1} + x_t + x_{t+1})/3$ , tendríamos que:  $m_t = (T_t + I_{t-1} + I_t + I_{t+1})/3$  es decir “descubriríamos” la tendencia subyacente.

Para un instante  $t$ , se define la media móvil de orden 3 de la serie como

$$m_t = \frac{x_{t-1} + x_t + x_{t+1}}{3}$$

Suponemos que la tendencia  $T_t$  satisface

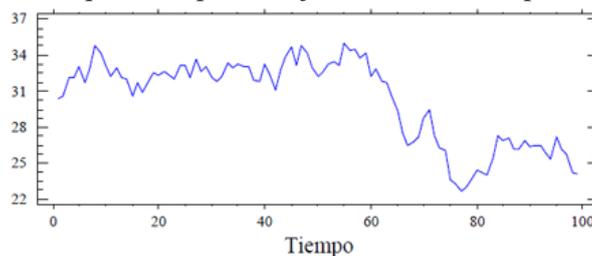
$$T_t = m_t - \frac{I_{t-1} + I_t + I_{t+1}}{3}$$

Como la media del componente irregular es cero, podemos suponer que la media de los tres valores  $(I_{t-1}, I_t, I_{t+1})$  es pequeña, de esta manera  $m_t$  recoge fundamentalmente la tendencia de la serie en el instante  $t$ . Se observa que realizando esta operación se pierde la primera observación y la última. Si calculamos las medias móviles de orden 5, perderemos las dos primeras observaciones y las dos últimas.

Es posible calcular medias móviles de órdenes más altos. Cuando crece el orden, el valor de  $m_t$  cambia más suavemente.

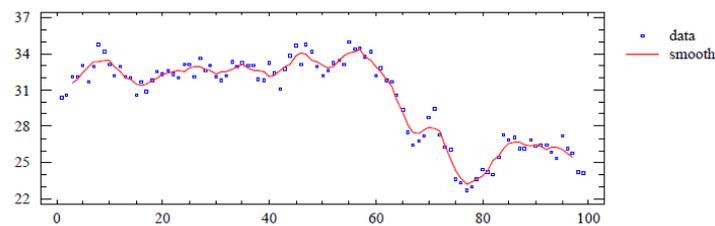
Con medias móviles de órdenes altos, suavizamos los efectos estacionales

En el siguiente ejemplo en el que no es posible ajustar la tendencia por medio de una recta:

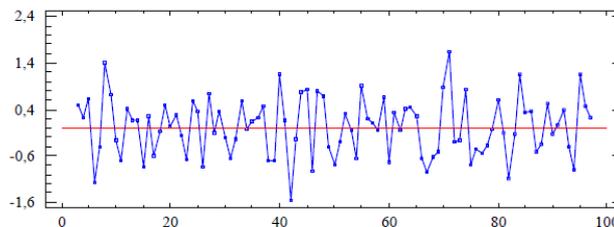


Si aplicamos una media móvil de orden 5 queda

medias móviles de orden 5



componente irregular



### Diferenciación de la serie

Un tercer método más general para eliminar la tendencia consiste en suponer que la tendencia evoluciona lentamente en el tiempo, de manera que en el instante  $t$  la tendencia debe estar próxima a la tendencia en el instante  $t-1$ . De esta forma, si restamos a cada valor de la serie el valor anterior, la serie resultante estará aproximadamente libre de tendencia. Esta operación se denomina *diferenciación* de la serie y consiste en pasar de la serie original  $x_t$  a la serie  $y_t$  mediante:

$$y_t = x_t - x_{t-1}$$

## Series temporales con R

*R* cuenta con numerosas funciones para el análisis de series temporales. El uso de tales funciones requiere que los objetos a los que se aplican sean de la clase *ts* (*time series*). La función *ts()* convierte un objeto a serie temporal.

La construcción de objetos *ts* requiere que los datos de partida estén distribuidos regularmente en la escala temporal utilizada; por ejemplo, que haya siempre un dato por día, o dos datos por mes o un dato cada dos años.

### Ejemplo 5.1

En el archivo *Tem.media.mensual.Retiro.txt* tenemos las temperaturas medias mensuales de la Estación del Retiro de Madrid desde el año 1940 hasta el 2017. Se hallan guardadas con cabecera (*temp*) y en orden anual, empezando por enero de 1940, febrero, ... hasta diciembre, a continuación el año 1941, enero, febrero..., aunque no hay ninguna mención a la fecha en el fichero.

Primero leeremos el fichero:

```
> temperaturas <- read.table('Tem.media.mensual.Retiro.txt', header=TRUE)
```

Después transformamos el *data.frame* a una serie temporal *ts*:

```
> temp.Retiro=ts(temperaturas$temp,freq=12,start=c(1940,1))
```

Hemos especificado *freq=12*, lo que indica que a cada unidad temporal le corresponden 12 observaciones; *R* asume entonces que la unidad temporal a considerar es el *año*. Además la opción *start=c(1940,1)* indica que la primera observación corresponde a enero de 1940.

