Homogeneización de series climáticas con Climatol

Versión 3.1.1

https://CRAN.R-project.org/package=climatol

José A. Guijarro

Agencia Estatal de Meteorología (AEMET), D.T. en Islas Baleares, España

Versión de esta guía: 1.3.1 (Agosto de 2018)

English version available at http://www.climatol.eu/homog_climatol-en.pdf

Índice

1. Introducción

Las series de observaciones meteorológicas son de capital importancia para el estudio de la variabilidad climática. 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.

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 una 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. Como esto no garantiza que la serie de referencia sea homogénea, otros métodos proceden a comparar todas las series disponibles por parejas, de modo que la repetida detección de una inhomogeneidad permita identificar las series erróneas. Para mayor información pueden consultarse trabajos como los de Peterson et al. (1998) y Aguilar et al. (2003), que pasan revista a estas metodologías.

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). La Acción COST ES0601 *(Advances in homogenisation methods of climate series: an integrated approach, "HOME")* financió un esfuerzo internacional de comparación de los mismos (Venema et al., 2012). Posteriormente el proyecto MULTITEST (http://www.climatol.eu/MULTITEST/) realizó otra comparación de los métodos actualizados que pudieran ejecutarse en modo totalmente automático. Hasta ese momento la atención estuvo centrada en la homogeneización de series mensuales, principalmente de temperatura y precipitación, pero se ha suscitado un interés creciente en la homogeneización de series diarias, necesarias para el estudio de la variabilidad de los fenómenos extremos, y actualmente el proyecto europeo INDECIS está tratando de aplicar varios métodos a series diarias de diversas variables climáticas.

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. La documentación estándar del paquete proporciona detallada información sobre cada una de sus funciones y sus parámetros de control, así como breves ejemplos de aplicación. Este manual es un complemento a dicha documentación pues, sin dar tantos detalles sobre cada una de las funciones disponibles, explica primero los fundamentos de las metodologías empleadas, y luego proporciona una guía práctica sobre cómo abordar la homogeneización de series diarias o mensuales de distintas variables.

2. Metodología

En sus inicios, este programa estaba enfocado a rellenar los datos ausentes mediante estimas calculadas a partir de las series más próximas. Para ello se adaptó el método de Paulhus y Kohler (1952) para rellenar precipitaciones diarias mediante promedios de valores de alrededor, normalizados mediante división por sus respectivas precipitaciones medias. Este método se escogió por su simplicidad y por permitir el uso de series próximas aunque no dispongan de un periodo común de observación con la serie problema, cosa que no permitiría ajustar modelos de regresión.

Además de normalizar los datos mediante división por sus valores medios, *Climatol* ofrece también hacerlo restando las medias o mediante una estandarización completa. Así, denominando m_X y s_X a la media y desviación típica de una serie X, tenemos estas opciones para su normalización:

1. Restar la media: $x = X - m_X$
2. Dividir por la media: $x = X/m_X$
3. Estandarizar: $x = (X - m_X)/s_X$

El principal problema de esta metodología es que las medias (y desviaciones típicas en el tercer caso) de las series en el periodo de estudio no se conocen si las series no están completas, que es lo más frecuente en las bases de datos reales. Entonces *Climatol* calcula primero estos parámetros con los datos disponibles en cada serie, rellena los datos ausentes usando estas medias y desviaciones típicas provisionales, y vuelve a calcularlas con las series rellenadas. Después se vuelven a calcular los datos inicialmente ausentes usando los nuevos parámetros, lo que dará lugar a nuevas medias y desviaciones típicas, repitiendo el proceso hasta que ninguna media cambia al redondearla con la precisión inicial de los datos.

Una vez estabilizadas las medias, se normalizan todos los datos y se procede a estimarlos (tanto si existen como si no, en todas las series) mediante la sencilla expresión:

$$\hat{y} = \frac{\sum_{j=1}^{j=n} w_j x_j}{\sum_{j=1}^{j=n} w_j}$$

en la que $\hat{y}$ es un dato estimado mediante los correspondientes n datos x_j más próximos disponibles en paso temporal, y w_j es el peso asignado a cada uno de ellos.

Estadísticamente, $\hat{y}_i = x_i$ es un modelo de regresión lineal denominado *Eje Mayor Reducido* o *Regresión Ortogonal*, en el que la recta se ajusta minimizando las distancias de los puntos medidas en dirección perpendicular a la misma (regresión tipo II) en lugar de en dirección vertical (regresión tipo I) como se hace generalmente (figura 1), cuya formulación (con series normalizadas) es $\hat{y}_i = r \cdot x_i$, siendo r el coeficiente de correlación entre las series x e y. Nótese que este tipo de ajuste se basa en la presunción de que la variable independiente x se mide sin error (Sokal y Rohlf, 1969), presunción que no se sostiene cuando ambas son series climáticas.

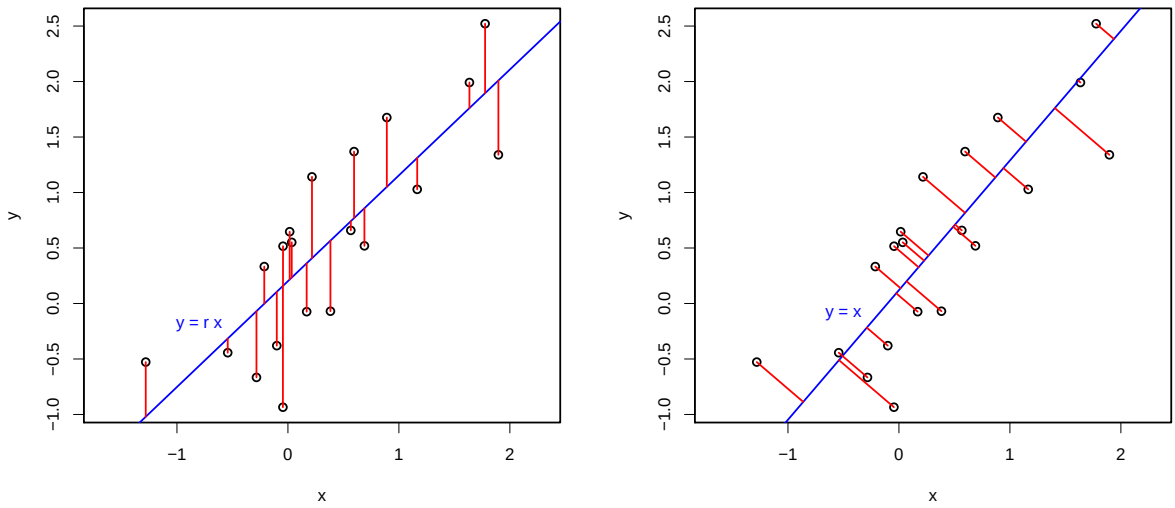

Figura 1: En rojo, desviaciones de la recta de regresión lineal (azul) minimizadas por mínimos cuadrados en los tipos I (izquierda) y II (derecha).

Las series estimadas a partir de las demás sirven como referencias para sus correspondientes series observadas, de forma que el siguiente paso es obtener series de anomalías restando los valores estimados a los observados (siempre en forma normalizada). Estas series de anomalías van a permitir:

- Controlar la calidad de las series y eliminar aquellas anomalías que superen un umbral prefijado.
- Comprobar su homogeneidad mediante la aplicación del *Standard Normal Homogeneity Test* (SNHT: Alexandersson, 1986).

