Centro Regional del Clima para el Sur de América del Sur

Centro Regional do Clima para o Sul da América do Sul

Serie Reportes Técnicos – Reporte Técnico CRC-SAS-2013-002

APROXIMACIÓN A LA HOMOGENIZACIÓN DE UNA RED REGIONAL DE SERIES CLIMÁTICAS A RESOLUCIÓN DIARIA

Dr. Enric Aguilar Anfrons

Centro en Cambio Climático, C3

Universidad Rovira i Virgili de Tarragona,

España

Taller sobre técnicas de homogenización de datos climáticos y uso de índices para el monitoreo de sequias. 9-13 de diciembre de 2013. Asunción, Paraguay.

1. Rationale

En las últimas décadas, el foco de atención en el estudio de las ciencias atmosféricas y de la Tierra ha basculado hacia la climatología, debido al efecto sobre los sistemas naturales y socioeconómicos de la variabilidad y el cambio climático. La información climática de base debe ofrecer a los usuarios de productos y servicios climáticos y a la comunidad científica información libre de sesgos artificiales. La implementación del Marco Global para los Servicios Climáticos (GFCS, según sus siglas en inglés), reconoce como capacidad básica de los proveedores de servicios climáticos el generar, gestionar y mantener una red de datos climáticos completa, de calidad y homogénea.

Abundando en los conceptos anteriormente mencionados, entendemos como "completa" aquella base de datos que ha sido sometida a los procesos de Rescate de Datos (DARE, ver www.omm.urv.cat) y la denominaremos "de calidad" si ha superado un proceso de control de calidad (QC), entendido como la supresión de errores puntuales. Finalmente, una serie climática es homogénea cuando sus fluctuaciones se relacionan únicamente con fluctuaciones reales del clima y está libre de sesgos artificiales. No en vano, homogéneo significa "de la misma naturaleza" y el objeto del homogeneizador es conseguir que todas las observaciones de una serie sean comparables. Las causas de inhomogeneidad son numerosas y entre las mismas cabe destacar aquellas que

Figura 1. Proyecto SCREEN. Garita Stevenson operacional y reconstrucción de garita Monstouris decimonónica. El reemplazo en España de la primera garita por la segunda, introdujo inhomogeneidades en las estaciones de la red.

producirán un cambio abrupto en la serie (relocalizaciones, cambios de instrumento, cambios de exposición del instrumento (ver Figura 1), cambios en el cómputo de parámetros, operaciones de mantenimiento, etc.) y aquellas que producirán alteraciones graduales (urbanización, crecimiento de vegetación alrededor del jardín meteorológico, descalibraciones o disfunciones progresivas de los instrumentos de medición, desgaste de la garita meteorológica, etc.)

2. Visión conjunta del proceso de homogeneización

La homogeneización de datos climáticos es una tarea que debe abordarse con posterioridad a los procesos de DARE y QC. Describimos a continuación un proceso cuyo objeto final es la obtención de series climáticas homogeneizadas a resolución diaria de las variables temperatura máxima diaria; temperatura mínima diaria y precipitación acumulada diaria. Nótese la utilización del vocablo “homogeneizadas” en lugar de la palabra “homogéneas”. Mientras que esta última implica la idea de un proceso perfecto, el primero refiere más fielmente la realidad: homogeneizamos, es decir, llevamos hacia un estado homogéneo, minimizamos el impacto de los sesgos artificiales que afectan a nuestras series, aunque siempre contaremos con un error remanente o al menos con una incertidumbre respecto a nuestra corrección.

El proceso de homogeneización se divide en cuatro fases principales:

1. Detección de inhomogeneidades: se realiza sobre promedios anuales y estacionales. Determina los puntos de cambio existentes, que aíslan segmentos homogéneos en cada serie (HSPs).
2. Ajuste del dato mensual: se realiza mediante diferencias (temperatura) o ratios de medias (precipitación) entre los distintos HSPs.
3. Ajuste del dato diario: se realiza aplicando una variedad de técnicas complejas, cuya viabilidad y validez dependen de la densidad y homogeneidad general de la red y de la variable sobre la que trabajamos.
4. Validación de resultados: tras la homogeneización es necesario validar nuestros resultados para asegurar que no hemos introducido artificios mediante nuestras correcciones.

3. Detección de inhomogeneidades y ajuste del dato mensual

El proceso de detección consiste, “simplemente”, en encontrar los puntos de cambio que contiene una serie climática. El entrecomillado anterior pretende llamar la atención sobre la simplicidad del concepto, que al mismo tiempo resulta complejo en su desarrollo. Efectivamente, los puntos de cambio se presentan mezclados con las fluctuaciones de la señal climática, que en ningún modo es estacionaria y con la propia varianza de las series.

Existen dos aproximaciones dominantes para la detección de inhomogeneidades: la segmentación jerárquica y la detección múltiple.

La primera de ellas (RSNHT – http://www.c3.urv.cat/data/manual/manual_rsnht, RhTest – http://etccdi.pacificclimate.org/software.shtml) busca el lugar más probable en el que ubicar un primer punto de cambio (BP en lo sucesivo). Si este no es estadísticamente significativo, el proceso finaliza. De lo contrario, el BP divide la serie en dos HSP. Si alguno de ellos contiene al menos un número predeterminado de datos (por ejemplo, 5), se testea de forma análoga. El proceso de subdivisión continúa mientras se encuentren BPs significativos estadísticamente que produzcan un segmento de suficiente longitud para ser evaluado.

Por el contrario, los métodos de detección múltiple, como los contenidos en HOMER (Mestre et al., 2013), pretenden identificar al mismo tiempo todos los BPs y sus posiciones. Son mucho más complejos matemática y computacionalmente y se basan en la aplicación de un factor de penalización que crece a medida que aumentamos el número de puntos de cambio. De forma muy simplificada, si tenemos una serie con BPs en sus valores 25, 50 y 75, determinará 4 HSPs con promedios distintos: 1-25; 26-50; 51-75; 76-100. Si computamos la raíz cuadrada del error cuadrático medio (RMSE) de dicha serie, será menor si lo hacemos con respecto a las cuatro medias de los 4 HSPs, calculadas exactamente con los datos de cada segmento, que si lo hacemos respecto al promedio conjunto (datos 1 a 100) de la serie. Pero, desafortunadamente, también podría si introducimos puntos de cambio adicionales; aún más, si llevamos el argumento al absurdo de utilizar 100 medias, el error sería "realmente pequeño". Es por ello que es necesario utilizar un factor de penalización que evite la introducción de un número excesivo de puntos de cambio. Actualmente, las aproximaciones de detección múltiple, que son ya posibles computacionalmente, han demostrado ser más eficientes.

Otra cuestión que es necesario tener en cuenta es sobre qué datos se aplicará el test que elijamos. Introducimos aquí dos conceptos: homogeneización absoluta y homogeneización relativa. Para explicar mejor estos conceptos debemos introducir una aproximación a la descomposición de las series climáticas, suponiendo un modelo aditivo, en:

$\mathbf{O_i = C_i + S_i + e_i}$, dónde O_i és una observación climática; C_i es una componente climática regional, común a todas las estaciones de una región climática; S_i una componente de estación y e_i, ruido.

De encontrarnos ante una serie homogénea, S_i es constante; de lo contrario, S_i es constante entre puntos de cambio y la expresión anterior puede transformase en:

$\mathbf{O_i = C_i + S_{ij} + e_i}$, siendo S_{ij} la componente de estación para el HSP j (j = 1,n; n = número de HSPs)

La homogeneización absoluta consiste en aplicar un test a la serie O (es decir, sobre los mismos datos de la estación). Las variaciones temporales en C asociadas a las fluctuaciones reales del clima, estarán mezcladas con las variaciones en S, asociadas a los cambios artificiales en el factor de estación. Si bien permiten detectar grandes inhomogeneidades, la tasa de falsos positivos y falsos negativos será elevada y los factores de ajuste dudosos, debido a la confusión anteriormente mencionada. Por ello, la aproximación denominada relativa es preferida.

La homogeneización relativa se basa en utilizar la diferencia (ratio para variables acumulativas, como la precipitación) entre la serie O y una serie de referencia, R, que comparte la misma señal climática regional, C. En términos estadísticos serán series bien correlacionadas y si ambas series son homogéneas, la diferencia (ratio)

entre ambas se deberá a la diferencia (ratio) entre las componentes de estación, S y tendrá un nivel medio constante, con variaciones determinadas por el ruido e. Existen dos aproximaciones principales para el cálculo de la serie de referencia R:

1. Serie de referencia compuestas: la serie R se computa a partir de un promedio ponderado de diversas estaciones que compartan señal climática con aquella que queremos homogeneizar (en adelante, serie candidata). Los promedios acostumbran a ponderarse por correlación, aunque otras extracciones como la extracción del primer componente principal son igualmente válidas. Se asume que esta serie de referencia es aproximadamente homogénea, puesto que el promediado entre series cancela las inhomogeneidades aleatorias que se produzcan. Así, si incluimos en el promedio 8 series, admitimos que las mismas tendrán distintos puntos de cambio pero que no se producirán en el mismo momento temporal, ni tendrán la misma magnitud, ni signo. Por ello, los BPs que se detecten al testar O-R, se atribuyen a O. Esta premisa – la homogeneidad de la serie de referencia – no se cumple frecuentemente y es necesario realizar procesos iterativos para evitar la propagación de las grandes inhomogeneidades dentro de la red.