Si listamos ahora vemos la estructura de *temp.Retiro*

> *temp.Retiro*

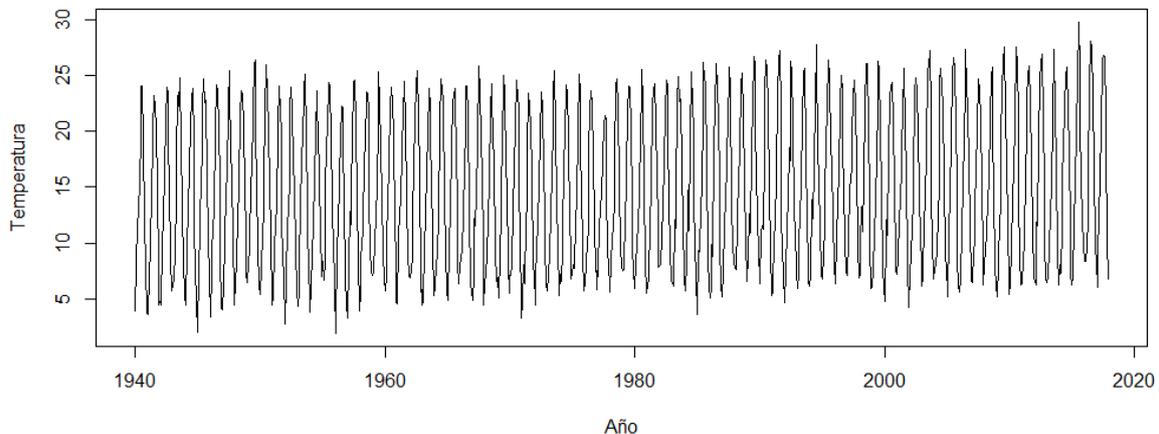
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1940	3.9	8.2	11.7	12.8	15.7	19.3	24.0	24.1	19.1	12.0	7.8	3.8
1941	3.6	7.0	9.6	11.1	13.2	20.1	23.2	22.3	20.2	15.0	7.5	4.5
1942	4.8	4.5	9.9	12.2	16.3	21.0	23.9	23.4	19.0	15.1	8.7	5.7
1943	6.3	6.8	9.2	14.2	17.8	23.2	22.2	24.8	18.8	13.5	7.7	5.9
1944	5.3	4.5	9.7	12.9	16.8	20.3	23.1	23.8	19.2	11.2	7.8	4.7
1945	2.0	7.5	11.2	16.5	17.2	22.2	24.7	22.6	22.8	16.0	10.5	6.9
1946	3.4	7.4	9.0	11.6	12.7	20.1	24.2	23.6	20.4	15.4	8.8	4.2
1947	4.0	6.1	10.1	14.7	16.0	22.5	25.4	23.1	19.1	15.2	11.2	4.4
1948	6.4	8.3	12.6	12.1	15.0	21.6	23.6	23.4	20.5	14.9	10.8	7.3
1949	6.5	8.3	9.4	15.9	16.0	21.9	26.0	26.4	19.9	15.1	9.5	6.3
1950	5.4	7.8	10.8	12.8	15.6	21.8	26.0	23.5	20.1	15.3	10.7	4.5
1951	5.2	5.6	9.1	12.2	12.9	20.5	24.1	21.9	19.7	12.3	8.7	7.2
1952	2.8	6.5	12.1	12.6	16.1	22.1	23.9	21.9	17.2	14.9	9.4	5.7
1953	4.3	5.7	9.4	12.1	18.4	18.8	23.8	25.1	19.5	13.3	10.2	8.9
1954	3.8	5.9	9.1	11.5	15.8	19.8	23.6	21.6	20.3	15.6	11.1	7.0
1955	8.4	6.7	8.4	14.8	18.6	20.3	24.4	24.0	19.4	13.9	8.8	7.7
1956	5.9	1.9	9.8	11.2	15.8	19.4	22.2	22.1	18.6	13.9	6.9	4.4
1957	3.3	8.3	12.8	11.7	15.3	19.0	24.5	24.6	20.9	13.7	8.5	3.9
1958	5.5	9.0	9.5	11.4	17.7	18.6	23.5	23.4	21.9	14.2	9.3	7.5
1959	7.1	7.4	10.2	12.4	15.3	20.7	25.3	22.8	18.4	14.1	9.2	6.7
1960	5.7	7.1	10.4	13.7	16.9	22.4	23.9	22.6	19.8	11.4	8.9	4.7
1961	4.6	10.2	12.9	13.7	18.4	21.6	24.5	24.4	21.0	13.7	9.0	7.0
1962	6.9	7.5	8.8	12.4	16.8	21.0	24.2	25.4	21.4	16.1	6.7	4.4

...