Cuando los máximos valores SNHT de las series son mayores que un umbral predefinido, la serie se divide por el punto de máximo SNHT, pasando todos los datos antes del cambio a una nueva serie que se añade a las demás con las mismas coordenadas pero añadiendo un sufijo numérico al código y al nombre de la estación. Este procedimiento se realiza de forma iterativa, partiendo solo las series con mayores valores SNHT en cada ciclo, hasta que no se encuentren más inhomogeneidades. Además, como SNTH es una prueba originalmente ideada para encontrar un solo punto de ruptura en una serie, la existencia de dos o más saltos en la media de un tamaño similar podría enmascarar sus resultados. Para minimizar este problema, en una primera pasada se aplica SNTH sobre ventanas temporales solapadas, y después en una segunda pasada se aplica SNHT a las series completas, que es cuando la prueba tiene más poder de detección. Finalmente, una tercera pasada se dedica a rellenar todos los datos ausentes en todas las series y sub-series homogéneas con el mismo procedimiento de estimación de datos explicado anteriormente. Por lo tanto, aunque la metodología subyacente del programa es muy simple, su operación se complica a través de una serie de procesos iterativos anidados, como se muestra en el diagrama de flujo mostrado en la figura 2.

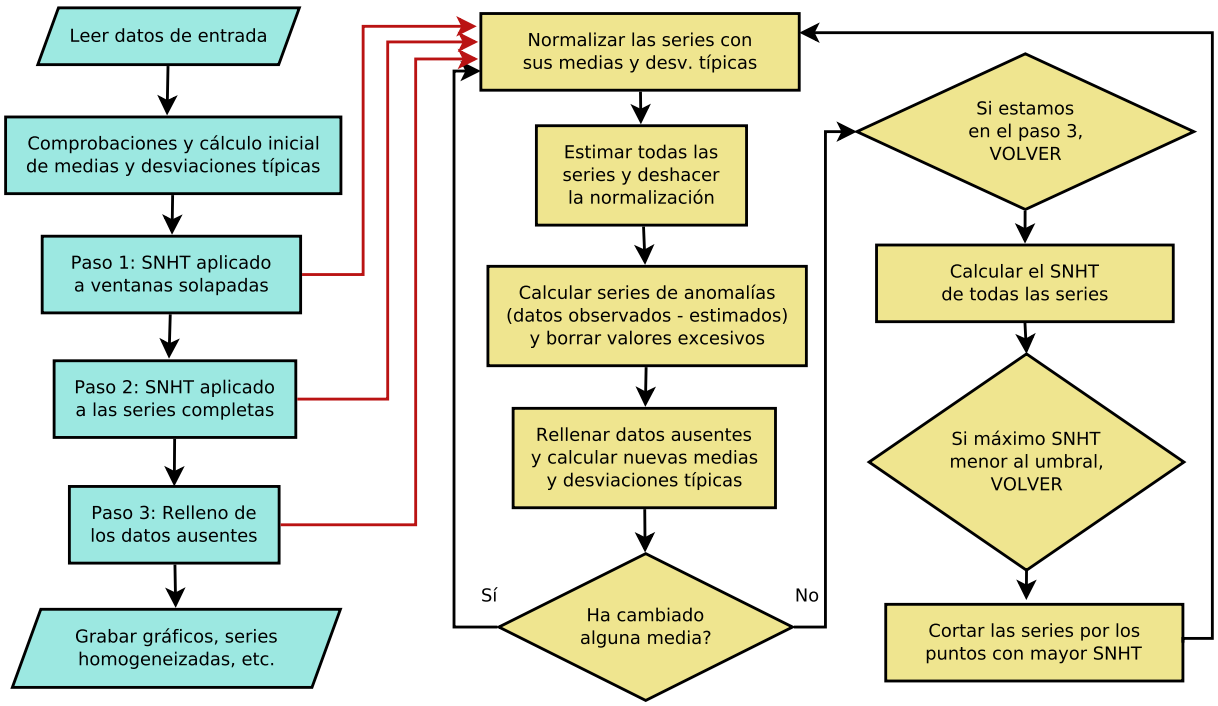

Figura 2: Diagrama de flujo del funcionamiento de *Climatol*, mostrando sus procesos iterativos.

Aunque se han publicado umbrales de SNHT para distintas longitudes de serie y niveles de significación estadística, la experiencia demuestra que esta prueba puede arrojar valores muy diferentes según la variable climática estudiada, el grado de correlación entre las series y su frecuencia temporal. *Climatol* adopta por defecto el valor SNHT = 25, apropiado para valores mensuales de temperatura aunque un poco conservador, tratando de no detectar falsos saltos en la media a costa de dejar pasar los de menor importancia. Sin embargo, para otras variables y en particular para valores diarios, es necesario elevar ese umbral por encima de 100 para evitar un excesivo número de cortes en las series. Lo mismo sucede con el umbral para rechazar datos anómalos, establecido por defecto en 5 desviaciones típicas, pues con datos diarios de precipitación, dada su gran variabilidad espacial, puede ser necesario elevarlo hasta 20 o más. Por todo ello, en lugar de fijar dichos umbrales según niveles de significación, imposibles de establecer con carácter general, se da al usuario la opción de escojerlos subjetivamente, inspeccionando los histogramas de los valores encontrados tras una primera aplicación de *Climatol* a su problema concreto.

3. Procedimientos de homogeneización

Tras haber expuesto la metodología seguida por el paquete *Climatol*, esta sección se dedicará a ilustrar su aplicación práctica a través de algunos ejemplos.

3.1. 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. Además, para evitar complicaciones, el periodo de estudio debería abarcar años completos, comenzando en enero (el día 1 si son datos diarios) del año inicial y terminando en diciembre (el día 31 en el caso de datos diarios) del año final. 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, de modo que los usuarios de Windows pueden asociar abrirlos con el bloc de notas u otro editor de texto plano. (Si se editan con LibreOffice o Word, téngase cuidado de grabarlos como texto sencillo para evitar problemas).

Solo con el propósito de realizar los ejemplos que siguen, estos archivos se pueden generar en el directorio de trabajo por medio de estas órdenes (todo lo que sigue a # son comentarios):

```
library(climatol) # cargar las funciones del paquete
data(Ttest) #cargar los datos de ejemplo en memoria
write(dat, 'Ttest_1981-2000.dat') #grabar los datos
#grabar el fichero de estaciones:
write.table(est.c, 'Ttest_1981-2000.est', row.names=FALSE, col.names=FALSE)
rm(dat, est.c) #borrar los datos cargados en memoria
```

Estos archivos contienen 20 años de temperaturas diarias de prueba de 12 estaciones inventadas. Se pueden inspeccionar para ver su estructura. Las primeras líneas del fichero de estaciones `Ttest_1981-2000.est` son:

```
-108.035 44.38 1169.5 "WY003" "Small Horn"
-108.9006 44.4139 1599.6 "WY018" "Narrow Canyon"
-108.5931 44.8919 1251.2 "WY020" "Wide Meadows"
-108.3906 44.4972 1355.8 "WY027" "Greenbull"
```

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 (no en grados, minutos y segundos) y con el signo adecuado para indicar Oeste, Este, Norte o Sur. Z es la altitud en metros.

