RELLENO DE LAGUNAS Y HOMOGENEIZACIÓN DE SERIES DE PRECIPITACIÓN EN REDES DENSAS A ESCALA DIARIA

Rafael CANO TRUEBA* y José Manuel GUTIERREZ LLORENTE**
** Instituto Nacional de Meteorología*
*** Dpto. de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria*

RESUMEN

El fin último de este trabajo es obtener series de referencia completas y homogéneas a partir de series largas en redes de observación cuya elevada densidad permite formular hipótesis aproximativas de gran utilidad en el tratamiento del dato diario. Para ello, este artículo se divide en dos partes claramente diferentes, aunque estrechamente relacionadas como son: el relleno de lagunas y la homogeneización en series temporales espacialmente distribuidas. En ambos casos se explotan las posibilidades derivadas de la utilización de series diarias de observaciones en redes densas como la red pluviométrica del Instituto Nacional de Meteorología. En primer lugar se aplica un método de interpolación gaussiana sobre variables como la precipitación, para rellenar lagunas. En segundo lugar se aplica un método de homogeneización basado en la eliminación de EOFs.

Palabras clave: Lagunas, homogeneización, EOF, interpolación gaussiana, dato diario, precipitación.

ABSTRACT

This paper has two clearly different parts, but closely related: filling missing data and homogeneization of temporal series. It exploits the very dense daily series of Spanish weather service (INM). First we apply an special version of gaussian interpolation to fill missing data in daily precipitation series. And finally a new method of EOFs elimination is described for homogeneization purposes.

***Key words**: Missing data, homogeneization, EOF, gaussian interpolation, daily series, precipitation.*

1. INTRODUCCIÓN

Un problema que se afronta en la práctica al trabajar con series temporales de observaciones es que es muy común encontrar defectos como la existencia de lagunas (falta de datos) y/o inhomogeneidades (derivadas de cambios en los parámetros de observación). Por una parte, las lagunas dificultan la aplicación de las distintas técnicas estadísticas (en la mayoría de los casos obligan a utilizar algoritmos iterativos costosos, del tipo de Esperanza-Maximización, EM). Por otra parte, las inhomogeneidades se reflejan en series no estacionarias que, en realidad, no representan a un único "modelo". Por todo ello, es muy útil disponer de series de referencia completas y homogéneas que permitan llevar a cabo estudios climatológicos de forma simple y consistente.

Los dos problemas anteriores, aunque distintos en esencia, comparten muchas características comunes. Por ejemplo, un método de homogeneización marcará las pautas para el desarrollo de métodos de relleno de lagunas que preserven la homogeneidad elegida. En la literatura, el interés por estos problemas se ha centrado casi exclusivamente en la escala mensual y han sido numerosos los métodos de homogeneización (ALEXANDERSSON, 1986) y relleno de lagunas propuestos

García Codron, J.C.; Diego Liaño, C.; Fdez. de Arróyabe Hernáez, P.; Garmendia Pedraja, C. y Rasilla Álvarez, D. (Eds.) (2004). **El Clima entre el Mar y la Montaña.** Asociación Española de Climatología y Universidad de Cantabria, Serie A, nº 4, Santander.

(una buena revisión de métodos se cita en: ALLEN *et al.,* 1998; software con algunos métodos se puede conseguir en: STEPANEK, 2001; etc.). La gran ventaja es que las series mensuales de precipitación pueden considerarse aproximadamente normales con correlaciones significativas a distancias de cientos de kilómetros.

Este artículo trata estos problemas a escala diaria, lo cual dificulta el análisis ya que no se cumple la hipótesis de normalidad de los datos, pues estos requieren una distribución apropiada (gamma, doble exponencial, etc.) y además, las correlaciones entre estaciones cercanas son más débiles, con un alcance significativo del orden de decenas de kilómetros. En este artículo mostramos que si se dispone de una red de observatorios suficientemente densa (como la red pluviométrica del Instituto Nacional de Meteorología, con más de 10.000 observatorios en España), se tienen correlaciones muy significativas entre estaciones cercanas incluso cuando se considera el dato diario; además, se obtienen muy buenos resultados considerando distribuciones normales truncadas. A partir de este hecho, se propone un nuevo método de relleno de lagunas basado en un único modelo normal multivariado para todas las estaciones (se opera considerando como unidad las cuencas parciales, o subcuencas, definidas según el indicativo de la estación, ver SMN, 1968), que permite obtener la esperanza de la precipitación para un día dado en cualquier estación a partir de cualquier subconjunto de las observaciones disponibles.