Así pues, *R* organiza automáticamente la variable en meses y años. La representación gráfica tiene en cuenta esta secuencia temporal:

> *plot(temp.Retiro, main="Temperatura media mensual en el observatorio del Retiro (Madrid)", xlab="Año", ylab="Temperatura")*

Temperatura media mensual en el observatorio del Retiro (Madrid)



Las funciones *start()* y *end()* muestran, respectivamente, los instantes correspondientes a la observación inicial y final de una serie temporal:

```
> start(temp.Retiro)
[1] 1940  1
```

```
> end(temp.Retiro)
[1] 2017 12
```

La función *window()* extrae los valores de la serie temporal comprendidos entre una fecha de inicio y otra final.

```
> window(temp.Retiro, start=c(1940,1),end=c(1942,12)) # 1 y 12 hacen referencia a los meses.
Se pueden poner solo el año también (start=1940)
```

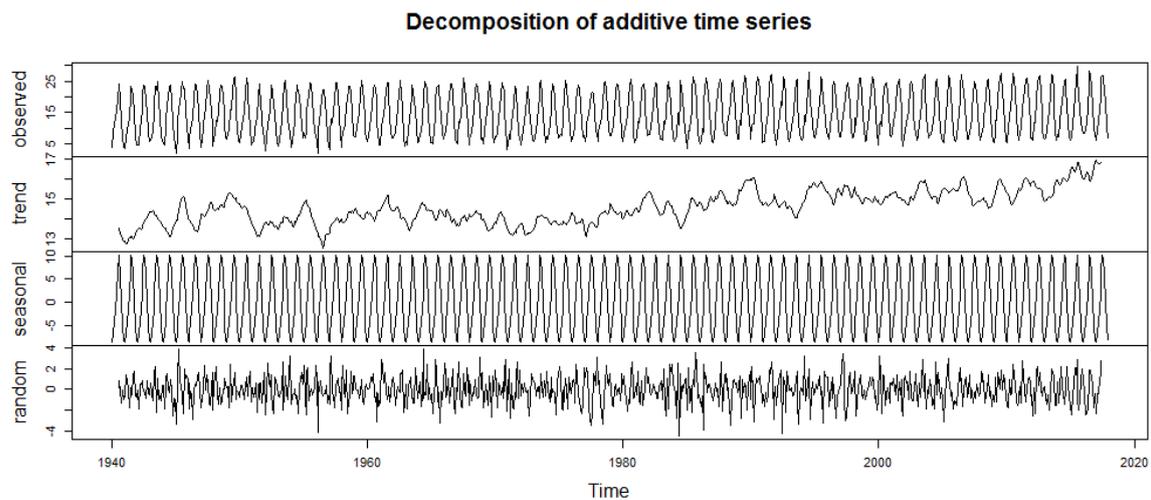
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1940	3.9	8.2	11.7	12.8	15.7	19.3	24.0	24.1	19.1	12.0	7.8	3.8
1941	3.6	7.0	9.6	11.1	13.2	20.1	23.2	22.3	20.2	15.0	7.5	4.5
1942	4.8	4.5	9.9	12.2	16.3	21.0	23.9	23.4	19.0	15.1	8.7	5.7

### Descomposición de una serie temporal

La función `decompose()` permite descomponer una serie temporal en sus componentes estacional, de tendencia y de ruido. En el caso de `decompose()` la descomposición se lleva a cabo mediante un ajuste de medias móviles. Existen otras funciones como `stl()` que realiza la descomposición mediante un ajuste polinómico local.

> `descomp=decompose(temp.Retiro)` # descompone en componentes

> `plot(descomp)` # dibuja el resultado

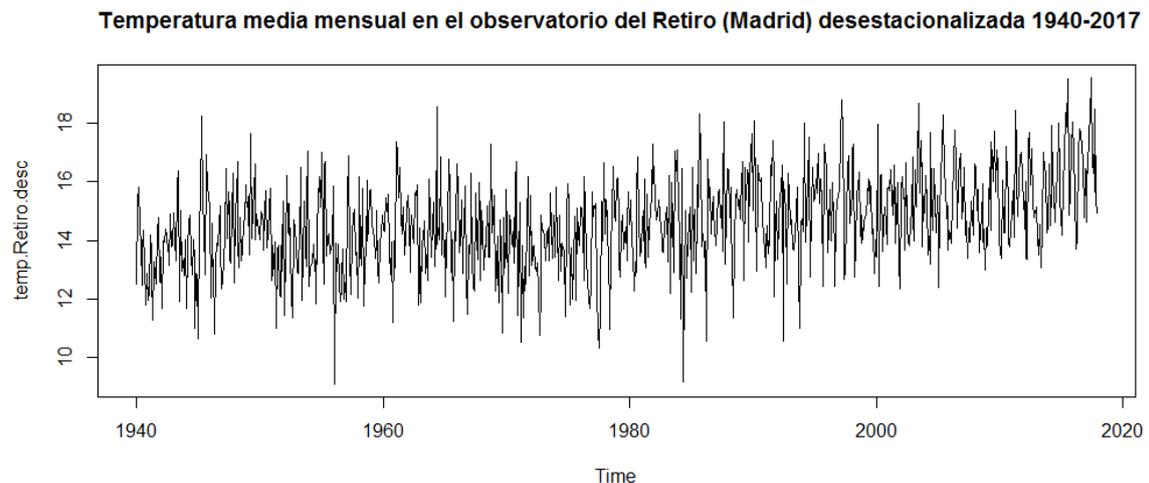


En esta gráfica se presenta la serie original, la componente de la tendencia, la componente estacional y la parte aleatoria.

