CONTROL DE CALIDAD DE DATOS DIARIOS EN LA PENÍNSULA IBÉRICA

A. H. ENCINAS[1], C. RODRIGUEZ PUEBLA[2], A. GARCÍA CASADO[3]
[1]*Dpto de Matemática Aplicada. Universidad de Salamanca*
[2]*Dpto de Física General y de la Atmósfera. Universidad de Salamanca*
[3]*Dpto de Matemáticas. Universidad de Salamanca*
{ascen, concha, lagc}@usal.es

RESUMEN
La calidad de los datos es un requisito necesario en estudios de variabilidad climática, principalmente cuando se analizan datos de gran resolución espacial y temporal. Para analizar los extremos climáticos es preciso utilizar datos diarios, pero estas series temporales evidencian algunos errores debidos a discontinuidades y otros al azar; por ejemplo, que la precipitación sea menor que cero y que la temperatura máxima sea menor que la mínima para un día determinado. El procedimiento para corregir estos errores resulta muy complicado y hemos adoptado los métodos desarrollados en http://cccma.seos.uvic.ca/ETCCDMI/software.shtml, programados en lenguajes R (RClimDex y RHTest) y Fortran (FClimDex y FHTest). Mostraremos algunos ejemplos aplicados a datos observados "in situ" y compararemos la calidad de estos datos frente a la de los datos de reanálisis del NCEP/NCAR.

Palabras clave:
Homogeneización de datos, cambios de tendencia, test de homogeneización, Climdex, Htest, península Ibérica.

ABSTRACT
Data quality has to be considered in studies about climate variability, especially when data have high spatial and temporal resolutions. The extreme climate research requires daily data which very often have errors due to discontinuities and other random causes, for example, we found some negative precipitation values, and maximum temperature lower than minimum. Therefore, we need to adjust the data to correct these errors. The methods to obtain homogeneous data are complicated. We have addressed the procedure developed in http://cccma.seos.uvic.ca/ETCCDMI/software.shtml, that provides software in R (RClimDex and RHTest) and Fortran (FClimDex and FHTest) languages. In this paper we present some examples about the adjustment technique applied to observed data "in situ" and we compare the quality of the observed against the NCEP/NCAR reanalysis data.

Key words:
Data homogenization, step-change points, ClimDex, Htest, Iberian peninsula.

1. INTRODUCCIÓN

Uno de los temas de interés en relación a la variabilidad climática es el análisis de los extremos climáticos, porque tienen un impacto negativo en aspectos socio-económicos. Estos estudios requieren datos diarios y a su vez el uso de estos datos presentan varias dificultades, por una parte la disponibilidad y, por otra, que éstos presentan numerosas inhomogeneidades. Los errores son inherentes a toda actividad de medida, y la Meteorología no es una excepción. Por ello la preocupación por la detección y corrección de errores e inhomogeneidades en las series climatológicas es tan antigua como la climatología misma. A pesar de que hay numerosos grupos de investigación que han dedicado esfuerzos para crear series homogéneas (ABAURREA *et al* (2004), BRUNET *et al* (2001), GUIJARRO (2004), STAUDT *et al* (2007)), cuando iniciamos el estudio sobre extremos climáticos, los datos diarios homogeneizados no estaban accesibles. Por ello, el paso previo para desarrollar la investigación consistió en la adecuación de los datos diarios mediante controles de calidad. Los estudios sobre homogeneización de datos han sido ampliamente desarrollados y el trabajo de Staudt (STAUDT *et al* (2007) proporciona un resumen de los mismos.

Existen dos fuentes de inhomogeneidad: la que se deriva de la no estacionariedad del sistema observado como resultado de su dinámica ya que la atmósfera varía de forma cuasi-cíclica en diferentes escalas; y la que se deriva de la propia observación por haber sido modificado el entorno, el instrumento de medida, su localización, la persona que lo utiliza, etc. En este trabajo identificaremos y corregiremos este segundo tipo de inhomogeneidades, para evitar señales no climáticas que puedan distorsionar y causar interpretaciones erróneas de los resultados.

El objetivo fundamental es obtener series de datos diarios completas y homogéneas a partir de las series de datos cedidas por el Instituto nacional de Meteorología (INM) y el European Climate Assessment & Dataset (ECA&D), además se comparará la calidad de los datos observados con la de los datos del reanálisis del NCEP/NCAR. Los controles de calidad deberían llevarse a cabo en los centros de registro y custodia de los datos, o advertir al posible usuario de los problemas existentes indicando los periodos no fiables. Debido a que los test de calidad no se han aplicado a los datos diarios hemos desarrollado este estudio y mostramos aquí los métodos que usamos, que fueron propuestos en http://cccma.seos.uvic.ca /ETCCDMI/software.shtml.