2. Comparaciones emparejadas (denominadas en inglés *pairwise comparisons*): la estación candidata se compara con un número predeterminado de estaciones de referencia, entendidas como estaciones individuales que comparten la misma señal climática. En cada emparejamiento O-R_i, no podemos determinar a qué estación corresponde el punto de cambio; pero si este se repite en las distintas comparaciones con referencias individuales, lo atribuiremos a O. Este proceso de decisión puede ser manual (por ejemplo, en PRODIGE (Caussinus y Mestre, 2004), antecesor de HOMER); mixto (HOMER) o automático (aproximación U.S. Historical Climatology Network NOAA).

Aunque ambas aproximaciones para el cómputo de series de referencia son válidas si se aplican correctamente, en la actualidad se piensa que las comparaciones emparejadas ofrecen mejores resultados. No obstante, tanto las referencias compuestas como las comparaciones emparejadas, encuentran dificultades ante inhomogeneidades que se presentan cuasi-simultáneamente a lo largo de una red e introducen un cambio de magnitud y signo similar. Sirva como ejemplo las sustituciones en años recientes de estaciones convencionales por estaciones automáticas en numerosas redes o la introducción de la garita Stevenson reemplazando stands abiertos a finales del siglo XIX y principios el XX.

Para estos casos, la denominada homogeneización directa, basada en la comparación de medidas emparejadas entre el estado A (i.e., estación automática; garita Stevenson) y el estado B (i.e. estación convencional; garita abierta) permite derivar ajustes específicos.

La detección de inhomogeneidades no debe descansar sólo en el test estadístico, sino que debe apoyarse en el estudio de los metadatos de la estación, es decir la información documental sobre los cambios que ocurrieron en un observatorio y también en la propia inspección visual de las series.

Una vez detectados los BP existentes, el ajuste es relativamente sencillo, y se limita utilizar como factores de corrección las diferencias O-R (ratios, O/P) entre distintos HSPs. La detección y ajuste de datos mensuales puede realizarse de forma automática aunque ello solo se recomienda para redes de gran tamaño que deban

homogeneizarse repetidamente ante la ingesta de nuevo dato. Los procesos semi automáticos, en los que el software ofrece una solución que puede ser modificada y mejorada por el climatólogo son óptimos para redes regionales de tamaño pequeño y medio que no necesiten rehomogeneizaciones frecuentes.

4. Preparación de datos para trabajo con HOMER

Dado que el objeto final es el ajuste, en la medida que sea posible del dato diario, iniciaremos nuestro proceso con esa resolución. Se va a utilizar el formato denominado RClimdex (http://etccdi.pacificclimate.org/software.shtml), que consiste:

- Utilizar un fichero por estación
- Utilizar para nombrarlos la convención *ssccccccc.txt*, dónde *ss* significa status y debe ser *ra* para datos originales o raw y *qc* para datos que han pasado control de calidad; *cccccccc* es el código numérico de la estación (por ejemplo, el código OMM), que será rellenado con ceros a la izquierda para completar los ocho dígitos. Así, la estación OMM 83247 controlada de calidad, se codificaría como *qc00083247.txt*.
- Dentro de cada fichero, se utiliza un registro por día y ocho campos por registro. Los campos son año (yyyy); mes (*mm*); día (*dd*); precipitación (en milímetros, con precisión de un decimal); temperatura máxima diaria (grados centígrados, con precisión de un decimal); temperatura mínima diaria (grados centígrados con precisión de un decimal). En todos casos, el separador de decimales debe ser el punto (.), nunca la coma (,) y los campos se delimitarán por tabulaciones. Los valores perdidos se expresan como -99.9.
- Se requiere un fichero que contenga el listado de estaciones (ver Figura 2) disponibles, de nombre libre, cuyo formato será un registro por estación y 9 campos por registro. Los campos son nombre del fichero; grados de latitud; minutos de latitud; segundos de latitud; grados de longitud; minutos de longitud; segundos de longitud; elevación; nombre de la estación (sin espacios, sin caracteres especiales como acentos ni letras propias de idioma, como la ñ o la ç).

Toda la información que se acaba de describir se ubicará en nuestro directorio de trabajo (**md**, a partir de ahora) que contará con un fichero de estaciones y tantos ficheros de datos como series diarias dispongamos.

Tras preparar los datos, procederemos a su conversión en promedios mensuales. Para ello, utilizaremos una función programada sobre **R**. Cargaremos **R** y estableceremos como directorio de trabajo **md** (ya sea mediante la opción de menú del GUI *file/change dir/* o mediante la ejecución de *setwd('ruta/md')*, siendo ruta el *path* hasta llegar a **md**. Cargaremos – desde allí dónde se encuentre – el código *utiles.R*, facilitado en este taller. De nuevo, podemos utilizar los menús del GUI *(file/source R code)* o la función *source('ruta/utiles.R')*, siendo ruta el path hasta la carpeta dónde se encuentra el código a ser cargado.

Para la conversión del dato diario al mensual utilizaremos la función ***makemonthly***, que acepta los siguientes argumentos:

- interactive: si es 1, requerirá el resto de argumentos interactivamente; si es 0 (valor por defecto), habrá que introducirlos del modo que se describe a continuación.
- dailystats: nombre del fichero de estaciones, facilitado entre comillas. Por defecto es "stations.txt".
- percent: porcentaje de datos faltantes que se permiten en un mes para calcular el promedio/acumulación mensual. Por defecto, 5.
- minyear: Año más lejano al presente para el que queremos calcular valores mensuales. Por defecto, 1900.
- maxyear: Año más cercano al presente para el que queremos calcular valores mensuales. Por defecto, 2013.

La ejecución de ***makemonthly()*** calcularía promedios mensuales 1900 a 2013 para las estaciones contenidas en stations.txt, permitiendo un 5% de valores faltantes por mes. De querer, por ejemplo, leer del fichero de estaciones "misestaciones.txt", de 1931 a 2010, con un 10% de valores faltantes permitido, ejecutaríamos:

```
makemonthly(dailystats='misestaciones.txt',minyear=1931,maxyear=2010,percent=10).
```

Si por comodidad se prefiere introducir los valores interactivamente, se introducirá ***makemonthly(interactive=1)*** y se nos pedirán progresivamente los parámetros en pantalla. Tras este proceso contamos con los ficheros requeridos para la homogeneización en HOMER, que, básicamente, son:

- un fichero de estación por variable, con el mismo formato que el fichero de estación descrito para los datos diarios:
  - 000001stations.txt para precipitación acumulada mensual
  - 000002stations.txt para medias mensuales de la temperatura máxima
  - 000003stations.txt para medias mensuales de la temperatura mínima
  - 000004stations.txt para medias mensuales de la temperatura media (definida como semisuma de máxima y mínima, según criterio OMM)
  - 000005stations.txt para medias mensuales de la amplitud térmica diaria o DTR.

- Un fichero de datos para cada estación y cada una de las cinco variables mencionadas en el punto anterior, cuyas características son:
  - Nombre: *ssppmcccccccd.txt*, dónde *ss* es el código de estatus (*ra* o *qc*); *pp* corresponde al código de parámetro (*rr*,*tx*,*tn*,*tm*,*rn* para las cinco variables indicadas en el apartado anterior); *m* es literalmente "m", que indica que se trata de datos mensuales; *cccccccc* corresponde con el código de estación; *d* es literalmente "d", indicando que se trata de un fichero de datos; *.txt* es literalmente ".txt" para indicar la extensión de texto del fichero. Por requerimiento de HOMER, si nuestros datos diarios son raw (estatus ra), se creará una carpeta *./md/ra* y los datos se situarán en la misma. Si, por el contrario, son datos de estatus qc, la carpeta de emplazamiento de las series será *./md/qc*.
  - Formato: un registro por año, trece campos por registro. El primer campo corresponde al año en cuestión, los doce siguientes a los valores de enero a diciembre. Los datos se

expresan en las mismas unidades que el valor diario, aunque el código de valor perdido cambia a -999.9.