Si queremos “eliminar” la componente estacional de nuestra serie temporal, y dejar solamente la componente de tendencia y el ruido, basta con restar a la serie original la parte estacional (`descomp$seasonal`) que acabamos de calcular:

> `temp.Retiro.desc = temp.Retiro - descomp$seasonal` # restamos la parte estacional a la serie original

> `plot(temp.Retiro.desc, main="Temperatura media mensual en el observatorio del Retiro (Madrid) desestacionalizada 1940-2017")`



También se podría eliminar la componente de ruido.

## 5.1 Homogeneización y relleno de lagunas

Las series de observaciones meteorológicas son de capital importancia para el estudio del clima. Sin embargo, estas series se ven frecuentemente contaminadas por eventos ajenos a dicha variabilidad: errores en la toma de medidas o en su transmisión, y cambios en el instrumental utilizado, en la ubicación del observatorio o en su entorno. Estos últimos pueden ser cambios bruscos, como el incendio de un bosque colindante, o graduales, como la posterior recuperación de la vegetación. Estas alteraciones de las series, denominadas *inhomogeneidades*, enmascaran los verdaderos cambios del clima y hacen que el estudio de las series conduzca a conclusiones erróneas.

Se considera que una serie es *homogénea* cuando sus variaciones *responden exclusivamente a causas climáticas*.

Aparte de las posibles *inhomogeneidades* presentes en las series, la falta de datos en éstas constituye otro de los problemas frecuentes que requiere el uso de ciertas técnicas de *relleno de lagunas*.

Para abordar este problema se han desarrollado desde hace muchos años metodologías de *homogeneización* que permitan eliminar o reducir en lo posible estas alteraciones indeseadas. Inicialmente consistían en comparar la serie problema con otra supuestamente homogénea, pero como esta suposición es muy arriesgada, se pasó a construir *una serie de referencia* a partir del promedio de otras seleccionadas por su proximidad o elevada correlación, diluyendo así sus posibles inhomogeneidades.

Bajo un contexto de Cambio Climático como el actual (IPCC) es necesario disponer de series completas y homogéneas de longitud suficientemente larga que permitan el análisis de la variabilidad natural, así como de otros comportamientos de las series, como el de sus valores medios o extremos.

Existen muchos paquetes de programación que implementan estos métodos para que puedan ser usados por la comunidad climatológica (<http://www.climatol.eu/tt-hom/index.html>).

El paquete de *R Climatol* (<https://CRAN.R-project.org/package=climatol>) contiene funciones para el control de calidad, homogeneización y relleno de los datos faltantes en un conjunto de series de cualquier variable climática.

## 5.2 Paquete Climatol

(<http://www.climatol.eu/>)

*Climatol* es un conjunto de rutinas de depuración y homogeneización de datos climatológicos, creado por un compañero de AEMET: *José Antonio Guijarro*.

*Climatol* está desarrollado para su uso dentro del paquete estadístico *R*. Las funcionalidades implementadas incluyen la comparación de cada serie original con una de referencia, estimada a partir de un promedio de las demás (modulado mediante una función inversa de la distancia y previa tipificación de los datos de cada estación), así como la inspección gráfica de los datos y sus anomalías. Los errores puntuales, saltos en las medias y tendencias se pueden detectar en las series de diferencias entre los datos originales y los estimados mediante las pruebas paramétricas habituales.

Trabajaremos con la versión **4.0.0** de *Climatol* (mayo 2023).

Primero se construye *una serie de referencia* para cada estación usando varias series reales. Una vez construida la serie de referencia, la comparación con la serie problema se puede realizar estudiando las diferencias entre las series a comparar.

Esta serie de diferencias debe estar ya libre de las oscilaciones que el elemento climático estudiado haya podido sufrir en la zona de estudio, así como de las características de persistencia y estacionalidad propias de las series climatológicas. Por tanto, en el caso ideal, esta serie debería comportarse como un *ruido blanco*, pudiendo considerarse como inhomogeneidad a cualquier anomalía que se aparte del mismo. (El ruido blanco es una señal aleatoria que se caracteriza por ser una señal que no guarda correlación estadística en distintos momentos temporales). Estas *anomalías* pueden agruparse en tres tipos:

1. *Errores aislados*, debidos a errores de lectura o transcripción, o a errores puntuales en el sensor (estaciones automáticas). Se pueden detectar como datos anómalos aislados (*outliers* en la terminología inglesa).
2. *Salto en las series*, atribuibles a cambios de instrumentación o de las condiciones de instalación de los aparatos. Se pueden detectar por cambios en los promedios anteriores y posteriores a los saltos.
3. *Derivas graduales* en las series, debidos a cambios progresivos en la respuesta de los sensores (descalibración) o en el entorno (crecimiento urbano, cambios en los usos del suelo, etc.). Detectables mediante análisis de tendencias o de desviaciones acumuladas.