En la segunda parte del trabajo, se analiza el problema de la homogeneización de datos diarios utilizando el método de las Componentes Principales (más popularmente conocido como EOF en este ámbito) para reducir el ruido (o inhomogeneidad). La idea de este método entronca con el método de relleno de lagunas, pues las EOF determinan direcciones óptimas donde el conjunto de estaciones expresan máxima varianza (y por tanto tienen una dinámica común).

Se verá que el método de relleno de lagunas descrito conlleva parejo un incremento de la homogeneidad de las series, ya que las ausencias se rellenan con combinaciones lineales óptimas de estaciones cercanas (favoreciendo así alguna de las EOFs).

El objetivo final de este artículo es la obtención de series completas y homogéneas (en la medida de lo posible), haciendo especial énfasis en el tratamiento masivo de series diarias; de acuerdo con MEKIS y HOGG (1999): "*El uso del dato diario permite aplicar métodos de relleno y de corrección que no son posibles con series de datos mensuales, además la disponibilidad de series largas, continuas y homogéneas ofrece grandes ventajas en la investigación climatológica"*. Por otra parte, según GROISMAN (1994): "*Los métodos generalizados de corrección permiten trabajar con un gran número de estaciones sin necesidad de recurrir al metadato"*. De acuerdo con estas premisas, algunos de los métodos descritos han sido elegidos anteponiendo la eficiencia a la idoneidad.

2. DESCRIPCIÓN DE LOS DATOS

Se ha utilizado la base datos de la red pluviométrica del INM, con más de 10.000 estaciones. Las series están compuestas por el dato observado diariamente de la precipitación acumulada en 24 horas entre las 07z y 07z y medido en milímetros. El análisis se lleva a cabo en cada una de las subcuencas (éstas contienen entre 5 y 50 estaciones). La distribución por cuencas hidrográficas de la red pluviométrica del INM es la siguiente:

García Codron, J.C.; Diego Liaño, C.; Fdez. de Arróyabe Hernáez, P.; Garmendia Pedraja, C. y Rasilla Álvarez, D. (Eds.) (2004). **El Clima entre el Mar y la Montaña.** Asociación Española de Climatología y Universidad de Cantabria, Serie A, nº 4, Santander.

Cuenca	Nº de estaciones
Catalana	667
Norte	1.366
Duero	1.225
Tajo	742
Guadiana	922
Guadalquivir	1.349
Sur	494
Segura	358
Levante	938
Ebro	1.712
Baleares	367
Canarias	1.080