Las primeras líneas del fichero de datos `Ttest_1981-2000.dat` son:

```
-1.8 2.7 0.4 8 2.4
 1.4 1.2 3.3 1.5 0.7
-0.8 -0.6 4 2.6 -1.6
-4.8 -3.1 -0.8 -0.6 -4
```

Estos 20 datos son las temperaturas medias de los primeros 20 días de enero de 1981 en la primera estación (Small Horn). Las siguientes líneas del fichero contienen el resto de datos de esta estación hasta el 31 de diciembre de 2000, seguidos por todos los datos de las otras estaciones relacionadas en el fichero `Ttest_1981-2000.est`.

Para ayudar en la preparación de los ficheros de entrada con este formato, *Climatol* provee algunas funciones útiles (ver la documentación de R para más detalles sobre su uso):

- `db2dat` genera los ficheros extrayendo las series de una base de datos a través de una conexión ODBC.
- `daily2climatol` puede usarse cuando cada estación tiene los datos diarios almacenados en archivos individuales.
- `rclimdex2climatol` puede convertir ficheros en formato RClimDex.

3.2. Primer análisis exploratorio de los datos

La función de homogeneización de *Climatol* se llama `homogen`, y su aplicación más trivial solo requiere especificar tres parámetros: el acrónimo de la variable, y los años inicial y final del periodo de estudio:

```
homogen('Ttest', 1981, 2000)
```