qc00087129.txt	27	46	0	64	18	0	199	Santiago
qc00087148.txt	26	49	0	60	27	0	92	Presidencia_Roque_SP
qc00087166.txt	27	27	0	58	46	0	62	Corrientes_Aero
qc00087244.txt	29	54	0	63	41	0	341	Villa_MRS
qc00087257.txt	29	53	0	61	57	0	88	Ceres
qc00087276.txt	28	26	0	58	55	0	1	Bella_Vista
qc00087328.txt	31	95	0	65	13	0	1	Villa_Dolores
qc00087349.txt	34	40	0	63	53	0	338	Pilar
qc00087374.txt	31	47	0	60	29	0	78	Parana_Aero
qc00087393.txt	30	16	0	57	39	0	54	Monte_Caseros_Aero
qc00087395.txt	31	30	0	58	2	0	1	Concordia
qc00087436.txt	33	16	0	66	21	0	713	San_Luis
qc00087453.txt	33	7	0	64	14	0	421	Rio_Cuarto_Aero
qc00087480.txt	32	55	0	60	47	0	25	Rosario
qc00087497.txt	33	0	0	58	37	0	21	Gualeguaychu_Aero
qc00087525.txt	35	45	0	60	88	0	1	Nueve_julio
qc00087548.txt	34	33	0	60	55	0	81	Junin_Aero
qc00087585.txt	34	40	0	58	39	0	22	Bueno_Aires_Obs
qc00087623.txt	36	34	0	64	16	0	191	Santa_Rosa_Aero
qc00087641.txt	36	50	0	59	53	0	147	Azul_Aero
qc00087648.txt	36	35	0	57	73	0	1	Dolores

Figura 2. Formato del fichero de estaciones de datos diarios, a partir del cual podremos preparar nuestros datos para HOMER y el resto del proceso.

5. Homogeneización de datos mensuales con HOMER

HOMER se carga requiriendo el script HOMER.R de cualquiera de las dos formas descritas para el código *utiles.R* y nos permitirá realizar el proceso de detección y ajuste de los datos mensuales tanto de temperatura como de precipitación. Otras variables, como presión atmosférica o velocidad del viento son potencialmente homogeneizables con este método.

HOMER funciona requiriendo parámetros en tiempo de ejecución, tanto para identificar la red y configurar los gráficos de salida, como para determinar si queremos realizar el proceso interactivamente desde la pantalla o ejecutarlo de corrido y revisar la información volcada en nuestro directorio de trabajo. También nos pedirá otros que serán de vital importancia para el éxito del proceso:

5.1 Determinación de la vecindad de comparación entre las distintas estaciones

Podemos escoger tres opciones: todas (recomendable para redes pequeñas con buena correlación); distancia geográfica (no recomendable); distancia estadística o correlación (recomendable para redes de 10 estaciones o más). Las dos últimas opciones permiten especificar el número mínimo de comparaciones a realizar, parámetro que se sobrepone al anterior criterio. Así, si tenemos una red de 30 estaciones e indicamos una correlación de 0.999 y un mínimo de 10 estaciones, se utilizarán las 10 mejores en las comparaciones que se realicen a partir de

este momento, sea para realizar el QC, para los distintos procesos de detección o para la corrección. Este parámetro, puede ser modificado en cualquier momento del proceso. De hecho, para la corrección – excepto en redes muy pequeñas – se recomienda reducir el número de series a utilizar en el momento de la corrección.

5.2 Selección del modelo de corrección

Aditivo (para variables físicas como temperatura o presión); multiplicativo (para variables acumulativas, como la precipitación o las horas de sol). El modelo multiplicativo ofrece dos versiones. Una de ellas se basa en ratios simples y la segunda, en el logaritmo de las ratios, válido y recomendado para variables siempre positivas que necesiten normalización (de nuevo, precipitación).

5.3 Escala temporal para la detección

Podemos detectar exclusivamente sobre los promedios anuales; sobre los promedios anuales y estacionales o sobre los promedios anuales, estacionales y mensuales. La segunda opción (promedios anuales y estacionales) es la recomendada.

5.4 Forma de la corrección en el modelo multiplicativo

Se puede seleccionar computar un factor anual y aplicarlo a todos los meses o un factor distinto para cada mes. La primera opción es la recomendada, puesto que el factor multiplicativo mensual es problemático en su aplicación.

Una vez introducidos los parámetros de entrada, HOMER nos ofrece dos familias de herramientas, las de Control de Calidad y las de Homogeneización:

- **Herramientas Control de Calidad:** aunque los datos deben haber sido controlados de calidad a nivel diario, HOMER cuenta con diversas herramientas de control de calidad. Se recomienda el uso de la opción “Fast QC” (ver Figura 3), que permite, de forma relativa (es decir, comparando con otras estaciones) la identificación rápida de outliers y una primera inspección visual y cualitativa de las inhomogeneidades potenciales de una estación. De detectar outliers muy evidentes, podremos pedirle a HOMER que los elimine de forma interactiva. Esos meses pasarán a ser valor perdido.

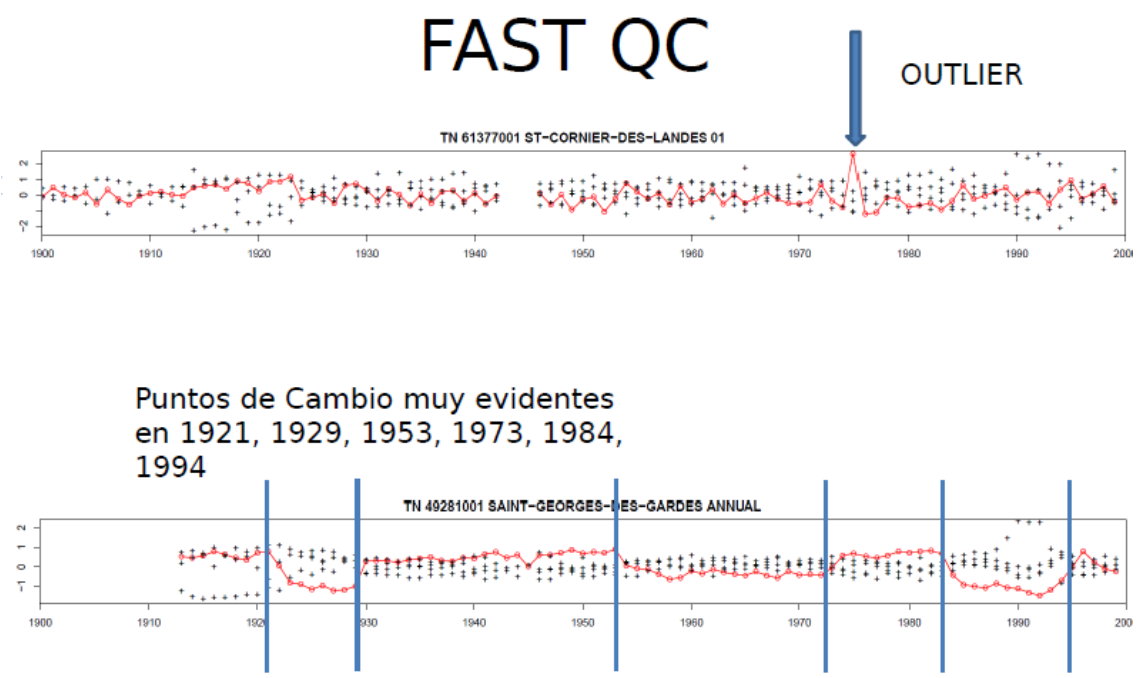

Figura 3. Ejemplo de la aplicación de la herramienta FASTQC. Los puntos rojos corresponden a la estación candidata; las cruces negras a sus distintas referencias. En el panel superior, se detecta un outlier (o valor fuera de rango); en el panel inferior, se visualizan cualitativamente los puntos de cambio que probablemente contenga la estación.

Una vez hayamos realizado este proceso, entraremos a utilizar las herramientas de homogenización que combinan detección y corrección, que se ejecutarán de forma sucesiva e iterativa:

- **Detección emparejada o Pairwise Detection** (ver Figura 4): proceso en el que se estiman los puntos de cambio existentes para la diferencia (ratio) entre una serie candidata y cada una de sus referencias. Ofrece un output gráfico para la determinación semi-objetiva de puntos de cambio. Para ello, deberemos analizar los gráficos que se habrán volcado en nuestro el directorio *./md/fig*, codificados de la siguiente forma: *detect_ssppccccccc_a.pdf*, dónde *detect* indica que se trata de un fichero de detección; *ssppccccccc* se define idénticamente que para el fichero de datos mensuales (estatus, parámetro, código), *a* indica que es la primera modalidad gráfica que se ofrece (no se describen las b y c, puesto que inicialmente no las usaremos). Por supuesto, .pdf indica el formato del fichero. Esta descripción corresponde a las detecciones sobre ficheros anuales. De realizarse, como se recomienda, detección sobre los ficheros estacionales, antes de la extensión, aparecerá el trío de letras

que identificará el trimestre estándar sobre el que se detectó (DJF, MAM, JJA, SON). Para utilizar estos ficheros, debemos establecer un criterio para identificar BPs. Si hemos determinado que cada estación va a ser comparada con otras 10, por ejemplo, este criterio puede ser que ya sea en el fichero de detecciones anuales, ya sea en alguno de los estacionales, este BP aparezca en 5 comparaciones con 1 año de margen. Anotaremos estos BPs detectados semi-objetivamente y los tendremos en cuenta en el siguiente paso. Hacer notar también que, aunque HOMER no detecta directamente tendencias artificiales (como las urbanas) estas tienen un patrón característico: diversos puntos de cambio consecutivos y no muy lejanos en el tiempo del mismo signo. De identificar muy claramente este patrón, deberemos anotarlo igualmente, puesto que existe una opción específica para su corrección.