En la figura 1 se ilustran estos tres tipos de inhomogeneidades. La serie 1 simula una serie aleatoria con distribución normal (ruido blanco), como sería el caso de una serie de diferencias entre dos series climáticas homogéneas de observatorios próximos. En la serie 2 se han simulado dos errores puntuales: +5s en el término 25, y -5s en el 65. La serie 3 presenta dos cambios bruscos de la media: +3s en el término 40, y -2s en el 75 (-5s respecto a los términos anteriores, 40 a 74). La serie 4 simula una tendencia desde -2s hasta +2s a lo largo de los 100 términos de la serie. Por último, en la serie 5 se han incluido todas las inhomogeneidades anteriores, observándose un notable enmascaramiento mutuo que dificulta su diagnóstico y corrección, como lamentablemente cabe esperar en las series climatológicas reales.

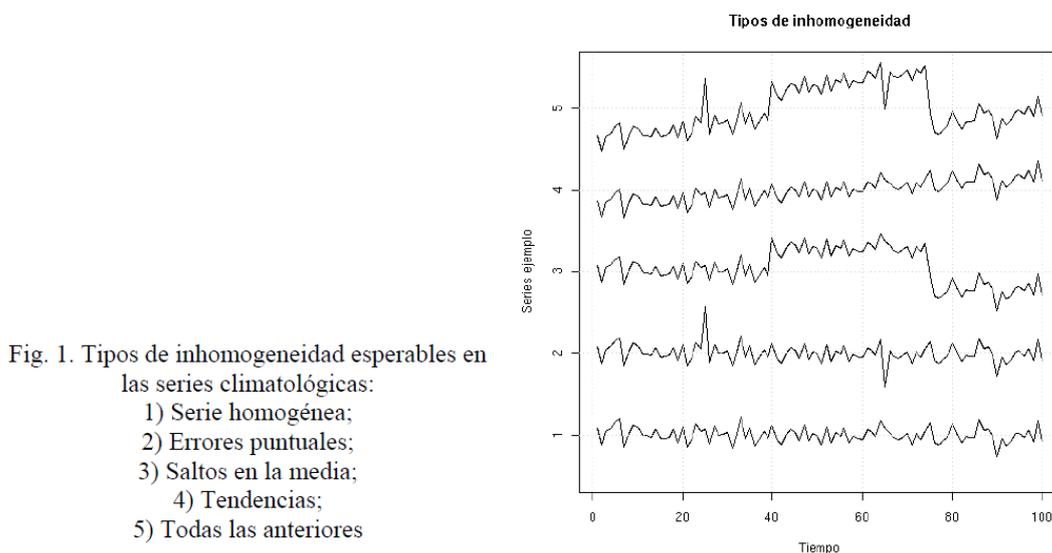


Fig. 1. Tipos de inhomogeneidad esperables en las series climatológicas:  
1) Serie homogénea;  
2) Errores puntuales;  
3) Saltos en la media;  
4) Tendencias;  
5) Todas las anteriores

## **Ejemplos prácticos del mismo paquete: Procedimientos de homogeneización**

Ahora veremos una aplicación práctica a través de unos ejemplos del mismo paquete *Climatol*.

### *Preparación de los ficheros de entrada*

*Climatol* solo necesita dos ficheros de entrada, uno con la lista de coordenadas, códigos y nombres de las *estaciones*, y otro con todos los *datos*, en orden cronológico desde la primera estación hasta la última. Como el fichero de datos carece de toda referencia temporal, todos los datos deben estar presentes, para todo el periodo de estudio, representando los datos ausentes con *NA* u otro código distintivo. Ambos archivos comparten el mismo nombre básico *VAR\_aaaa-AAAA* donde VAR es un acrónimo de la variable a estudiar, aaaa el primer año y AAAA el último de los datos, pero tienen distintas extensiones: *dat* para los datos y *est* para las estaciones. Ambos son ficheros de texto plano.

Los archivos los generamos desde el mismo *climatol* para este ejemplo.

```
> install.packages("climatol") # instalamos el paquete
> library(climatol) # cargamos las funciones del paquete
> data(climatol_data) # cargamos los datos de ejemplo en memoria
> write(Temp.dat, 'Temp_1961-2005.dat') # grabamos los datos a un fichero
> write.table(Temp.est, 'Temp_1961-2005.est', row.names=FALSE, col.names=FALSE) #
grabamos los datos de estaciones a otro fichero

> rm(dat, est.c) # borrar los datos cargados en memoria
```

Estos archivos contienen 45 años de temperaturas diarias de prueba de 5 estaciones inventadas. Se pueden inspeccionar para ver su estructura con un editor de texto. Las líneas del fichero de estaciones *Temp\_1961-2005.est* son:

```
-2.5059 39.0583 210 "st01" "Station 1"
-2.7028 38.9808 112 "st02" "Station 2"
-2.63 38.8773 111 "st03" "Station 3"
-2.5699 38.9205 112 "st04" "Station 4"
-2.4663 38.9885 125 "st05" "Station 5"
```

Como se puede ver, cada línea tiene, en formato libre separado por espacios, las coordenadas X, Y, Z de la estación, seguidas por el código y el nombre. Normalmente X e Y son la longitud y la latitud, en grados con decimales y con el signo adecuado para indicar Oeste, Este, Norte o Sur. Z es la altitud en metros.