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. Pero como se explica en la sección de metodología, 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('Ttest', 1981, 2000, expl=TRUE)
```

Ahora podemos abrir el archivo de salida `Ttest_1981-2000.pdf` para revisar sus diferentes gráficos de diagnóstico. Primero vemos la disponibilidad de datos, en todas las estaciones y globalmente (figura 3). 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. En este caso se deberían añadir nuevas series que tuvieran datos en esos momentos "huérfanos", o reducirse el período de estudio para evitar esa condición.

Cuando se trabaja con variables limitadas por cero y con una distribución de probabilidad sesgada (como la precipitación o la velocidad del viento), la normalización por proporción respecto a la media (establecida con `std=2`) es preferible a la estandarización por defecto.

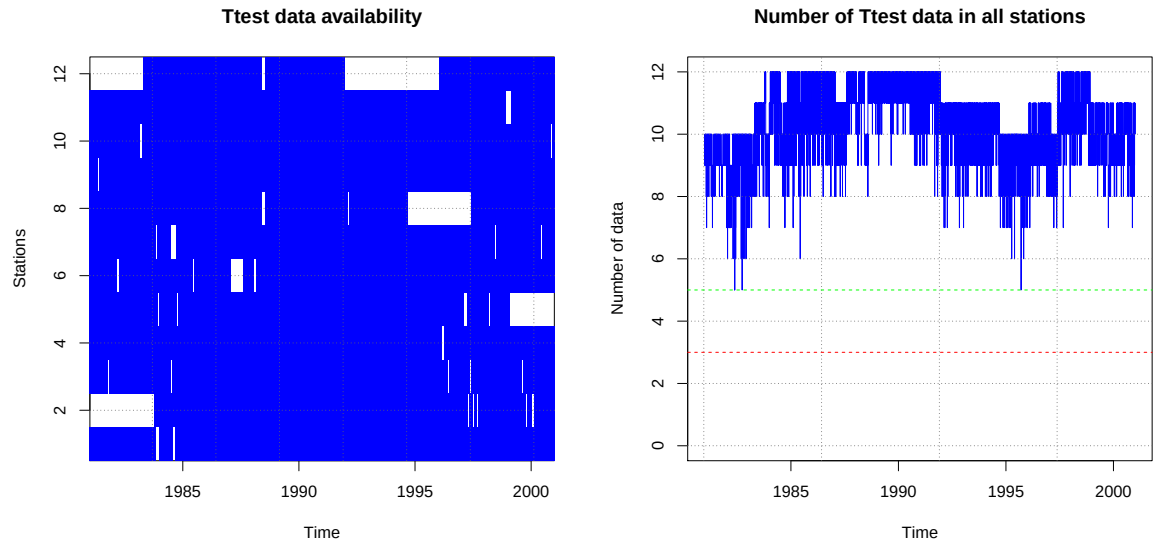

Figura 3: Disponibilidad de datos, por estaciones (izquierda) y globalmente (derecha).

Es importante ejecutar estos análisis exploratorios sobre los datos originales para un control de calidad confiable, ya que la detección de valores atípicos en series derivadas puede enmascarar los errores de observación. Por ejemplo, si hay un error de 10 °C en una temperatura máxima diaria, se reducirá a 5 °C en la media diaria si se calcula como $(T_{max} + T_{min})/2$, y a alrededor de 10/30 = 0.33 °C en la máxima media mensual o 5/30 = 0.17 °C en la media mensual.

Los siguientes gráficos muestran diagramas de caja de los datos en cada estación y un histograma del conjunto de todos los datos (figura 4). 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 (figura 5). 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. En áreas de topografía compleja y/o baja densidad de estaciones, las correlaciones pueden estar lejos de ser óptimas. En esta situación, los datos rellenados se verán afectados individualmente por errores importantes, pero es de esperar que sus propiedades estadísticas generales sean aceptables.

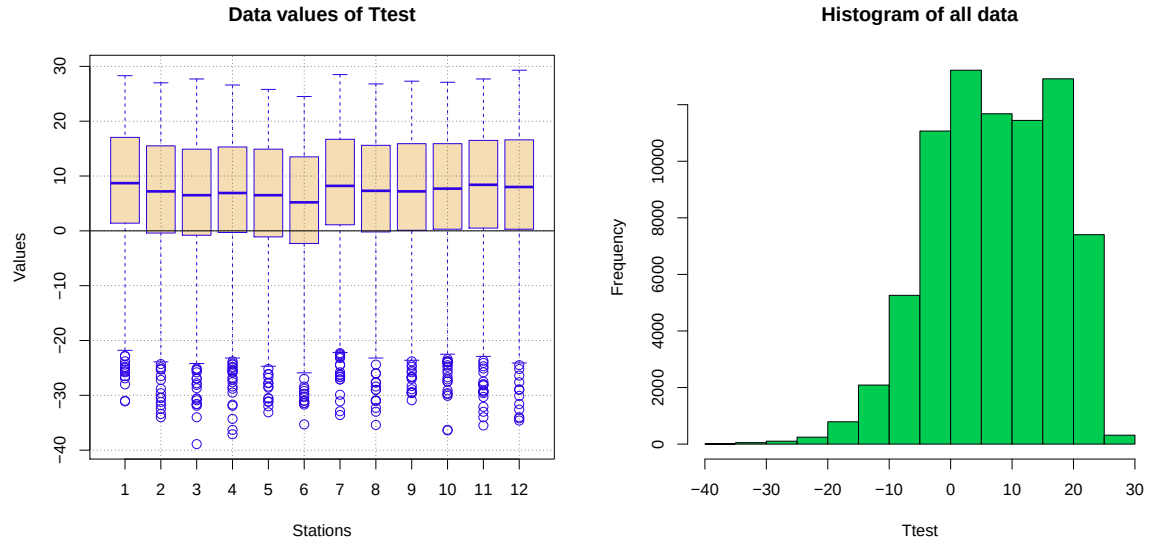

Figura 4: diagramas de caja de los datos en cada estación (izquierda) e histograma de todos los datos (derecha).

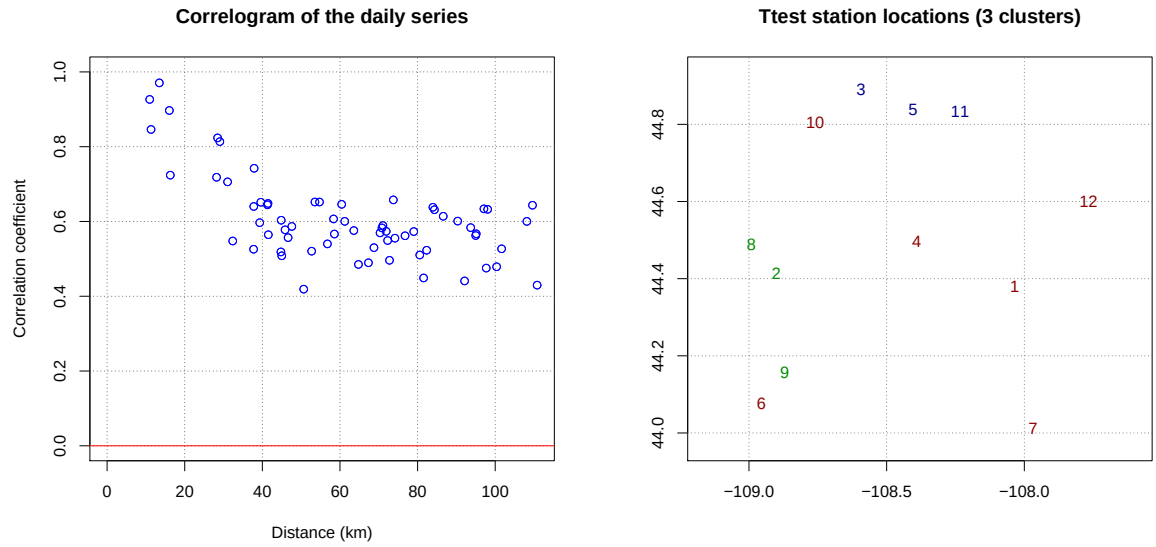

Figura 5: Correlograma de las series (izquierda) y mapa de las estaciones (derecha; los colores identifican grupos de estaciones con variabilidad similar).

Para evitar el procesamiento de matrices de correlación demasiado grandes, el número de series utilizado para este análisis de conglomerados está limitado por defecto a 100, y se utiliza una muestra aleatoria de este tamaño cuando el número de series es mayor, pero el usuario puede modificar este número.

Después de estos gráficos iniciales dedicados a verificar los datos, las siguientes páginas del documento muestran gráficos de anomalías estandarizadas. Cuando se opera normalmente, estos gráficos se muestran para cada una de las etapas: 1, detección en ventanas escalonadas superpuestas; 2, detección en las series completas; y 3, anomalías finales de las series homogeneizadas. Las gráficas de las dos primeras etapas muestran las series de anomalías de las inhomogeneidades detectadas, marcando los puntos de ruptura por donde son cortadas, pero en

este modo exploratorio se omiten las dos primeras etapas, y solo se muestran las anomalías de todas las series originales.