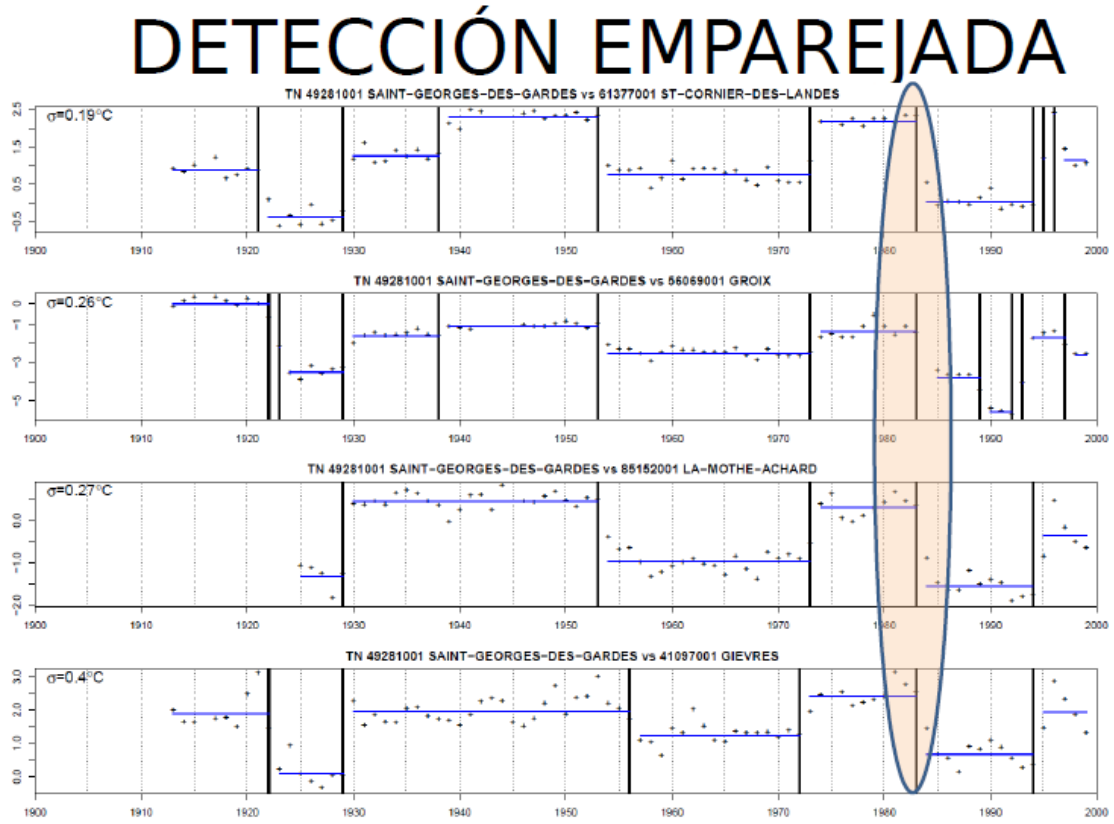

Figura 4. Ejemplo de detección emparejada. En el caso enmarcado con una elipse, la estación candidata presenta un BP en su comparación con las cuatro referencias disponibles. Debemos aceptar ese punto de cambio. Por el contrario, otros no aparecen más que en una o dos comparaciones. No serán, inicialmente, tenidos en cuenta.

- **Detección conjunta o Joint Detection** (ver Figura 5): mediante la detección simultánea entre todos los pares de estaciones, se determina para cada una de ellas el número y posición más probable de puntos de cambio. En modo interactivo (recomendado para este paso) los

puntos de cambio se muestran en pantalla de forma gráfica. El gráfico, mediante un simple *click*, nos permite añadir o eliminar puntos de cambio de acuerdo al propio gráfico (que muestra tanto los resultados de la detección emparejada como de la detección conjunta), nuestra detección semi-objetiva y los metadatos de los que dispongamos. Los puntos de cambio que hayamos aceptado o añadido se registrarán en *./md* un fichero cuyo nombre es *detectednnnnnn.txt*, siendo *nnnnnn* el número de red. Este fichero contiene un registro por punto de cambio y 6 registros por campo: el código de la estación; la palabra clave BREAK; el año del punto de cambio, el mes del punto de cambio (inicialmente fijado a 12), un carácter que indica si el punto de cambio está fijado por metadatos (y) y no debe ser alterado en futuras iteraciones o, si por el contrario (n) no está fijado por metadatos y puede ser alterado en futuras iteraciones; el nombre de la estación. Naturalmente, si hemos decidido no correr el proceso en modo interactivo, podemos trabajar directamente sobre este fichero, ya sea editándolo con la opción que para ello HOMER ofrece, ya sea mediante un editor de texto. Debe siempre respetarse el formato del fichero. También, si hemos identificado alguna inhomogeneidad gradual, tipo tendencia urbana, podemos consignarlo en este fichero, con una entrada en el año inicial de la tendencia, que contendrá la palabra clave BEGTR y otra en el punto final de la tendencia que contendrá la palabra clave ENDTR.

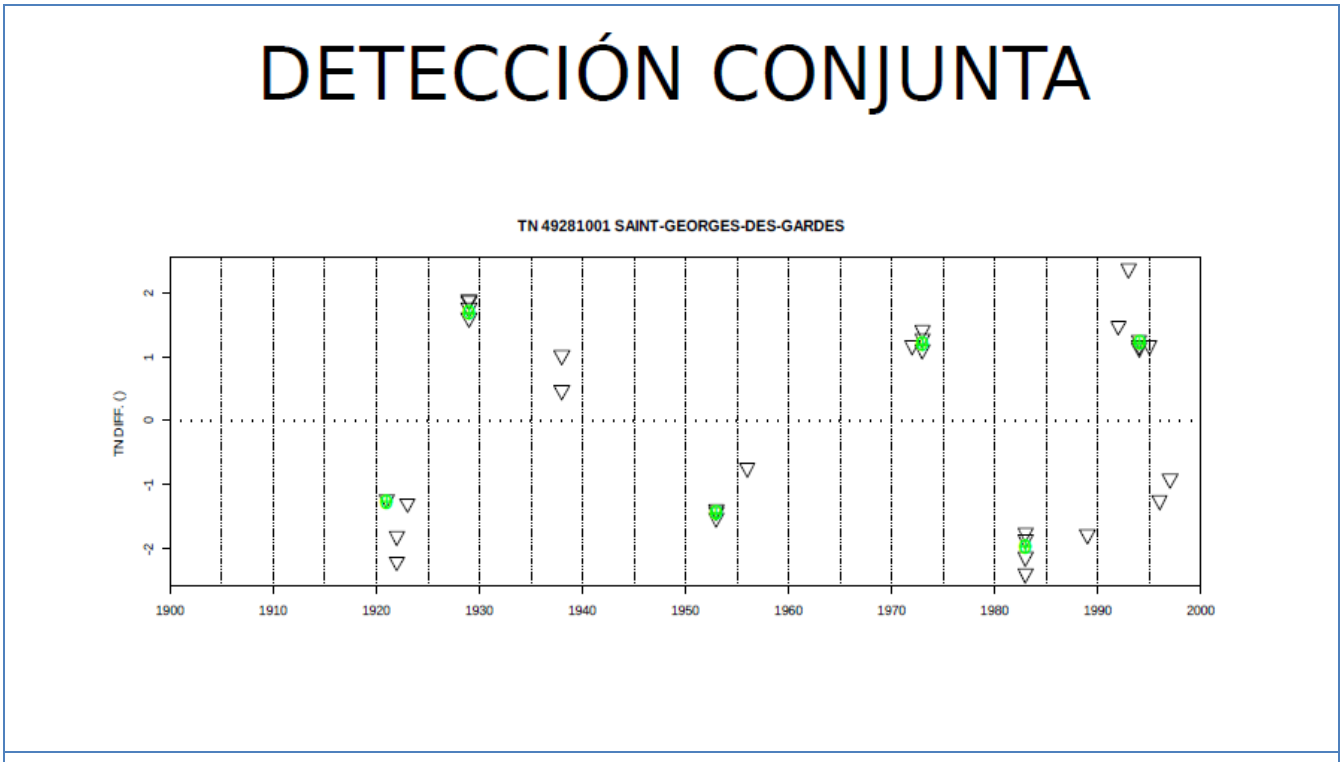

Figura 5. Detección Conjunta. En la figura, aparecen como triángulos los resultados de la detección emparejada y en forma de círculos verdes los de la detección conjunta. En formato interactivo, un click sobre la pantalla añadirá o quitará un punto de cambio en ese lugar.