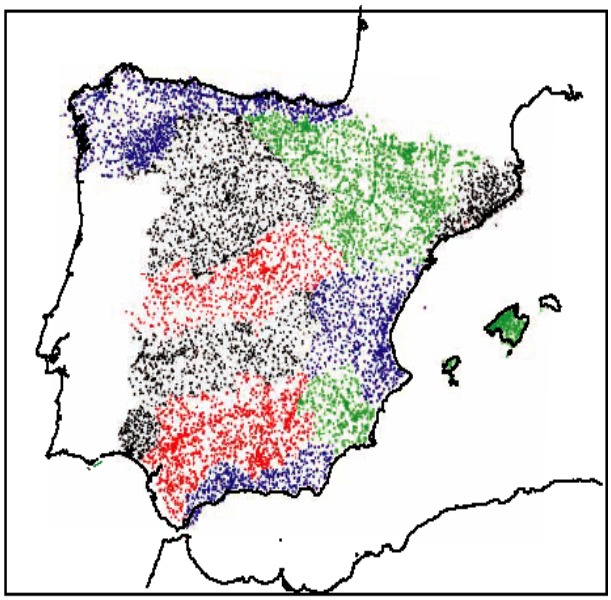

Fig. 1. Red pluviométrica del INM

3. RELLENO DE LAGUNAS

Desde un punto de vista estadístico, el problema de relleno de lagunas a escala diaria es muy similar al problema de la predicción, con la salvedad de que en el primer problema se dispone de más información que en el segundo. Para ilustrar estos problemas, considérese el valor diario acumulado de la precipitación, $X=\{x_1, ..., x_n\}$, en un conjunto dado de estaciones. Ambos problemas tratan de estimar el valor de la precipitación $x_k(T)$ en una estación k y en un instante de tiempo T. Para ello, en un problema de predicción se dispondría como única información, del valor de la precipitación en tiempos anteriores $t<T$ en todas las estaciones (en este caso se supone que no hay lagunas). En cambio, en el caso del relleno de lagunas, se dispondría del valor de la precipitación $x_i(T), \forall\ D(T),$ en el subconjunto de estaciones $D(T)$, con dato para el día T, así como los valores en fechas anteriores, $x_i(t), \forall\ D(t), t<T$.

Este aspecto de disponibilidad o no de la información "circundante" resulta de capital importancia para realizar la estimación de la laguna, ya que este problema puede ser extremadamente difícil en el caso de parámetros y localidades con gran variación espacial y sin información cercana. Existen numerosos métodos de relleno de lagunas en series temporales; sin embargo, la mayoría de ellos utilizan de una u otra forma la hipótesis de normalidad y, por tanto, se aplican preferentemente en datos mensuales (CONRAD y POLLACK, 1962). El más popular de los métodos de relleno es la regresión multivariada. Sin embargo, cuando las lagunas están repartidas por la serie de forma inhomogénea (es decir, los conjuntos $D(t)$ presentan mucha variabilidad) sería necesario estimar toda una batería de modelos de regresión para cada estación, a fin de poder estimar el valor en el tiempo T a partir de cualquiera que fuese $D(T)$. Cuando se trabaja con una red de estaciones muy densa, este es un esfuerzo que resulta inabordable.

García Codron, J.C.; Diego Liaño, C.; Fdez. de Arróyabe Hernáez, P.; Garmendia Pedraja, C. y Rasilla Álvarez, D. (Eds.) (2004). **El Clima entre el Mar y la Montaña.** Asociación Española de Climatología y Universidad de Cantabria, Serie A, nº 4, Santander.

En la siguiente sección se propone un nuevo método que utiliza un único modelo conjunto de todas las variables para rellenar las lagunas (obsérvese que este hecho garantiza ya una cierta coherencia en el relleno de las lagunas).

3.1 Modelo Gaussiano Truncado

El principal inconveniente de trabajar con datos diarios de precipitación es la carencia de normalidad e incluso de correlación significativa entre los datos. Para constatar este hecho se han calculado las correlaciones entre el observatorio de Santander y los observatorios de la red secundaria circundantes, para precipitación y temperatura máxima diarias. La figura 2 muestra que el alcance de la correlación para un umbral del 95% es de 20 km para la precipitación y de 60 km para la temperatura máxima. Esto es válido para la banda litoral del Cantábrico oriental donde el tipo de precipitación típico es de escala frontal y las temperaturas en la franja litoral son muy uniformes; en otras zonas esto es diferente, aunque se ha constatado un comportamiento similar. Los alcances más pequeños corresponden a observatorios aislados y/o con precipitaciones típicamente convectivas para los cuales la metodología propuesta es menos adecuada; aún así, ni en los casos más extremos se han detectado errores superiores al error actual en la predicción a corto y medio plazo.

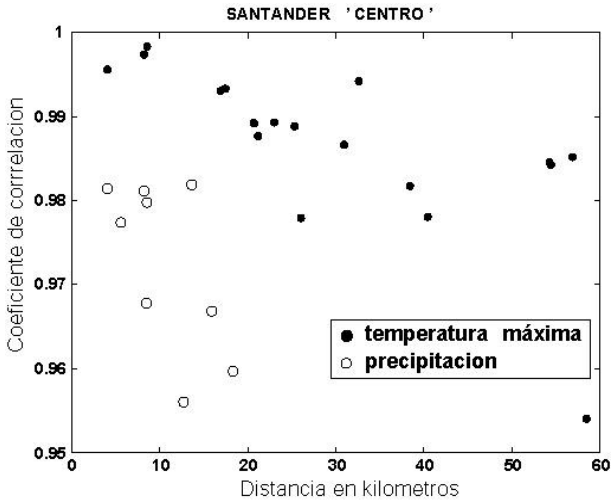

Fig. 2. Correlación entre el observatorio de Santander y los observatorios circundantes en función de la distancia, para precipitación y temperatura