Algunas líneas del fichero de datos *Temp\_1961-2005.dat* son:

```
> write(Temp.dat[41:80,4], stdout(), ncolumns=10) # mostrar un fragmento del fichero de datos:
.....
20.4 23 25.8 26.1 24.8 18.9 14.9 11.7 10.6 8.5
12.6 14.2 19.3 22.4 26.4 26.1 NA NA 15.9 12.5
13.2 NA 12.6 17 NA NA NA 25.5 23.1 18.2
12.5 10.1 10.7 11.2 13.5 NA 18.1 19.5 NA 24.9
.....
```

Estos 40 datos son las temperaturas medias de los 40 días siguientes al 1 de febrero de 1981 en la cuarta estación (Station 4). Las siguientes líneas del fichero contienen el resto de datos de esta estación hasta el 31 de diciembre de 2005, seguidos por todos los datos de las otras estaciones relacionadas en el fichero *Temp\_1961-2005.est*.

### *Procedimiento para las series diarias*

La función de homogeneización de *Climatol* se llama *homogen*, y su aplicación más trivial y automática, solo requiere especificar tres parámetros: el acrónimo de la variable, y los años inicial y final del periodo de estudio. Se haría como se expone a continuación, aunque primero vamos a realizar un análisis exploratorio.

```
> homogen('Temp', 1961, 2005)
```

Esta orden se puede aplicar tanto si los datos son diarios, mensuales, bimestrales, trimestrales, semestrales o anuales: la función estimará la frecuencia a partir de la cantidad de datos presentes. Sin embargo, los umbrales para el rechazo de valores atípicos y la detección de punto de inflexión pueden ser muy diferente dependiendo de la periodicidad de los datos y las correlaciones cruzadas de las series. Por lo tanto, *es aconsejable hacer una primera aplicación en modo exploratorio*:

> *homogen('Temp', 1961, 2005, expl=TRUE)*

Ahora podemos abrir el archivo de salida *Temp\_1961-2005.pdf* para revisar sus diferentes gráficos de diagnóstico. Primero vemos la disponibilidad de datos, en todas las estaciones y globalmente (figura 5). Idealmente, debería haber 5 o más datos disponibles en cada paso temporal, o un mínimo de tres, niveles marcados con líneas de trazos verdes y rojos en la parte derecha de la figura, pero la función no se parará excepto cuando no haya datos disponibles en ninguna estación en uno o más pasos temporales, situación que detendrá el proceso con un mensaje de error.

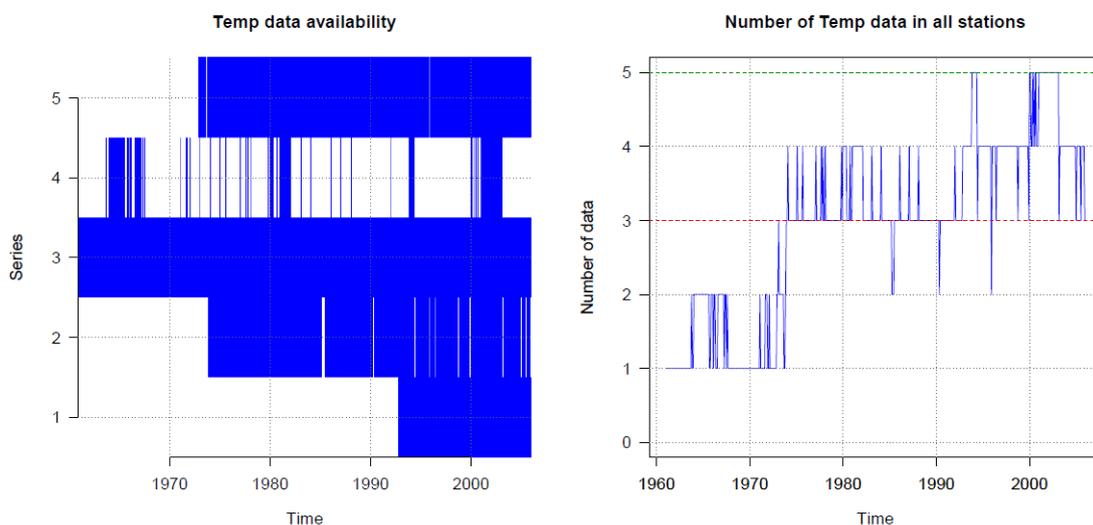


Figura 5: Disponibilidad de datos a lo largo del periodo estudiado.