- Corrección preliminar: una vez se ha realizado la detección emparejada y la conjunta y refinado el fichero de detección, se puede realizar la corrección preliminar de los datos. Tras la misma, HOMER realizará automáticamente una detección emparejada sobre los datos corregidos, que – de momento – obviaremos. Téngase en cuenta que este paso es imprescindible, ya que el próximo utiliza los datos corregidos preliminarmente. Para todas las correcciones (preliminar y no preliminar) se recomienda reducir la red estaciones. Para correr este paso, se recomienda desactivar la opción interactiva.
- Ajuste del mes de cambio: HOMER, sobre los datos corregidos preliminarmente, puede inspeccionar los datos mensuales desestacionalizados entre dos puntos de cambio para tratar de refinar su detección. Esta podrá variar hasta tres años del punto detectado originalmente. Cuando HOMER no encuentra una solución clara o no tiene suficientes datos entre puntos de cambio para hacer esa estimación, el punto de cambio se mantendrá en el año en que se detectó originalmente y en el mes 12.
- Corrección: tras el ajuste del mes de cambio, se realiza una nueva corrección. Nótese que las correcciones se realizan siempre sobre los datos originales, nunca sobre correcciones previas. De nuevo, la corrección lanzará una detección emparejada sobre los datos corregidos, que aprovecharemos en el siguiente paso.
- Iteración: inspeccionaremos los ficheros resultantes de la detección emparejada sobre los datos corregidos siguiendo el mismo criterio que establecimos con anterioridad. Ello nos permitirá comprobar si quedan puntos de cambio que no detectamos previamente o sí, por el contrario hemos "aplanado" las diferencias (ratios) entre candidatas y referencias, y podemos considerar la red como homogénea. De ser así, el proceso ha terminado. De encontrarnos en el primer caso – lo más común – iteraremos nuestro proceso desde la detección emparejada, que en este caso se realizará sobre los datos corregidos. De necesitar más de tres iteraciones, o de notar que en cada iteración la homogeneidad de la red no mejora, deberemos asumir que, en algún momento del proceso, nos equivocamos e introdujimos falsos puntos de cambio. En ese caso, deberemos empezar de nuevo.

La Figura 6 y la Figura 7 muestran respectivamente una red y una estación bien homogeneizadas. La Figura 8, el proceso de trabajo con HOMER y la **¡Error! No se encuentra el origen de la referencia.** los directorios que el software genera. La Figura 10 muestra el proceso completo de homogeneización.

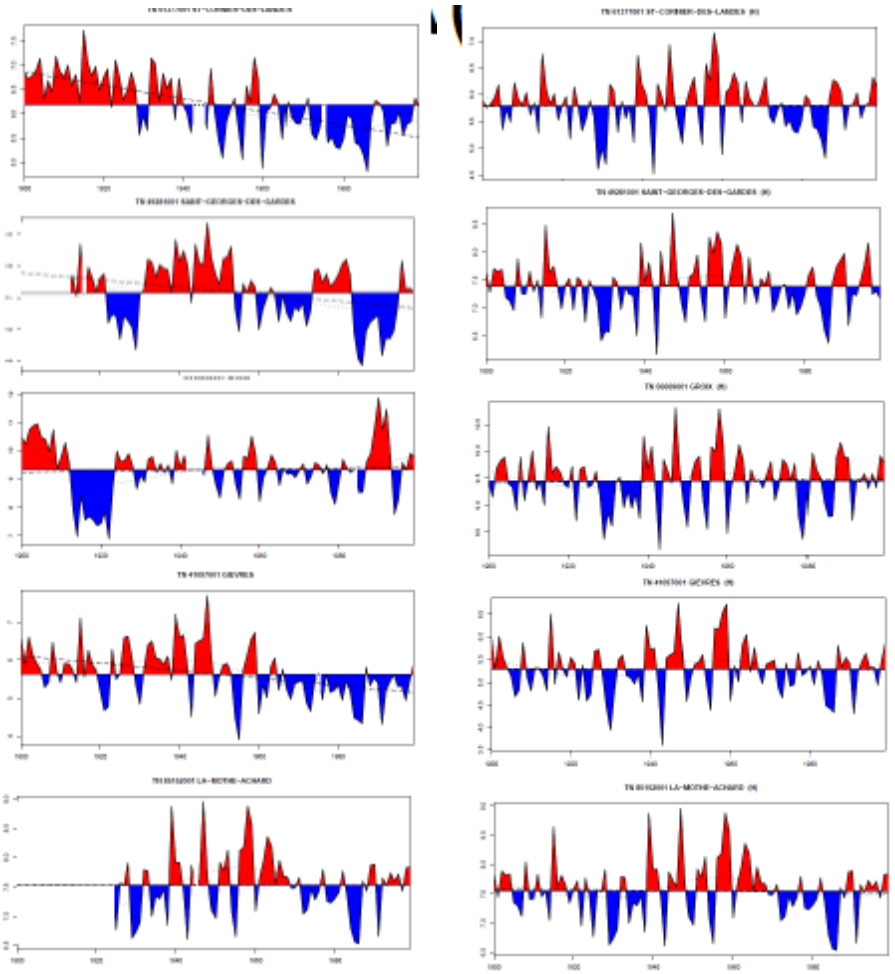

Figura 6. Red original (izquierda) y red homogeneizada (derecha). La coherencia entre estaciones es evidente en el segundo caso. El relleno de datos no afecta al proceso de homogeneización, ya que se realiza con posterioridad al mismo. La Figura 8 resume el proceso de trabajo con HOMER.

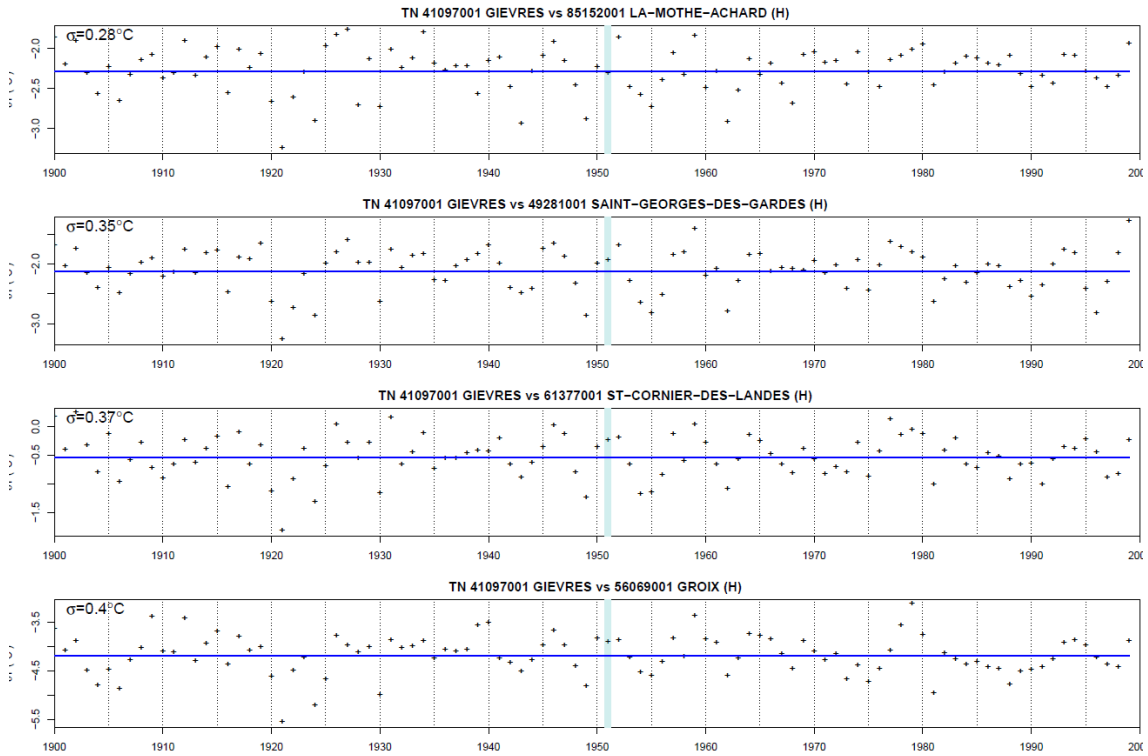

Figura 7. Ejemplo de estación que, tras la corrección de un único punto de cambio, queda homogénea respecto al resto de la red.