Por tanto, dada la gran densidad de observatorios disponibles se propone, para cada subcuenca, un modelo Gaussiano conjunto para modelizar la precipitación, donde la probabilidad asociada al estado de precipitación nula (único valor con probabilidad no nula) se supone distribuido en una cola izquierda ficticia de la densidad de probabilidad. Es decir, la densidad de probabilidad de la precipitación se aproxima por una distribución Gaussiana truncada (GLASBEY y NEVISON, 1997). Una vez estimado el modelo Gaussiano truncado conjunto $N(X,)$, interesa obtener las distribuciones condicionadas $N(X\backslash D(t)|D(t))$ para cada día t (donde $X\backslash D(t)$ es el conjunto diferencia en X y $D(t)$). De esta forma, se podrá estimar el valor de la precipitación en las estaciones con lagunas $X\backslash D(t)$, a partir de la información existente para esa misma fecha (por ejemplo, se puede considerar la media de la distribución condicionada de cada estación). Supongamos que un día dado queremos calcular la función de probabilidad de $Y= X\backslash D(t)$ condicionado al valor conocido de $Z=D(t)=z$. Entonces, dado que la distribución conjunta es normal, la probabilidad de Y

condicionada a *Z*, también es normal y su media y matriz de covarianzas viene dada por (CASTILLO, GUTIÉRREZ y HADI, 1997, Capt. 6):

$$\bar{Y}_{|Z=z} = \bar{Y} + \Sigma_{YZ}\,\Sigma_{ZZ}^{-1}\,(z - \bar{Z})$$

$$\Sigma_{Y|Z=z} = \Sigma_{YY} - \Sigma_{YZ}\,\Sigma_{ZZ}^{-1}\,\Sigma_{ZY}$$

El algoritmo de relleno de lagunas está basado en estas fórmulas. Aplicado al dato diario de la red pluviométrica del INM, para la serie 1950-1999, consta de los siguientes pasos:

1. Se divide la red pluviométrica en subcuencas hidrográficas; de ésta manera solamente se relacionarán series pertenecientes a la misma subcuenca hidrográfica, independientemente de su distancia.
2. Se determina un umbral para limitar el número máximo de lagunas de cada serie; por ejemplo, seleccionando series con más de 12.000 observaciones válidas en el periodo 1950-1999; es decir, si hay más de un 35% de lagunas, la serie se rechaza.
3. Se calcula la matriz de covarianzas a partir de la muestra, ignorando las lagunas.
4. Para cada fecha se extrae la submatriz de covarianza de las series sin laguna (que actuarán como predictores) y el correspondiente vector de covarianzas cruzadas para generar la matriz de regresión que servirá para estimar simultáneamente las lagunas del día.

Con éste método las lagunas son inferidas utilizando todas las observaciones válidas en observatorios circundantes. El método es equivalente a una regresión múltiple, aunque es más eficiente ya que contiene todas las relaciones de dependencia lineal entre todas las variables, y por tanto todas las posibles regresiones múltiples que se pueden plantear sobre un conjunto de variables. Esta ventaja cobra especial importancia en el caso del problema del relleno de lagunas donde las variables dependientes y las independientes cambian en función de la disponibilidad observacional de la base de datos, exigiendo un tipo diferente de regresión para cada caso.

Se han hecho pruebas de relleno de lagunas en series de precipitación utilizando este método con éxito en áreas donde la proximidad entre estaciones es tan grande que la linealización de la precipitación es, a escala local, una buena aproximación (Fig. 2), ya que aunque la correlación espacial de la precipitación tiene menor alcance que la de la temperatura, la gran densidad de estaciones pluviométricas de la base datos del INM es una ventaja que permite aplicar la interpolación gaussiana en buena parte de los casos, con errores inferiores al error operativo (Fig. 3).