La figura 6 muestra dos de estos gráficos. La serie de la izquierda parece bastante homogénea, con un SNHT máximo de 12 sobre ventanas escalonadas superpuestas marcadas en verde sobre una línea a trazos del mismo color en el punto donde se alcanza ese máximo, y un SNHT máximo de 17 en toda la serie debajo de una línea negra en su paso de tiempo correspondiente. Por el contrario, la serie de la derecha es claramente heterogénea, con SNHT máximos de 117 y 1561 alcanzados en el mismo punto. Dos líneas adicionales en la parte inferior informan sobre la distancia mínima de los datos vecinos (en verde) y el número de datos de referencia utilizados (en naranja), ambos utilizando la escala logarítmica del eje derecho.

Después de los gráficos de anomalías se encuentran las gráficas de las series ajustadas y las correcciones aplicadas, pero como no se realizan modificaciones en las series en modo exploratorio, aparte de completar todos los datos ausentes, estas gráficas se explicarán más adelante.

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 (figura 7) ayuda a elegir umbrales adecuados para rechazar datos muy anómalos, suponiendo que son errores y pueden eliminarse. Nuestro histograma de ejemplo muestra algo de sesgo a la izquierda, pero no es muy pronunciado, y por tanto podrían aceptarse todos los datos configurando `dz.max=9`, ya que el valor predeterminado eliminaría los datos con anomalías absolutas superiores a 5 desviaciones típicas.

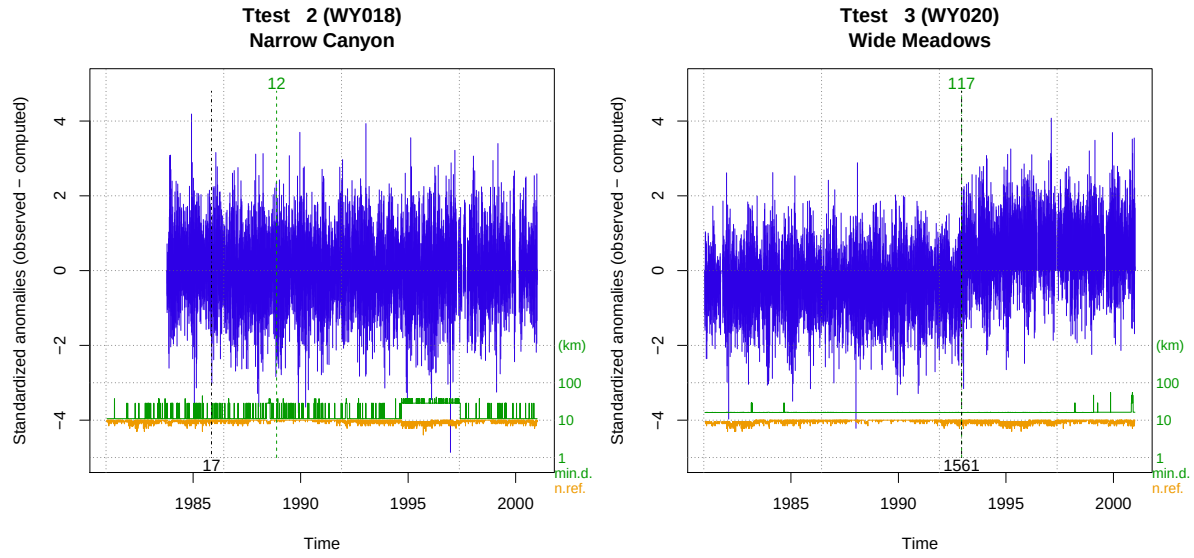

Figura 6: Anomalías de una serie homogénea (izquierda) y una muy inhomogénea (derecha).

Los histogramas de SNHT máximo (ya sea en ventanas o completos) tienen como objetivo elegir los umbrales de detección de cambios en el promedio de las series. Si estuviéramos procesando un gran número de series, estos histogramas mostrarían una alta frecuencia de valores bajos, correspondientes a series bastante homogéneas, y uno o más grupos secundarios de barras debidos a casos no homogéneos. Cuando hay una separación (o un mínimo claro) entre estas condiciones, es muy fácil establecer un valor entre ellas como umbral para las etapas de detección. En nuestro caso, con solo 12 series, las barras de frecuencia están separadas en varios sitios, lo que dificulta la decisión. Para la etapa de ventanas solapadas, la configuración

`snht1=60` parece razonable, pero está lejos de ser clara en el histograma de SNHT aplicado en las series completas. En este caso, la inspección visual de los diagramas de anomalías puede ayudar a elegir `snht2=70` como valor adecuado.

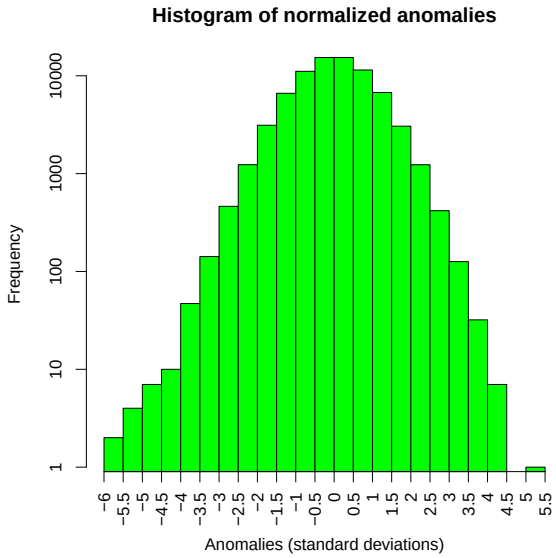

Figura 7: Histograma de anomalías (todos los datos conjuntamente).

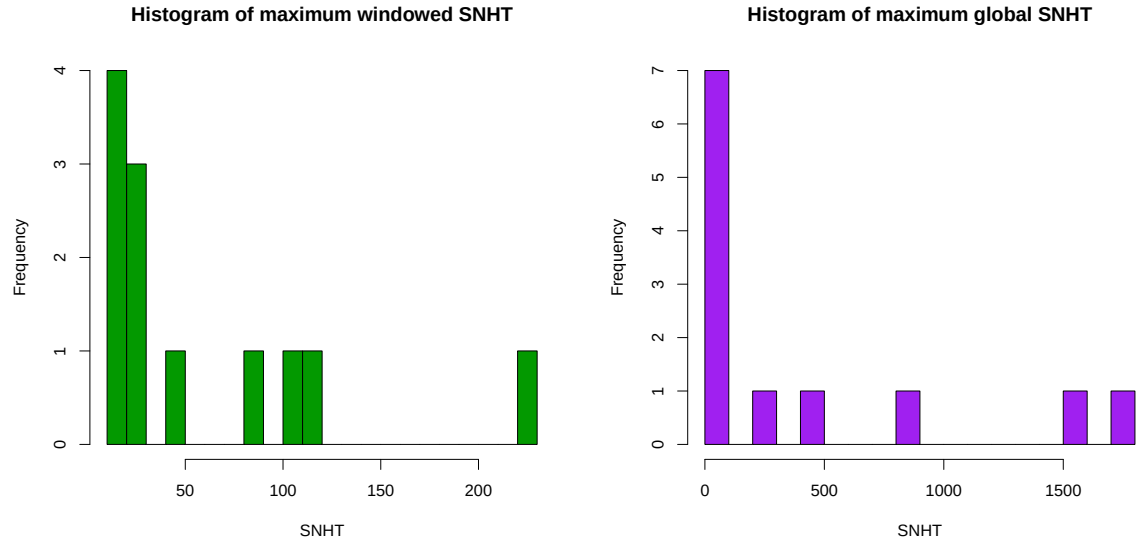

Figura 8: Histogramas de los valores máximos de SNHT encontrados en ventanas escalonadas superpuestas (izquierda) y en las series completas (derecha).

La última página del documento muestra un diagrama de números de estación (su orden en el archivo `Ttest_1981-2000.est`) de acuerdo con sus errores típicos (RMSE por sus siglas en inglés) finales y los valores de SNHT (figura 9). 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.

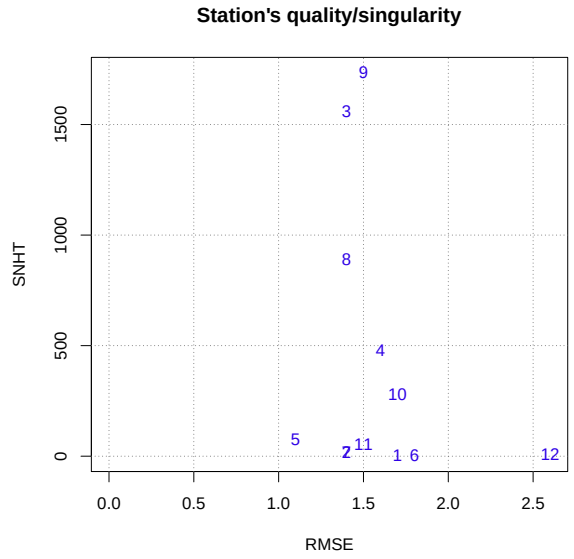

Figura 9: Gráfico de la calidad/singularidad de las series finales.

Después de todas estas consideraciones, procederíamos a homogeneizar la serie aplicando:

```
homogen('Ttest', 1981, 2000, dz.max=9, snht1=60, snht2=70)
```

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. *Climatol* ayuda a obtener datos mensuales de la serie diaria mediante la función `dd2m`, que podemos aplicar aquí de esta manera:

```
#(Con precipitaciones, añadir el parámetro valm=1 para
# calcular totales mensuales en lugar de valores medios)
dd2m('Ttest', 1981, 2000)
```

Esta orden guarda en `Ttest-m_1981-2000.dat` y `Ttest-m_1981-2000.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).

3.3. Homogeneización de las series mensuales

Si el usuario está trabajando con datos mensuales, no hay necesidad de usar el sufijo `-m`, pero aquí vamos a homogeneizar las series mensuales que se obtuvieron a partir de los valores diarios en el apartado anterior. Podríamos comenzar con una aplicación exploratoria de `homogen` como hicimos con los datos diarios, pero aquí probaremos la función con sus valores predeterminados:

```
homogen('Ttest-m', 1981, 2000)
```

La inspección de los gráficos de salida `Ttest-m_1981-2000.pdf` revela que los valores predeterminados de `snht1=snht2=25` parecen apropiados para los valores mensuales. La mayoría de los gráficos ya se han discutido anteriormente. La única diferencia es que ahora tenemos gráficos de anomalías para las etapas de detección 1 y 2, donde los cambios detectados en la media están marcados en rojo (ver un ejemplo en la figura 10 izquierda). Los histogramas de SNHT al final de estas etapas se refieren a los valores de la prueba después de que las series se hayan dividido en los puntos de ruptura detectados.

Después de los gráficos de anomalías de la etapa 3, los "gráficos finales" ilustran la reconstrucción de series completas a partir de cada sub-periodo homogéneo. La figura 10 (derecha) muestra un ejemplo de una serie que se dividió en dos fragmentos. La parte superior del gráfico traza las medias anuales móviles de las series reconstruidas, con los datos originales en negro y los rellenados en diferentes colores para cada serie resultante. (Téngase en cuenta que en presencia de datos faltantes, las medias móviles de los datos originales que no se puedan calcular no aparecerán en el gráfico). La parte inferior muestra las correcciones aplicadas a las series, trazadas en diferentes colores. Como se puede ver, las correcciones presentan variaciones estacionales (se pueden lograr correcciones constantes en este caso si no se usan las desviaciones típicas en la normalización estableciendo `std=1`), y las puntas se deben a rechazos de valores atípicos.

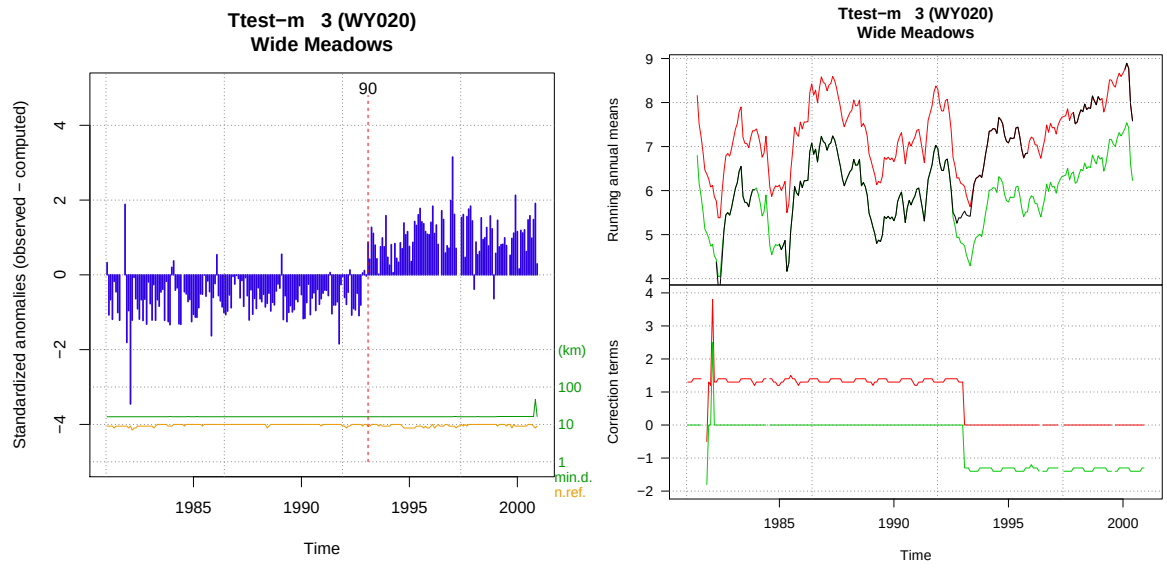

Figura 10: Ejemplo de detección de un cambio en la media de una serie con SNHT=90 (izquierda) y reconstrucción de series completas para ambos sub-periodos homogéneos (derecha).

Después de ejecutar con éxito la función `homogen`, el usuario puede encontrar los siguientes archivos en su directorio de trabajo R (sin el sufijo `-m` si la serie original tuvo una periodicidad mensual o mayor):

- `Ttest-m_1981-2000.txt` : Fichero de texto con todos los mensajes emitidos en la consola durante el proceso. Incluye los grupos de estaciones y resúmenes finales de los valores de SNHT and RMSE de las series resultantes.

- `Ttest-m_1981-2000_out.csv` : Fichero de texto (valores separados por comas, CSV) con la lista de valores atípicos corregidos. Nótese que los valores sugeridos ("Suggested") solo son primeras estimas en el momento del rechazo del valor. Por tanto, los valores finales pueden diferir, e incluso haber más de uno (cuando las series se han cortado).
- `Ttest-m_1981-2000_brk.csv` : Fichero de texto (CSV) con la lista de puntos de corte y sus correspondientes valores SNHT.
- `Ttest-m_1981-2000.pdf` : Los gráficos de diagnóstico comentados previamente.
- `Ttest-m_1981-2000.rda` : Fichero binario de R conteniendo los resultados de la homogeneización. (Ver la documentación de `homogen` para más detalles).

Cuando el usuario tiene acceso directo a los datos originales, vale la pena inspeccionar la lista de valores atípicos rechazados en el archivo `Ttest-m_1981-2000_out.csv` y verificar si son errores o valores creíbles. Después de haber corregido la base de datos, se pueden volver a compilar los archivos de entrada para *Climatol* y repetir todo el procedimiento.

Además, si hay metadatos sobre cambios históricos en los observatorios, es muy conveniente editar el archivo `Ttest-m_1981-2000_brk.csv` para ajustar las fechas de los puntos de corte detectados a los de los eventos que puede haber alterado las observaciones y ejecutar de nuevo la función `homogen` con el parámetro `metad=TRUE` (y `sufbrk=''` si las series originales estaban compuestas de datos mensuales). Pero téngase en cuenta que no todos los cambios deben necesariamente tener un impacto en la variable climática estudiada, y que lo más frecuente es que los meta-datos estén incompletos o se carezca de ellos.

3.4. Ajuste de las series diarias con los puntos de corte mensuales

Si los datos iniciales eran diarios en lugar de mensuales, procederemos a su ajuste empleando los puntos de corte detectados con los agregados mensuales, para lo que aplicaremos nuevamente la función `homogen` con el parámetro `metad=TRUE`:

```
homogen('Ttest', 1981, 2000, dz.max=7, metad=TRUE)
```

De esta forma, `homogen` omite las dos etapas de detección y procede a dividir la serie diaria por los puntos de corte listados en el archivo `Ttest-m_1981-2000_brk.csv`, y luego pasa directamente a la tercera etapa de reconstrucción de todas las series a partir de sus sub-periodos homogéneos por medio de la rutina de relleno. Este proceso crea los archivos de salida habituales, excepto el `Ttest_1981-2000_brk.csv`, ya que esta vez no se ha realizado ninguna detección de saltos en la media.

4. Obtención de productos con los datos homogeneizados

El usuario puede cargar los resultados de la homogeneización en la memoria de trabajo de R para su posterior procesado manual mediante la orden:

```
load('Ttest_1981-2000.rda')
```

Pero *Climatol* provee las 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, que se pueden generar con:

```
#(Con precipitaciones, añadir el parámetro valm=1 para
# calcular totales mensuales en lugar de valores medios)
dd2m('Ttest', 1981, 2000, homog=TRUE)
```

Con el parámetro `homog=TRUE` se generarán agregados mensuales a partir de las series homogeneizadas, y no con las series originales como se hizo anteriormente. Ahora los nuevos ficheros creados son `Ttest-mh_1981-2000.dat` y `Ttest-mh_1981-2000.est`, que contienen los agregados mensuales de las series diarias ajustadas.

4.1. Series homogeneizadas y resúmenes estadísticos

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

```
dahstat('Ttest', 1981, 2000, stat='series')
```

Uno de los ficheros, `Ttest_1981-2000_series.csv`, contiene todas las series homogeneizadas, y el otro, `Ttest_1981-2000_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('Ttest',1981,2000) #medias de las series diarias
dahstat('Ttest',1981,2000,mh=TRUE) #medias de los valores mensuales
dahstat('Ttest',1981,2000,mh=TRUE,stat='tnd') #tendencias y p-valores
dahstat('Ttest',1981,2000,stat='q',prob=.2) #primer quintil (diarios)
```

Esta función incluye parámetros para escojer un subconjunto de las series, bien dando una lista con los códigos deseados (como con `cod=c('WY020','WY055')` o especificando que queremos las series reconstruidas desde el último sub-periodo homogéneo (`last=TRUE`), desde el subperiodo más largo (`long=TRUE`), etc.