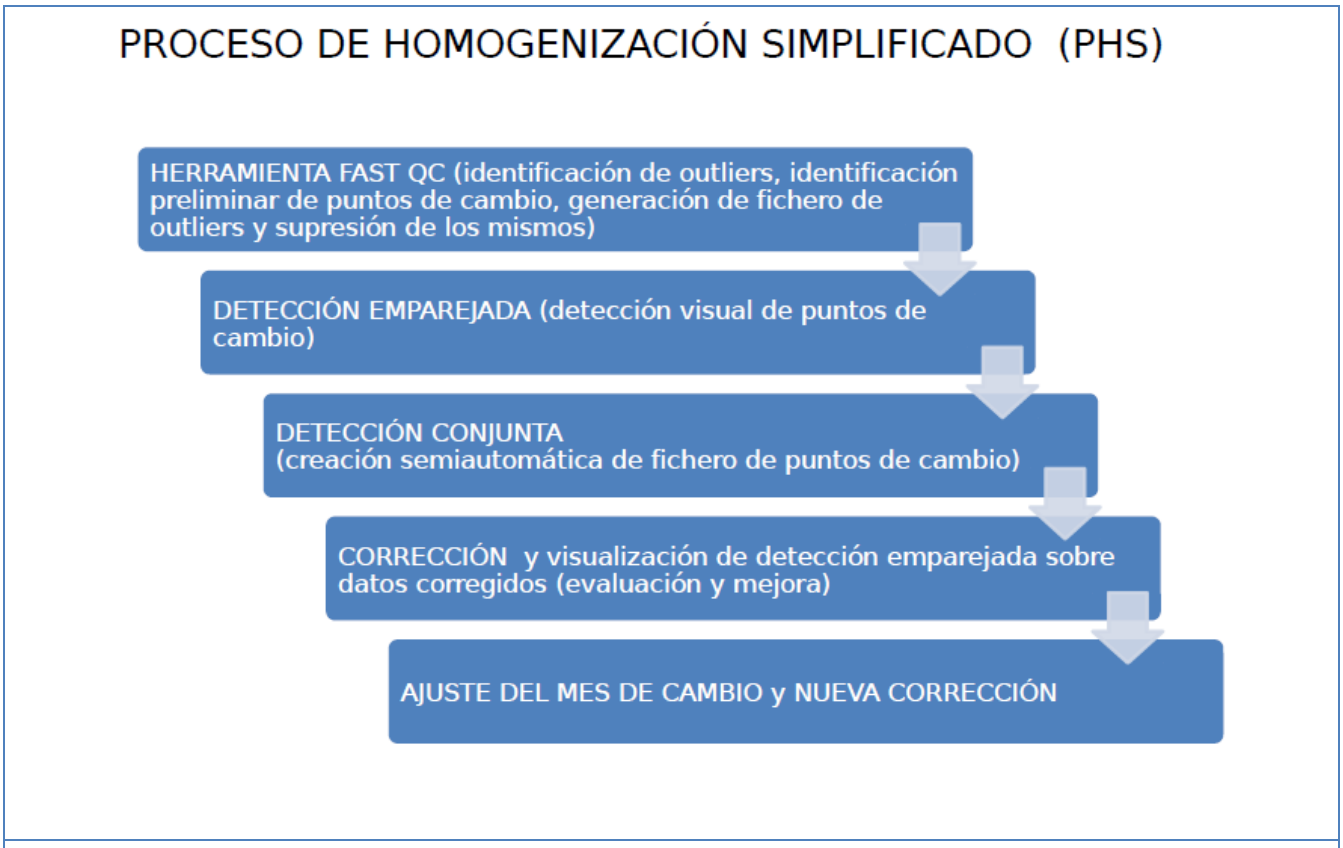

Figura 8. Proceso de Homogeneización simplificado. Una vez llegados al último paso, debemos iterar desde la detección conjunta (no más de 3-4 ocasiones) hasta conseguir que la red aparezca homogénea.

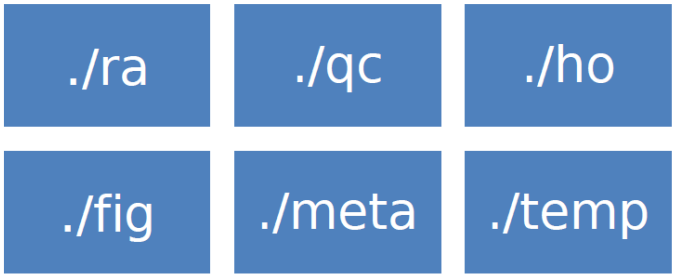

Figura 9. Estructura de directorios de HOMER. En ./ra se ubican los datos raw; en ./qc los datos que han pasado control de calidad (automáticamente desde makemonthly); en ./ho los datos mensuales homogeneizados por HOMER; en ./fig los gráficos que genera y en ./meta los gráficos de control de calidad más estadísticas y otros metadatos del proceso de homogeneización.

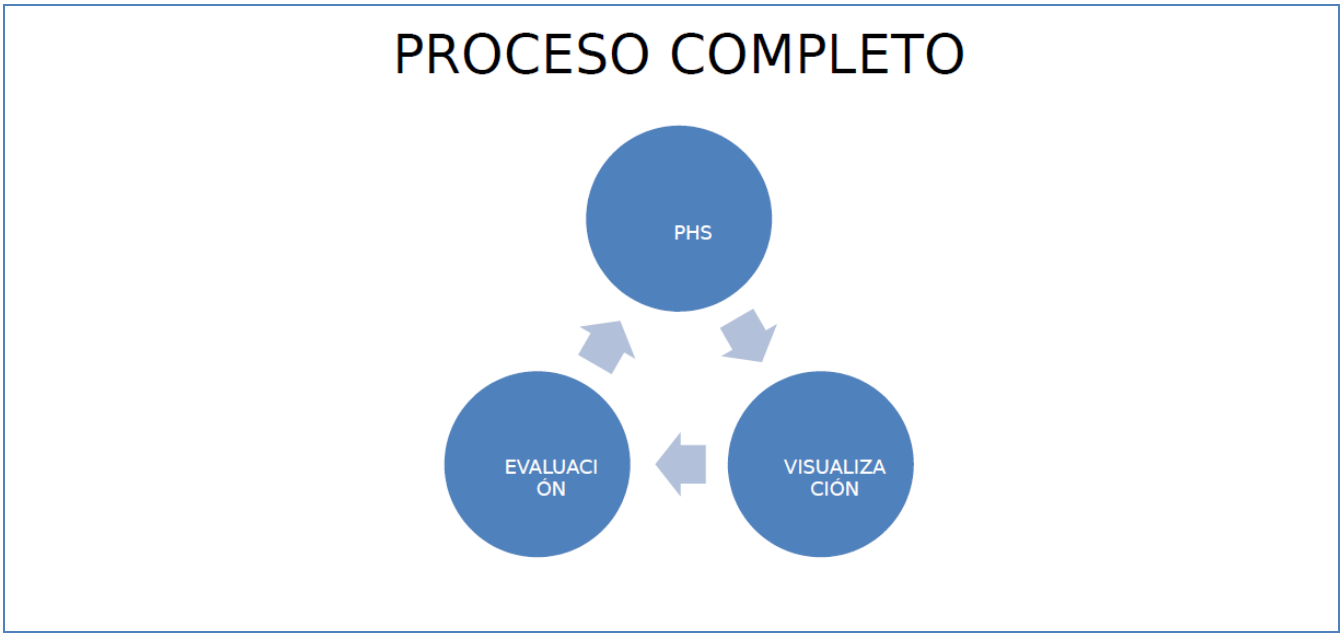

Figura 10. Proceso completo de homogeneización. El proceso simple (PHS) debe ser validado visualizando y evaluando estadísticamente los resultados.

6. Homogeneización de datos diarios

La homogeneización del dato diario es aún más compleja que la homogeneización del dato mensual, debido tanto a las propias características estadísticas del dato a esa resolución menor, como de la necesidad de abordar la homogeneización no solo de la media, sino también la de la distribución completa de los datos, dada la importancia del trabajo con extremos. Ya de inicio debemos decir que esta opción no siempre va a ser posible, puesto que va a requerir que nuestra base de datos posea una gran densidad espacial de estaciones completas, altamente correlacionadas y con espacio sin solapamiento entre los puntos de cambio. Por ello, existen técnicas "de seguridad". En el caso de temperatura, se basan en la interpolación del factor mensual al dato diario, que van a acercar los datos diarios hacia un estado más homogéneo, corrigiendo la media y – a través de esa corrección – mejorando levemente la detección de extremos (al menos de extremos moderados como los contenidos en los paquetes de índices del Expert Team on Climate Change Detection and Indices o ETCCDI). En el caso de precipitación, simplemente aplicaremos el factor anual a los datos diarios. Obviamente, no es una solución óptima y la investigación en otras aproximaciones es necesaria. Entre las mismas podemos destacar las que se basan en el ajuste de frecuencias y en el filtrado de bajas cantidades (por ejemplo, convertir a 0 todos los valores por debajo de 1 mm o de 3 mm) puesto que es en esas frecuencias dónde se concentran las inhomogeneidades. Si hemos realizado un buen control de calidad que nos haya librado de extremos falsos (como los derivados de la precipitación acumulada en varios días y falsamente reportada en un día final), podremos realizar una detección de extremos (de nuevo, al menos los ofrecidos por el ETCCDI más sólida).