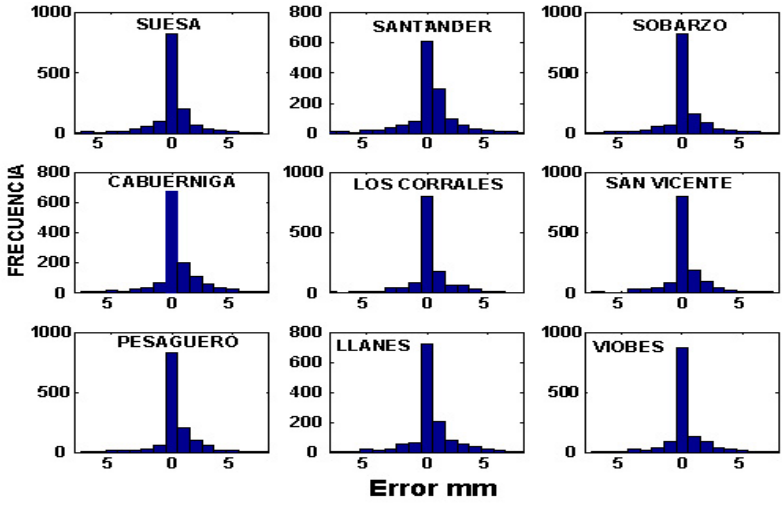

Fig. 3. Distribución de los errores en el relleno de lagunas de la precipitación diaria para varias estaciones de la Cuenca Norte aplicando el método de relleno de lagunas descrito. El conjunto de test consta de 1.500 días tomados al azar de la serie diaria 1950-1999

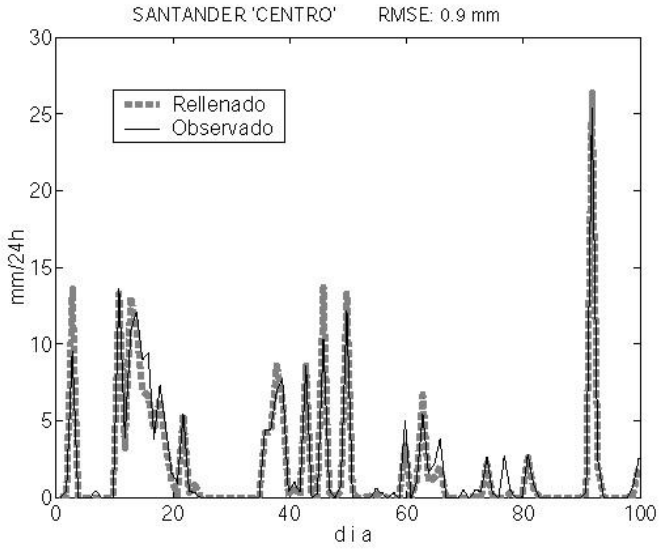

Fig. 4. Relleno de lagunas de precipitación para Santander. Se han tomado como conjunto de test, los primeros 100 días de 1979

4. HOMOGENEIDAD

Los registros históricos de observaciones son en realidad muestras locales de una población extendida espacialmente (la región considerada homogénea). En este artículo se consideran como regiones homogéneas las distintas subcuencas de la Península, pero el método que se presenta a continuación sería igualmente aplicable a otras zonas. Por tanto, cualquier muestra de la subcuenca ha de ser localmente representativa de la población y, por ello, un requisito importante es que las observaciones sean homogéneas a lo largo de toda la serie. Según CONRAD y POLLACK (1962): "*Una muestra es homogénea si sus variaciones responden exclusivamente a las variaciones de la atmósfera*".

La homogeneidad se puede separar en dos componentes. Por una parte respecto de los parámetros de la población: La media y varianza han de ser constantes a lo largo de toda la muestra; y por otra parte, respecto de los parámetros de observación: El instrumento de medida, la ubicación y el

entorno del observatorio han de ser los mismos en todas y cada una de las medidas de la muestra (por ejemplo, en España, a partir de 1910 se adoptó un pluviómetro estándar para todos los observatorios, ALMARZA -1996-). Existen, por tanto, dos fuentes de inhomogeneidad: la que se deriva directamente de la no estacionariedad del sistema observado como resultado de su dinámica (la atmósfera posee ciclos a casi todas las escalas) y la que se deriva del sistema observador (modificación del entorno, del instrumento o de su ubicación). Interesa detectar y, en su caso, eliminar las inhomogeneidades del segundo tipo. Para este propósito han sido desarrollados distintos métodos (BUISHAND, 1982; ALEXANDERSSON, 1986; etc.). Estos métodos están orientados a series mensuales o anuales; para series diarias, que pueden tener 25.000 registros, se plantea un fuerte problema de estimación de valores críticos. Por ello se utilizarán los tests de homogeneidad relativa (THR) y absoluta (THA) adaptados a dato diario descritos en GUTIÉRREZ *et al.* (2004).