4.2. Series de rejillas homogeneizadas

La otra función de post-proceso, `dahgrid`, provee rejillas calculadas a partir de las series homogeneizadas (sin usar datos rellenados). Pero antes de aplicar esta función, el usuario debe definir los límites y la resolución de la rejilla, como en este ejemplo:

```
grd=expand.grid(x=seq(-109,-107.7,.02), y=seq(44,45,.02)) #rejilla deseada
library(sp) #paquete necesario para la siguiente orden:
coordinates(grd) <- ~ x+y #convertir la rejilla en un objeto espacial
```

La función de R `expand.grid` se ha usado para definir las secuencia de coordenadas X e Y, y luego se aplica `coordinates` (del paquete `sp`) para convertir la rejilla, guardada como `grd` (se podría haber usado cualquier otro nombre), en un objeto de clase espacial.

Ahora las rejillas pueden generarse (en formato NetCDF) con:

```
dahgrid('Ttest', 1981, 2000, grid=grd) #grids with daily time steps
dahgrid('Ttest', 1981, 2000, grid=grd, mh=TRUE) #id. with monthly steps
```

Estas rejillas se han construido con valores normalizados, adimensionales. Se pueden obtener nuevas rejillas con las temperaturas en °C por medio de herramientas externas, como las *Climate Data Operators* (CDO):

```
#Esto no son órdenes de R! Ejemplo para una terminal linux/unix:
cdo add -mul Ttest-mh_1981-2000.nc Ttest-mh_1981-2000_s.nc \
  Ttest-mh_1981-2000_m.nc Ttest-mu_1981-2000.nc
```

Pero las nuevas rejillas contenidas en `Ttest-mu_1981-2000.nc` solo estarán basadas en interpolaciones geométricas, y por tanto deberían obtenerse mejores rejillas de medias y desviaciones típicas en `Ttest-mh_1981-2000_m.nc` y `Ttest-mh_1981-2000_s.nc` mediante métodos geoestadísticos antes de usarlas para deshacer la normalización de las rejillas proporcionadas por `dahgrid`.

5. Recetas adicionales

Los ejemplos anteriores muestran y comentan las aplicaciones más frequentes de las funciones de *Climatol*. No obstante, pueden surgir preguntas relativas a cómo proceder cuando tratamos con otras variables climáticas o resoluciones temporales. Esta sección está dedicada a responder esas preguntas, y pueden añadirse más respuestas en el futuro fruto de la interacción con los usuarios.

5.1. Cómo modificar los pesos y el número de referencias

Los pesos w_j dados a los datos próximos para estimar los valores de las series dependen de las distancias d_j a través de la función $w_j = 1/(1+d_j^2/h^2)$, donde h es la distancia a la que el peso se reduce a la mitad. Por defecto, $h = 100$ km, pero se puede cambiar asignando otro valor al parámetro `wd` (distancia de ponderación), que es como se llama h dentro de la función `homogen`. Por defecto, `wd=0` en las dos primeras etapas de detección, indicando que no se van a aplicar pesos diferentes a los datos, puesto que correríamos el peligro de asignar mucho peso a una estación muy próxima pero potencialmente inhomogénea. Pero el usuario puede especificar `wd` para las tres etapas, como fijando `wd=c(0,1000,25)`, que no pondera los datos en la primera etapa, asigna $h = 1000$ a la segunda y $h = 25$ a la tercera. La figura 11 muestra cómo varían los pesos en función de la distancia para diferentes valores de h (`=wd`).

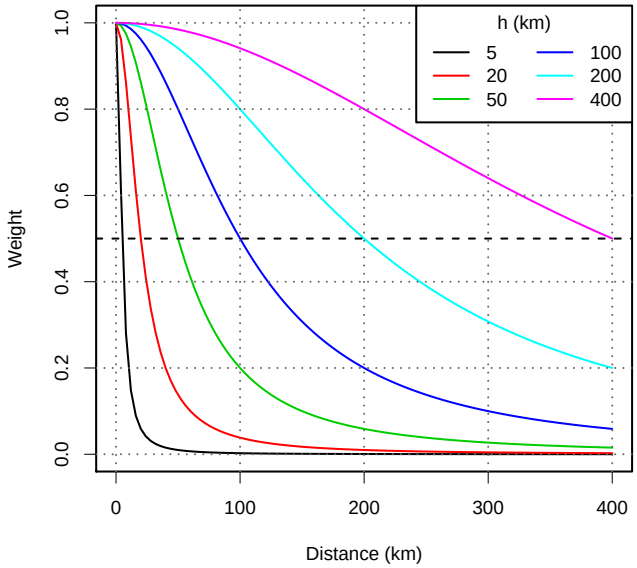

Figura 11: Variación de los pesos para distintos valores de h (parámetro `wd`).

En cuanto al número de datos más próximos usados en cada paso temporal, por defecto se usan hasta 10 (si los hay) en las etapas de detección, y 4 en la última etapa de reconstrucción de series. Esto se puede cambiar con el parámetro `nref`, como en `nref=c(8,8,2)`.

Los parámetros escogidos pueden ser óptimos o no dependiendo del objetivo final del análisis de las series. Por ejemplo, si se quiere obtener normales climáticas, los ajustes de la varianza no tienen importancia, mientras que serán cruciales para calcular periodos de retorno de

valores extremos. En el último caso se puede limitar la disminución de las estimas ponderadas estableciendo distancias ponderales más cortas, especialmente en la tercera etapa (ejemplo: `wd=c(100,100,15)`), y/o reduciendo el número de referencias, e incluso usando solo una, en la última etapa (`nref=c(5, 5, 1)`), como puede ser preferible al ajustar precipitaciones diarias.

5.2. Cómo guardar los resultados de diferentes pruebas

Si se ejecuta `homogen` con diferentes parámetros para explorar cuáles dan mejores resultados, puede evitarse sobre-escribir las salidas anteriores renombrándolas con la ayuda de la función `outrename`. Por ejemplo, la orden

```
outrename('Ttest-m', 1981, 2000, 'old')
```

renombrará todos los ficheros de salida `Ttest-m_1981-2000*` a `Ttest-m-old_1981-2000*`.

5.3. Cómo cambiar el nivel de corte en el análisis de agrupamiento

Climatol aplica un análisis de agrupamiento en su comprobación inicial de los datos, pero el número de grupos se determina automáticamente. Mirando el dendrograma cerca del principio del documento de salida en PDF se puede elegir un nivel de corte mejor. En el ejemplo de la figura 12 se generan tres grupos de estaciones al cortar por el nivel de disimilaridad 0.058. Pero podríamos preferir cortar el dendrograma por el valor 0.04 para obtener cinco grupos. En ese caso, lo único que hay que hacer es repetir la orden de homgeneización añadiendo el parámetro `cutlev=0.04`. Por defecto se usan hasta 100 series en el análisis de agrupamiento, y se extrae una muestra aleatoria de ese tamaño si el número de series estudiado excede ese límite. Pero si ese número no es muy grande, se puede forzar a usarlas todas estableciendo, por ejemplo, `nclust=136`. Nótese también que el número de grupos será de nueve como máximo..