Estos métodos pueden aplicarse siempre que hayamos podido homogeneizar el dato mensual.

Si se ha realizado la homogeneización de temperaturas máximas y mínimas, puede interpolarse ésta directamente al dato diario siguiendo el proceso que se describe a continuación:

- Debe cargarse el código utiles.R tal como se describió anteriormente
- Se ejecuta la función hdwlv()
- Aparecerán en ./md series en formato *hocccccc.txt*, de conteniendo los datos ajustados de temperatura máxima y mínima en formato RClimdex. Los datos de precipitación, por el momento, no se alteran.

Otras técnicas más complejas se basan en modelizar la relación entre una estación candidata y otra de referencia y trasladarla al periodo anterior al punto de cambio mediante distintas aproximaciones:

- Quantile Matching: empareja los quantiles de la distribución empírica de las diferencias. Corresponden a esta familia el método descrito por Trewin (2013), y sus variantes (disponibles en R bajo petición); la familia de ajustes para dato diario del paquete RhTestV4, de Xiaolan Wang (http://etccdi.pacificclimate.org/software.shtml).
- Ajuste de distribuciones de probabilidad: empareja los cuantiles de una distribución de probabilidad, seleccionada entre múltiples opciones. Pertenece a esta familia el método HOM, contenido en el paquete HOMSPLIDHOM.R (Mestre et al., 2011).
- Ajuste de regresiones no lineales: ajusta una regresión no lineal basada en cubic spline. Pertenece a esta familia SPLIDHOM, incluido en el paquete HOMSPLIDHOM.R
- Correcciones en función de otras variables/tipos de tiempo: en desarrollo, requieren mayor cantidad de información.

Respecto a estos métodos debemos saber que:

- Tan solo RhTestV4 ofrece la posibilidad (no recomendada) de ajustar el dato diario de forma absoluta
- El método de Trewin sugiere el uso de tres referencias e indica que correlaciones superiores a 0.7 son aceptables, dado que el ajuste proviene de la mediana de tres ajustes. Esta afirmación – probada para Australia – no ha sido comprobada para otras latitudes/países. En cualquier caso, la dificultad de obtener tres estaciones bien correlacionadas es notable.
- Los métodos contenidos en HOMSPLIDHOM requieren correlaciones de al menos 0.8 (óptimamente 0.9) para mejorar el rendimiento de la interpolación del dato mensual para la variable temperatura.
- Si existe, entre dos estaciones, una excelente correlación pero tiene BPs cercanos, no podrán ser utilizadas para la corrección.

Dada la dificultad de aplicar los métodos complejos anteriormente descritos, debemos considerar la posibilidad de mantener distintas bases de datos: una, con mayor número de estaciones y mayor longitud temporal, a partir de la traslación del factor mensual al diario; una segunda – que contendrá menor número de estaciones y/o un período temporal más corto – aplicando métodos más complejos. Finalmente, es también válida la aproximación

de descartar los tramos inhomogéneos detectados, siempre y cuando nuestro objetivo y densidad de red nos lo permita.

7. Validación de la corrección

Todo proceso de homogeneización, como en general cualquier procedimiento científico, necesita una validación de resultados. Recordemos que, al homogeneizar, pretendemos librar nuestras estaciones de sesgos artificiales, hacer todas sus observaciones comparables al último dato registrado y - mediante ello - aumentar la coherencia regional de nuestra red. No obstante, debemos ser cuidadosos y monitorizar diversos factores ya sea con herramientas gráficas o numéricas. Deberíamos realizar inicialmente un control de calidad a las series para:

- Evaluar la introducción de *outliers* tras la aplicación de los factores de corrección.
- Detectar y corregir la presencia de *overshooting*, que consiste en llevar por la aplicación de distintos factores de corrección la temperatura mínima de un día por encima de su máxima correspondiente. Este fenómeno ocurrirá con casi inevitablemente en unas pocas observaciones cuando corregimos una red de estaciones, especialmente si no hemos tratado de aplicar factores similares en temperaturas máximas y temperaturas mínimas. Podemos realizar una corrección simple basada en la conservación de la ratio de cambio del DTR inhomogéneo mensual respecto al DTR homogéneo mensual, definiendo los siguientes elementos y aplicando los siguientes cálculos a los días que incurran en este problema:
    - DTRa = DTR para los datos mensuales originales mes y año del dato que incurre en overshooting
    - DTRb = DTR para los datos mensuales homogeneizados para el mes y año del dato que incurre en overshooting
    - DTRc = DTRb/DTRa
    - DTRdia = DTR original del día que incurre en overshooting
    - DTRerr = DTR ajustado y negativo (erróneo) del día que incurre en overshooting
    - DTRtarget= DTRc*DTRdia-DTRerr
    - Sumar DTRtarget/2 a la TX y restar DTRtarget/2 a la TN
- Evaluar la integridad del ciclo estacional: dado que habremos aplicado factores a nivel mensual y/o diario, deberemos asegurarnos que no hemos introducido un artificio que destruya en su aplicación el ciclo anual lógico de la variable.
- Evaluar la relación entre las series mensuales homogeneizadas y las series diarias homogeneizadas. Si se ha recurrido a métodos otros que la traslación del factor mensual, van a diferir.

Evaluaremos también la coherencia regional de las series. Una buena homogeneización debe aumentar la correlación entre las mismas y ofrecer mapas de tendencias más coherentes. Si cartografiamos las series en paralelo o mapeamos valores de tendencias temporales, el impacto positivo de la homogeneización debe ser evidente.

8. Propuesta de trabajo

El Grupo de Trabajo 1 del CRC-SAS se enfrenta a la tarea producir una base de datos homogénea de diversas variables. El proceso sugerido para la homogeneización de los datos de temperatura máxima y mínima y precipitación, para una red regional de unas 500 estaciones es:

1. Aplicar el control de calidad al dato diario. El Propio CRC-SAS está desarrollando su propia herramienta de QC, óptima para este trabajo. El proceso se estima finalizará en 6 meses. (mes 6)
2. Realizar la homogeneización del dato mensual con HOMER, siguiendo el proceso descrito con anterioridad. Aunque sería posible realizar este proceso en una red única, se recomienda realizar homogeneizaciones en redes sub-regionales. Estas redes pueden coincidir con las fronteras nacionales, o mejor aún, con fronteras climáticas. En cualquier caso, cada red debe contar con la posibilidad de añadir estaciones adyacentes para mejorar al proceso de detección/corrección. La homogeneización de las distintas redes puede realizarse tanto de manera centralizada (i.e. desde el CRC-SAS) o de manera distribuida (i.e., desde cada uno de los SMHNs). En cualquier caso y, especialmente, en el segundo, deben fijarse a priori una serie de criterios claros para considerar/rechazar los puntos de cambio que las distintas herramientas de HOMER detectan. También, en el caso de homogeneización distribuida será necesaria una punto focal que centralice la recolección de los datos y metadatos del proceso (es decir, las carpetas generadas por HOMER) para su revisión. (mes 7-12)
3. Realizar la interpolación del factor mensual al dato diario en la temperatura; trasladar el factor anual en el caso de la precipitación. Este punto no ofrece ninguna dificultad: simplemente, aplicar el código y obtener las series ajustadas. (mes 13)
4. Validar el proceso de homogeneización realizado hasta este momento. Siguiendo los pasos descritos en el apartado anterior, deberá ser realizada preferentemente por cada país. En aquellas estaciones que – por solapamiento – hayan sido homogeneizadas en más de una red, será necesario comparar los resultados y decidir cuál de ellas nos ofrece la mejor solución. El proceso puede realizarse tanto de forma centralizada como de forma distribuida, siempre existiendo una figura responsable de coordinar y finalizar dicha tarea. (mes 14)
5. Estudiar la viabilidad de aplicar, a un subconjunto de estaciones, técnicas más avanzadas de homogeneización. De nuevo, puede realizarse de forma centralizada o distribuida, con redes solapadas y mediante criterios claros preestablecidos. Se recomienda de cómputos estacionales y una correlación mínima de 0.8. (meses 15-18)
6. Valorar los resultados de esta segunda homogeneización. Idéntica recomendación que la realizada en el punto 4. (mes 19)

9. Otros retos y resultados e investigaciones complementarias

Como hemos visto a lo largo de este documento, el proceso de homogeneización permite realizar de forma sólida algunas tareas, como la detección de inhomogeneidades pero se necesita investigación en otros campos,

como la corrección sólida del dato diario, tanto en temperatura, como muy especialmente en otras variables. Existen iniciativas internacionales como la International Surface Temperature Initiative (IST, www.surfacetemperatures.org) que, persigue la compilación de un banco de datos global en el momento presente de temperatura media mensual. Esta iniciativa está respaldada por la OMM e integra un grupo de experimentación en homogeneización que genera bancos de datos sintéticos homogéneos e inhomogeneizados a partir del mejor conocimiento que poseemos sobre la frecuencia y características de las inhomogeneidades, incompleta especialmente en las regiones tropicales y en el Hemisferio Sur (excepto Australia). Otra iniciativa internacional, en estado más embrionario, es el Parallel Data Inititative (https://ourproject.org/moin/projects/parallel, http://tinyurl.com/paralleldata) que pretende recoger y analizar medidas paralelas para estudiar el impacto que la sustitución de un sistema por otro (garita A por garita B; automática por convencional, etc.), así como la incertidumbre. Dicha iniciativa, reconoce la prioridad de aquellos contribuidores de datos que deseen realizar sus estudios iniciales.