4.1. Homogeneización por eliminación de EOFs

Se suele aceptar que dentro de una zona climatológicamente homogénea las fluctuaciones del sistema observado generalmente afectan de una forma conjunta a todos o a la mayoría de los observatorios. Por ello, si se consideran las EOFs de un conjunto de estaciones de la misma subcuenca (por ejemplo, las 30 estaciones en la subcuenca de Santander), entonces sólo serán necesarias unas pocas EOFs para explicar la variabilidad total de la subcuenca. Los modos (EOFs) que componen la variabilidad total, además de ser ortogonales entre sí, son construidos de manera que acumulen sucesivamente la máxima cantidad de varianza del sistema; es decir, el primer modo es el dominante y es el que acumula mayor varianza, etc. De esta manera, las últimas EOFs estarán asociadas a las fluctuaciones del sistema observador de cada observatorio. Por ejemplo, la figura 5 muestra la distribución de la varianza de Ontoria y Santander para cada una de las 30 EOFs. Mientras Ontoria sólo tiene valores significativos en las primeras EOF, Santander presenta varios picos en distintas EOFs que indican un comportamiento inhomogéneo con el resto de estaciones.

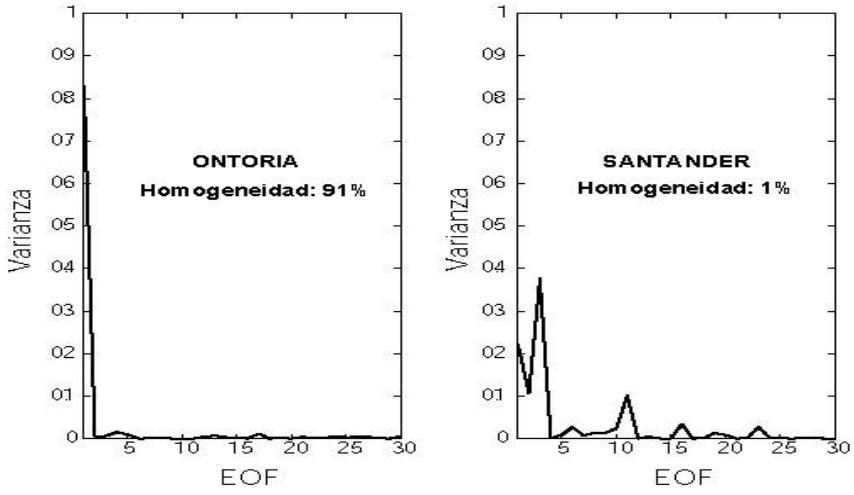

Fig. 5. Distribución de la varianza en las sucesivas EOF para la serie completa diaria de precipitación 1950-1999 para dos estaciones de la Cuenca Norte. La falta de homogeneidad en la serie del observatorio de Santander es fundamentalmente debida a la ubicación del pluviómetro -demasiado cerca del observatorio- y a las variaciones en el entorno urbano donde se encontraba dicho observatorio

Por ello, se propone en esta sección un método basado en el filtrado de las últimas EOFs, ya que éstas contienen las inhomogeneidades de estaciones aisladas. El umbral de filtrado vendrá dado por el porcentaje de varianza explicada que se desea conservar (cuanto mayor sea el porcentaje, menos efecto tendrá el proceso de homogeneización). Al ser el sistema climatológicamente homogéneo, los modos dominantes darán cuenta de las características globales del sistema, mientras que los modos medios y finales mostrarán aproximadamente las características no globales y particulares de cada observatorio. El problema es más complejo cuando se trata de errores sistemáticos durante periodos largos como, por ejemplo, modificaciones en el aparato o cambios de emplazamiento. Por supuesto, la eliminación de estos modos singulares tiene el problema de que junto con las heterogeneidades del sistema observador pueden irse algunas propiedades del sistema observado, razón por la cual no se recomienda el uso de series homogeneizadas en el tratamiento de valores extremos. Una importante ventaja que incorpora este método es que aumenta el fundamento de los métodos de homogeneidad absoluta sobre las series homogeneizadas. Un esquema similar a éste se utiliza para otros fines en el UK Meteorological Office (RAYNER *et al.,* 1996) y en el NOAA Climate Diagnostics Centre (SMITH *et al.,* 1996).