Los siguientes gráficos muestran diagramas de caja de los datos en cada estación y un histograma del conjunto de todos los datos . La presencia de valores muy anómalos sería evidente en estos gráficos, lo que permitiría al usuario tomar medidas correctivas. También el histograma de frecuencias será útil para decidir si la distribución de probabilidad es casi normal o muy sesgada. En el segundo caso, puede ser preferible utilizar la normalización por proporción respecto a la media (utilizando el parámetro  $std=2$ ) en lugar de la estandarización por defecto.

Los gráficos que siguen se centran en las correlaciones entre las series y su clasificación en grupos con variabilidad similar, que luego se representan en un mapa. Las correlaciones son generalmente más bajas cuando la distancia entre estaciones es mayor, como en este ejemplo. Cuanto más altas sean las correlaciones, mayor será la fiabilidad de la homogeneización y el relleno de datos ausentes. En particular, las correlaciones deben ser siempre positivas, al menos dentro de un rango de distancias razonables. De lo contrario, probablemente haya discontinuidades geográficas que produzcan diferencias climáticas (por ejemplo, una cresta montañosa puede producir regímenes de precipitación opuestos a ambos lados de la misma). Esto puede confirmarse con el mapa de estaciones, en el que los grupos de variabilidad similar se ubicarían en distintas zonas.

La figura 6 muestra un ejemplo, con las anomalías de la serie 1 a la izquierda y la reconstrucción de series completas a partir de los dos fragmentos homogéneos a la derecha. El gráfico de anomalías presenta dos líneas adicionales en la parte inferior que informan sobre la distancia mínima de los datos vecinos (en verde) y el número de datos de referencia utilizados (en naranja), usando ambos la escala logarítmica del eje derecho. En el gráfico de series reconstruidas se representan sus medias

anuales móviles, salvo que las series sean muy cortas (hasta 120 términos), en cuyo caso se dibujan todos los valores. Las series originales se trazan en color negro, y en colores las reconstruidas.

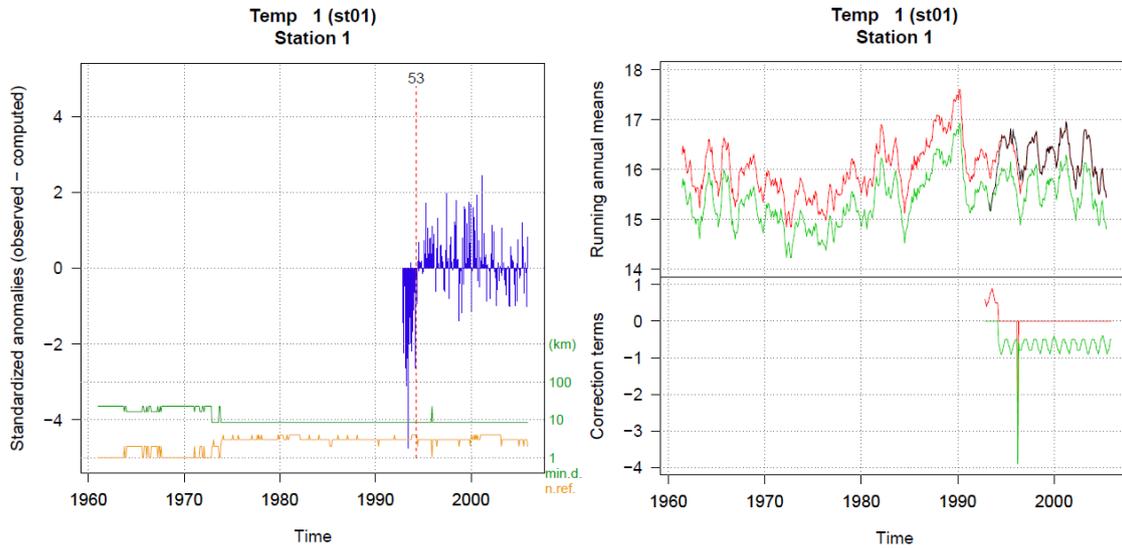
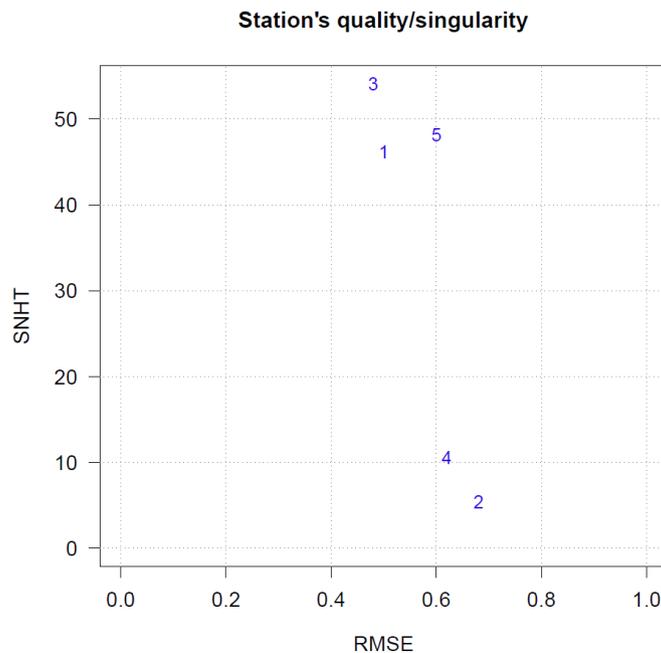