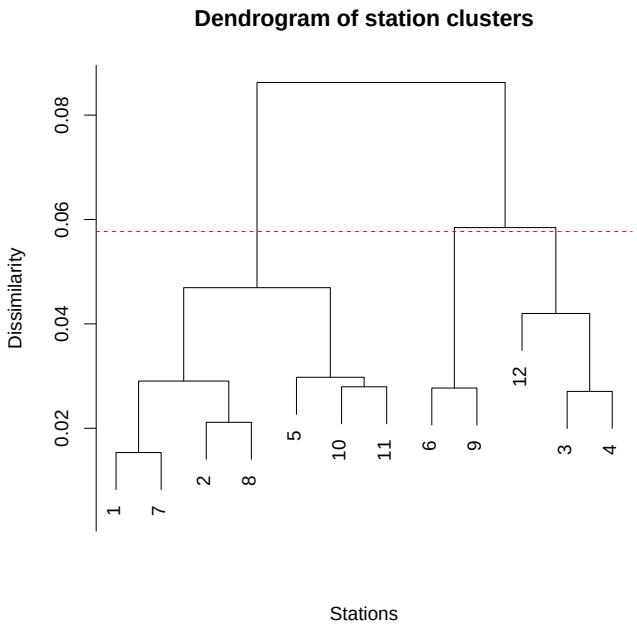

Figura 12: Dendrograma de las estaciones, basado en sus coeficientes de correlación.

5.4. Las coordenadas de mis estaciones son UTM

Climatol supone que las coordenadas están en grados si los valores absolutos de X e Y no superan 180 y 90 respectivamente. De lo contrario, si la media de X o Y es mayor que 10000 supondrá que están en metros, y las convertirá a kilómetros para el resto del proceso.

5.5. Cómo aplicar una transformación a mis datos sesgados

La función `homogen` puede aplicar transformaciones $log(x+1)$ (`trf=1`) o de cualquier raíz (`trf=2` para raíz cuadrada, `trf=3` para la cúbica, etc. Se admiten valores fraccionarios). Reduciendo suficientemente su sesgo, los datos se podrían estandarizar (`std=3`, la opción por defecto), pero los resultados de pruebas intensivas con series de precipitación mensual durante el proyecto MULTITEST mostraron resultados claramente mejores al normalizar los datos dividiéndolos por sus valores medios (`std=2`) sin necesidad de transformarlos.

5.6. Cómo limitar los valores posibles de una variable

Se pueden usar los parámetros `vmin` and `vmax` para que `homogen` limite el rango de valores posibles. Esto puede ser útil si tratamos humedades relativas (poner `vmin=0` y `vmax=100`) o cualquier otra variable con un rango truncado de posibles valores. Nótese que cuando se usa la normalización `std=2`, automáticamente se establece `vmin=0`, porque la esa normalización normalmente se aplica a variables con la precipitación o la velocidad del viento, que no pueden tener valores negativos.

5.7. ¿Pueden usarse salidas de reanálisis como series de referencia?

Cuando los datos están muy fragmentados y algunos pasos temporales de nuestro periodo de estudio no tienen datos en ninguna serie, una posible solución es usar series procedentes de productos de reanálisis para servir como referencias que provean datos en esos huecos críticos. Esas series deben estar correlacionadas positivamente con nuestra variable. Ejemplo: Las temperaturas a 2 m sobre el suelo pueden añadirse a nuestras series termométricas, pero si no están disponibles también podrían servir los espesores geopotenciales próximos a la superficie o variables similares. Las series de precipitación pueden no tener equivalencias en los reanálisis, y entonces podría probarse alguna variable derivada (¿advección de vorticidad? ¿velocidad vertical? ¿una combinación de ambas?), pero su correlación con nuestras precipitaciones debería ser comprobada antes de usarlas como referencias.

Aunque la aparición de nuevos sistemas de observación (como los satélites) introduce inhomogeneidades en la cantidad de datos disponibles para su asimilación por los modelos, podemos considerar que los productos de reanálisis son más homogéneos en general que las series observacionales. Para usar estos productos como referencias, las series de uno o más puntos de rejilla localizados en el dominio de estudio deben añadirse al fichero de datos `*.dat`, y las coordenadas de esos puntos añadirse al fichero de estaciones `*.est`. Sus códigos deben empezar por

un asterisco (ejemplo: `*R43`) para que los controles de calidad y homogeneidad se salten esas series más confiables.

5.8. ¿Con qué series cortadas debería quedarme?

La mayoría de métodos de homogeneización devuelven las series ajustadas desde el último sub-periodo homogéneo, pero *Climatol* genera reconstrucciones completas de cada sub-periodo (a no ser que sea demasiado corto para que dicha reconstrucción sea fiable). Por tanto, el usuario puede preguntarse cuál utilizar en su estudio climático. La respuesta depende del objetivo de la investigación. Para obtener valores normales con los que calcular las anomalías de nuevos datos entrantes para monitorización climática, esas normales deberían estar ajustadas al último sub-periodo homogéneo. Pero si el objetivo es crear un mapa de valores normales, deberían usarse los de todas las series, puesto que algunas pueden ajustarse mejor a la variabilidad espacial a la escala del mapa, mientras que otras pueden estar afectadas por microclimas locales y obtendríamos un mapa más ruidoso.

5.9. ¡Tengo tantas series diarias largas que el proceso dura días!

Una posibilidad para acortar el tiempo de cálculo es aplicar `homogen` a sub-áreas, tratando de agrupar aquellas estaciones que compartan los mismos factores climáticos. Si no hay claras discontinuidades climáticas (por ejemplo, crestas montañosas cruzando el dominio de estudio), `homogsplit` puede generar sub-áreas solapadas y homogeneizarlas automáticamente. El usuario solo tiene que suministrar las coordenadas X e Y de división del territorio, teniendo especial cuidado en que ninguna sub-área tenga muy pocas estaciones (aunque pueden estar vacías). También aumenta la posibilidad de que en alguna sub-área haya pasos temporales sin ningún dato en sus estaciones, lo que detendría el proceso. (Esta función puede considerarse experimental).

6. Bibliografía

Aguilar E, Auer I, Brunet M, Peterson TC, Wieringa J (2003): *Guidelines on climate metadata and homogenization*. WCDMP-No. 53, WMO-TD No. 1186. World Meteorological Organization, Geneve.

Alexandersson H (1986): A homogeneity test applied to precipitation data. *Jour. of Climatol.*, 6:661-675.

Khaliq MN, Ouarda TBMJ (2007): On the critical values of the standard normal homogeneity test (SNHT). *Int. J. Climatol.*, 27:681687.

Paulhus JLH, Kohler MA (1952): Interpolation of missing precipitation records. *Month. Weath. Rev.*, 80:129-133.

Peterson TC, Easterling DR, Karl TR, Groisman P, Nicholls N, Plummer N, Torok S, Auer I, Böhm R, Gullett D, Vincent L, Heino R, Tuomenvirta H, Mestre O, Szentimrey T, Salinger J, Førland E, Hanssen-Bauer I, Alexandersson H, Jones P, Parker D (1998): Homogeneity Adjustments of 'In Situ' Atmospheric Climate Data: A Review. *Int. J. Climatol.*, 18:1493-1518.

Sokal RR, Rohlf PJ (1969): *Introduction to Biostatistics*. 2^{nd} edition, 363 pp, W.H. Freeman, New York.

Venema V, Mestre O, Aguilar E, Auer I, Guijarro JA, Domonkos P, Vertacnik G, Szentimrey T, Stepanek P, Zahradnicek P, Viarre J, Müller-Westermeier G, Lakatos M, Williams CN, Menne M, Lindau R, Rasol D, Rustemeier E, Kolokythas K, Marinova T, Andresen L, Acquaotta F, Fratianni S, Cheval S, Klancar M, Brunetti M, Gruber C, Prohom Duran M, Likso T, Esteban P and Brandsma T (2012): Benchmarking homogenization algorithms for monthly data. *Clim. Past*, 8:89-115.