Un ejemplo de los resultados logrados con el método se ilustra en la figura 6, donde se muestran las series de anomalías acumuladas para 2 estaciones de la misma subcuenca, antes y después de aplicar el proceso de homogeneización (conservando un 80% de la varianza). En esta figura puede observarse que el proceso de homogeneización llevado a cabo (demasiado extremo) elimina parcialmente las características singulares de las estaciones.

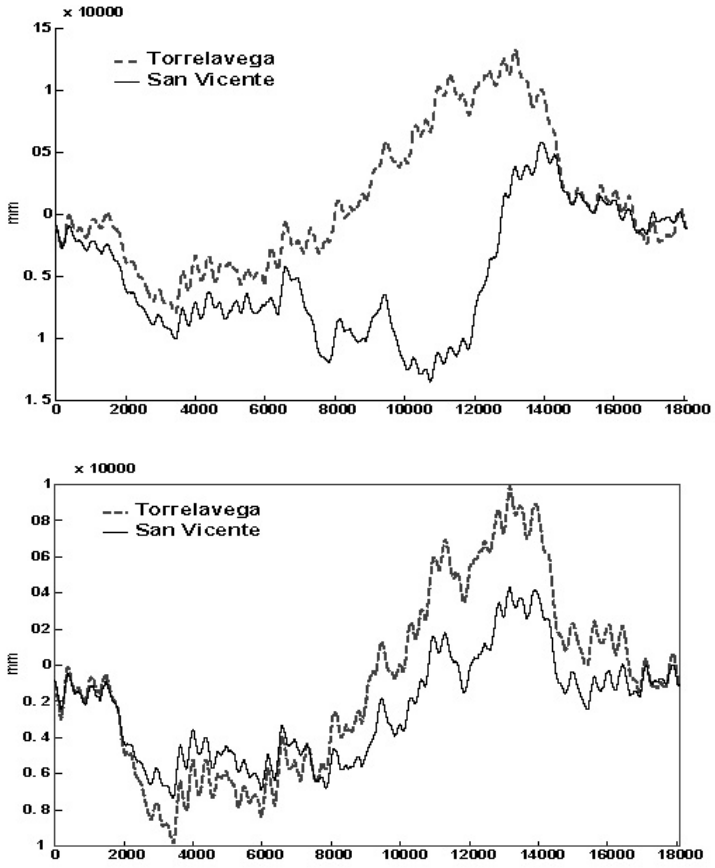

Fig. 6. Series de anomalías acumuladas de algunas estaciones de la Cuenca Norte antes (superior) y después (inferior) del proceso de homogeneización por eliminación de EOFs

En una aplicación real, como la que se muestra a continuación, el porcentaje de varianza retenida oscila entre el 90-95% generando un impacto sobre el dato original que muy raramente supera 1mm (Fig. 7), donde se ilustra la conexión entre el método de relleno de lagunas y el método de homogeneización. Primero se rellenan los datos ausentes (indicados en gris claro) y a continuación se homogeneiza toda la serie. Se puede observar que los errores más pequeños corresponden al segmento rellenado de la serie (pues el proceso de rellenado ya implica cierto grado de homogeneidad con los vecinos utilizados en el relleno).