Figura 6: Ejemplo de detección de salto en la media (izquierda) y reconstrucción de las series a partir de cada fragmento homogéneo (derecha).

El documento gráfico termina con histogramas de anomalías estandarizadas y SNHT de las series finales, y una figura que indica su calidad o singularidad. El histograma de anomalías ayuda a elegir umbrales adecuados para rechazar datos muy anómalos, suponiendo que son errores y pueden eliminarse.

La última página del documento muestra un diagrama de números de estación (su orden en el archivo *Temp\_1961-2005.est*) de acuerdo con sus errores típicos (RMSE por sus siglas en inglés) finales y los valores de SNHT. Los RMSE se calculan al comparar los datos estimados y los observados en cada serie. Un valor alto puede indicar una mala calidad, pero también podría deberse a que la estación se encuentra en un sitio peculiar con un microclima distinto. De todos modos, las series homogéneas de estaciones que comparten el clima común de la región tenderán a agruparse en la parte inferior izquierda del gráfico.



Pero dado que nuestro ejemplo se basa en datos diarios y este tipo de series muestran una alta variabilidad que reduce la eficiencia de la detección de sus inhomogeneidades, es mejor agregarlas y homogeneizar primero las series mensuales. Los saltos en la media, por ejemplo, se detectan mejor en las mensuales, ya que las series diarias son muy ruidosas, y posteriormente estos saltos se aplicaran a las diarias. *Climatol* ayuda a obtener datos mensuales de la serie diaria mediante la función *dd2m*, que podemos aplicar aquí de esta manera:

```
> dd2m('Temp', 1961, 2005)
```

Esta orden guarda en *Temp-m\_1961-2005.dat* y *Temp-m\_1961-2005.est* las series mensuales, listas para ser homogeneizadas. (El sufijo *-m* se ha agregado al nombre de la variable para evitar sobrescribir la serie diaria original).

No vamos a explicar la homogeneización de las series mensuales, que es bastante similar a las diarias, pero nos interesa un fichero que se obtiene con los valores de los saltos de la media detectados que podemos usar para mejorar la homogeneización de las series diarias.

Basándonos en el análisis de las series mensuales obtenemos unos valores en un fichero: *Temp-m\_1961-2005\_brk.csv* con los saltos de la media detectados en estas series, de forma que volvemos a correr el comando *homogen* para las series diarias usando este fichero, con la opción *metad=TRUE*.

```
> homogen('Temp', 1961, 2005,metad=TRUE)
```

### *Obtención de productos con los datos homogeneizados*

Las series diarias homogeneizadas y otros datos del proceso quedan almacenados en el fichero binario *Temp\_1961-2005.rda*, que pueden ser cargados en la memoria de trabajo de R mediante la orden:

```
> load("Temp_1961-2005.rda")
```

*Climatol* tiene unas funciones de post-proceso *dahstat* y *dahgrid* para facilitar la obtención de productos de uso corriente a partir de las series homogeneizadas, bien directamente de las diarias, bien de sus agregados mensuales homogeneizados.

Las series homogeneizadas pueden volcarse a dos ficheros de texto *CSV* de este modo:

```
> dahstat('Temp', 1961, 2005, stat='series')
```

Se obtienen así dos ficheros *CSV* para cada serie reconstruida: *Temp\_1961-2005\_series.csv*, contiene todas las series homogeneizadas, y el otro, *Temp\_1961-2005\_flags.csv*, códigos que indican si los datos son *observados* (0), *rellenados* (1, ausentes originalmente) o *corregidos* (2, por inhomogeneidades o por excesiva anomalía).

Los resúmenes estadísticos se crean con la misma función. Aquí se presentan algunos ejemplos (más información en la documentación de R de *dahstat*):

```
> dahstat('Temp', 1961, 2005) # medias de las series
# el parámetro mh=TRUE se refiere a parámetros estadísticos mensuales
```

El comando *dahgrid* se usa para obtener automáticamente datos homogeneizados en rejilla espacio-temporal en formato *netcdf*, procedentes del fichero *\*.rda* que genera la función *homogen* vista anteriormente.

La documentación completa, metodología y ejemplos, puede verse en la página de la aplicación:  
<http://www.climatol.eu/>  
[http://www.climatol.eu/homog\\_climatol-es.pdf](http://www.climatol.eu/homog_climatol-es.pdf)