1. Recolección de metadatos de las estaciones a homogeneizar: es muy importante disponer de metadatos completos para guiar el proceso de homogeneización. No sólo para mejorar el proceso de detección, sino también para validar los ajustes. Por ejemplo, si conocemos que una estación cambió en el año 1950 del centro de la ciudad al aeropuerto y podemos afinar una hipotética detección en 1951 (dentro del margen aceptable de error); también, si conocemos que la zona del aeropuerto es más fresca que el centro de la ciudad, si los factores de ajuste fueran negativos (es decir, implicaran hacer más fría la sección del aeropuerto) deberíamos pensar que hemos cometido algún tipo de error (i.e. mala selección de referencias, algún problema como códigos de valor perdidos o mal especificados, etc.)
2. Análisis de los metadatos de la homogeneización: se puede contribuir a responder a preguntas como:
   - ¿Cuál es la frecuencia de BPs por cada cien años?
   - ¿Cuál es el factor de ajuste medio? ¿Es cero o por lo contrario tiene algún sesgo positivo o negativo?
   - ¿Cuál es el ciclo estacional de los ajustes?
   - Si disponemos de buenos metadatos y conocemos las causas de algunas inhomogeneidades, como por ejemplo el traslado ciudad/aeropuerto, podemos conocer su firma o patrón.
3. Estudio de medidas emparejadas de datos diarios. Permiten resolver problemas específicos como el impacto de la automatización de las series mediante:
   - Conocimiento de la magnitud, ciclo estacional e impacto en la distribución de la sustitución.
   - Posibilidad de aplicación de métodos de homogeneización de datos diarios standard (como los descritos anteriormente) adaptados a datos emparejados (Ver Figura 11).
   - Desarrollo de métodos específicos generalizables.
   - Estudio de la incertidumbre introducida por los distintos métodos de corrección.
4. Aplicaciones a otras variables: conocemos relativamente poco respecto a cómo homogeneizar otras variables que la temperatura. Dado que la base de datos compilada y

controlada de calidad por parte del Grupo de Trabajo 1 del CRC-SAS contiene múltiples. La aplicación y desarrollo de técnicas específicas puede desarrollarse mediante:

- Aplicación de técnicas sobre datos reales y evaluación subjetiva experta de los resultados.
- Generación de bancos de datos simulados mediante diversas técnicas y evaluación objetiva:
  - Composición de estaciones: consiste en, dentro de una base de datos generar un punto de cambio conocido realizando una composición artificial entre dos estaciones bien correlacionadas.
  - Generación de bancos de datos artificiales, mediante técnicas complejas que se resumen en generar datos "homogéneos" e introducirles distintas inhomogeneidades.

5. Valorar la posibilidad de ceder fuerza de trabajo y datos a las iniciativas internacionales anteriormente mencionadas (ISTI) para colaborar en un mejor conocimiento global de los problemas que hemos descrito.

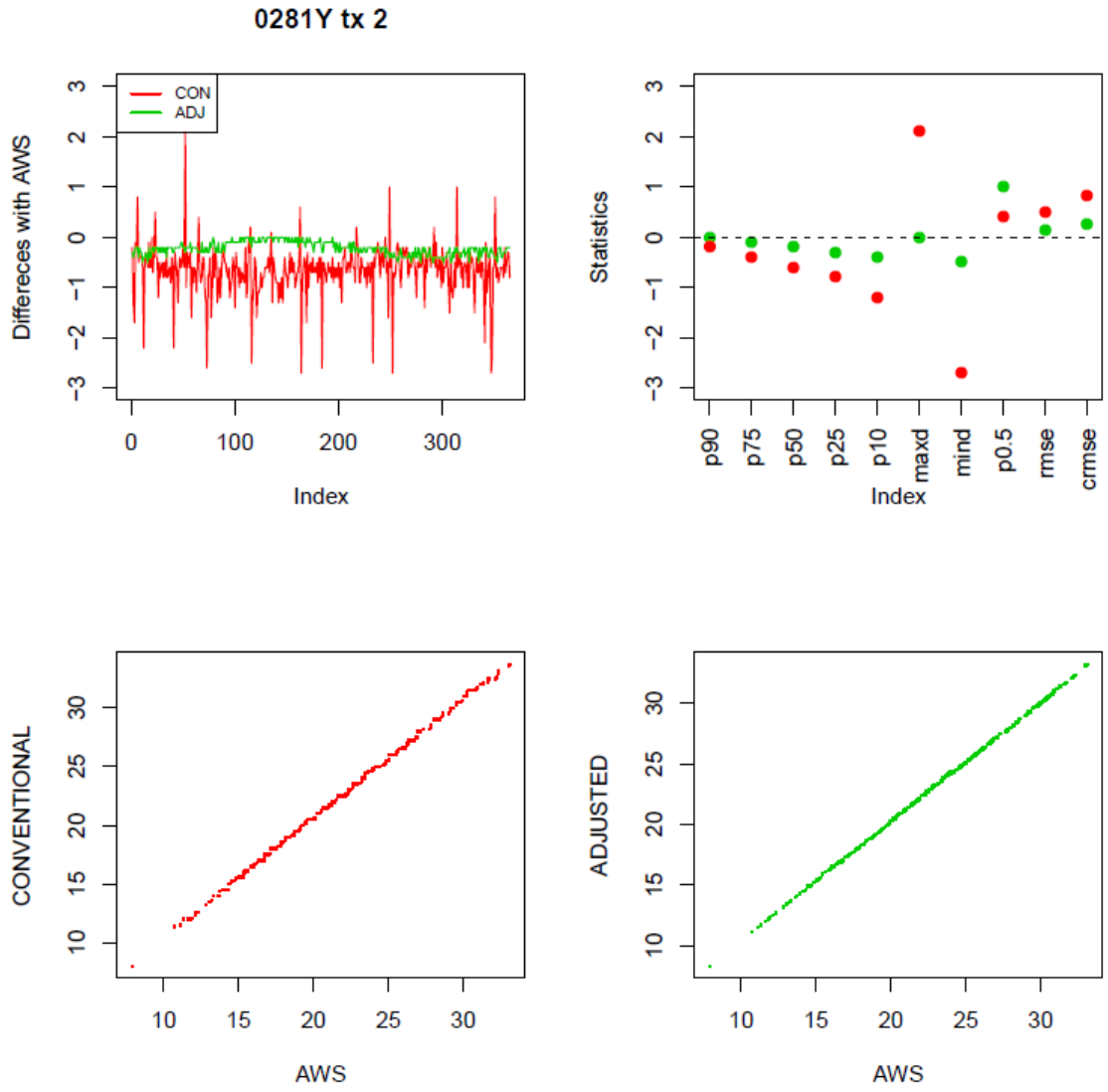

Figura 11. Corrección de diferencia entre automática y convencional en una estación española. Se han realizado los ajustes mediante una versión modificada del método QM de Trewin.

Referencias

Caussinus, H. y Mestre, O., 2004. Detection and correction of artificial shifts in climate series. Journal of the Royal Statistical Society, 53: 405-425.

Mestre, O. et al., 2013. HOMER : HOMogenisation softwarE in R- methods and applications. IDÖJÁRÁS - Quarterly Journal of the Hungarian Meteorological Service, 117(1): 47-67.

Mestre, O., Gruber, C., Prieur, C., Caussinus, H. y Jourdain, S., 2011. SPLIDHOM: A Method for Homogenization of Daily Temperature Observations. Journal of Applied Meteorology and Climatology, 50(11): 2343-2358.

Trewin, B., 2013. A daily homogenized temperature data set for Australia. International Journal of Climatology, 33(6): 1510-1529.