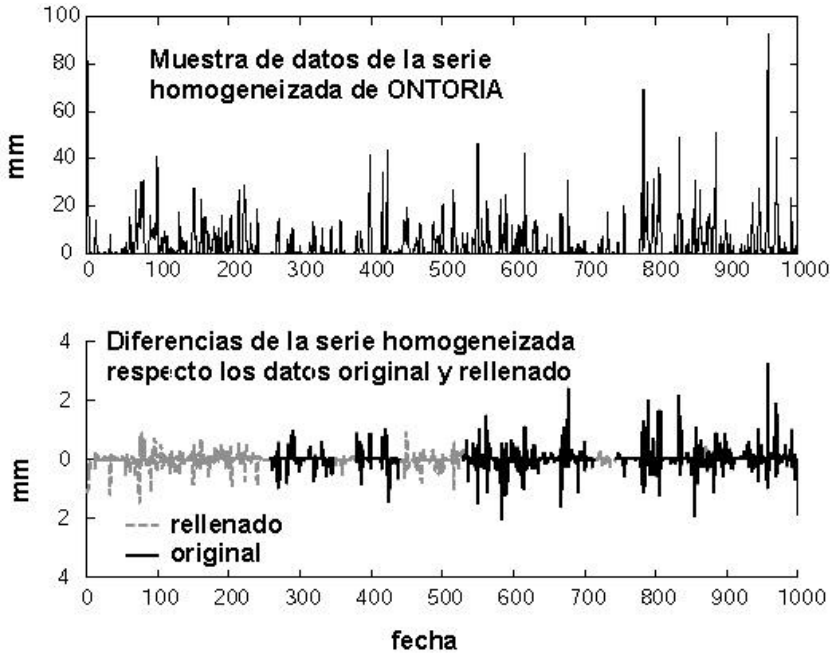

Fig. 7. Ilustración del impacto del proceso completo de relleno y homogeneización. El proceso de homogeneización afecta más a los datos originales que los datos rellenados

5. REFERENCIAS

ALLEN, R.G. *et al.* (1998). *Crop evapotranspiration - Guidelines for computing crop water requirements.* FAO Irrigation and drainage, Paper 56 , M-56 of FAO Technical Papers, FAO, Rome, 1998.

ALEXANDERSSON, H. (1986). "Homogeneity test applied to precipitation data". *Journal of Climatology*, 6, pp. 661-675.

ALMARZA, C.; LÓPEZ, J.A. y FLORES, C. (1996). *Homogeneidad y variabilidad de los registros históricos de precipitación de España*. INM, Monografías, vol. A-143.

BUISHAND, T. (1982). "Some methods for testing the homogeneity of rainfall records". *Journal of Hydrology,* 58, pp. 11-27.

CASTILLO, E.; GUTIÉRREZ, J.M. y HADI, A.S. (1997). *Sistemas expertos y modelos de redes probabilísticas*. Academia de Ingeniería.

CONRAD, V. and POLLACK, L.D. (1962). *Methods in Climatology*. Harvard University Press, vol. 4.2.

CRADDOCK, J. (1979). "Methods of comparing annual rainfall records for climatic purposes". *Weather,* 34, pp. 332-346.

GLASBEY, C.A. and NEVISON, I.M. (1997). Rainfall modelling using a latent Gaussian variable. In: GREGOIRE, T.G. *et al.,* (Eds.). *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions.* Lecture Notes in Statistics 122, Springer, New York, pp. 233-242.

GROISMAN, P.Y. and EASTERLING, D.R. (1994). "Variability and trends of total precipitation and snowfall over the United States and Canada". *Journal of Climate*, 7, pp.184-205.

GUTIÉRREZ, J.M. *et al.* (2004). *Redes Probabilísticas y Neuronales en las Ciencias Atmosféricas.* INM, Monografías, Madrid. 350 pp. (en prensa).

MEKIS, E. and HOGG, W.D. (1999). "Rehabilitation and analysis for Canadian daily precipitation time series". *Atmosphere-Ocean*, 37(1), pp. 53-85.

RAYNER, N.A. *et al.* (1996). *Version 2.2 of the global sea-ice and sea surface temperature data set, 1903-1994.* Hadley Centre for Climate Prediction and Research, report crtn 74.

SMN (Servicio Meteorológico Nacional) (1968). *Situación geográfica e indicativos de las estaciones pluviométricas españolas*. Serie C (37).

SMITH, T.M. *et al.* (1996). "Reconstruction of historical sea surface temperatures using empirical orthogonal functions". *Journal of Climate*, 9, pp. 1403-1420.

STEPANEK, P. (2001). *Anclim-software for time series analysis*. Dept. Geography, Fac. of Natural Sciences, MU, Brno.