2. DATOS

Para el análisis de índices de extremos se necesitan datos diarios de precipitación acumulada, temperatura máxima y temperatura mínima. Vamos a determinar la calidad de los datos observados en determinados lugares de la península Ibérica y compararla con la de los del reanálisis del NCEP/NCAR, con el fin de comprobar la posibilidad que ofrecen estos datos regularmente distribuidos para caracterizar los extremos climáticos y relacionarlos con sus causas.

Los datos observados corresponden a los lugares indicados en la Figura 1a. Son 34 lugares distribuidos por toda la península Ibérica. El periodo usado va desde 1949 hasta 2005. Algunos de estos datos fueron proporcionados por los Institutos de Meteorología de España, otros los

distribuye el "European Climate Assessment & Dataset" (ECA&D) (TANK et al. 2002), los simbolizaremos con el acrónimo INM-ECA.

Los datos del proyecto de reanálisis NCEP/NCAR (National Centres for Environmental Prediction/National Centres for Atmospheric Research) los simbolizaremos con el acrónimo NNR (KALNAY et al., 1996; KISTLEr et al., 2001). Se trata de datos distribuidos en una malla regular sobre la península (Figura 1b) de tamaño de rejilla aproximada de 1.9 x 1.9º.

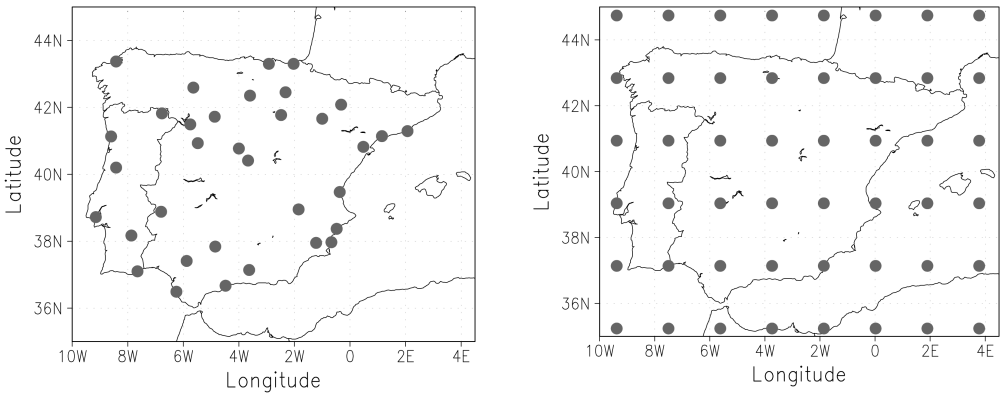

Fig. 1. a) Distribución de datos del INM-ECA; b) Distribución de datos del NCEP/NCAR.

3. METODOLOGÍA

La calidad de los datos se evalúa para cada lugar individualmente. Primero eliminamos aquellas series que tienen un porcentaje de datos faltantes de alguna de las variables, generalmente temperaturas máximas o mínimas. Después las sometemos a diferentes test de homogeneidad.

3.1 Test de homogeneidad Htest

El objetivo de este test es identificar puntos de cambio en los valores medios de las series temporales e indicar la significación de dichos cambios. Este control de calidad se lleva a cabo con el programa Htest. Este test detecta la componente de tendencia. La significación de la componente de tendencia se corrige en función de la autocorrelación de la serie temporal con la versión 2 del Htest (Wang et al. 2007, Wang 2007a). Es decir, con este test se puede determinar los cambios en la componente de tendencia (c.t.) que podrían existir en una serie de datos que además pueden ajustarse a una función autorregresiva de orden 1. Hay veces que los datos presentan variación en los valores medios debido a distintas causas: modificación del aparato de medida, del lugar del observatorio, etc. Estos cambios pueden determinarse mediante el Htest, y una vez conocidos, investigaremos si esos cambios en la media son de origen climático o debido a factores externos. Para ello se puede utilizar una serie de referencia homogénea que esté bien correlacionada con la serie base. La serie de referencia se puede construir de diferentes formas, bien mediante una combinación de series usando diferentes pesos para las estaciones que la componen, o usando una única serie. Sin embargo, HtestV2 detecta c.t. incluso cuando no se dispone de dicha serie de referencia, que podrían ser significativos al nivel del 5%.

Se deben ajustar los cambios significativos de media antes de usar la serie de referencia para comprobar la homogeneidad de la base.

Para las series de precipitación acumulada hemos realizado una transformación logarítmica, para que los datos tengan un comportamiento Gaussiano, que es uno de los requisitos para utilizar el Htest.

Los c.t. que resultan significativos incluso sin metadatos se llaman Tipo-1, y los que resultan de la comparación con los metadatos se llaman Tipo-0, estos sólo habrá que tenerlos en cuenta si los metadatos son fiables.

Si no hay c.t. significativos, la serie temporal comprobada se puede considerar homogénea, no es necesario aplicar más tests y se podrá trabajar con ella.

Nosotros hemos utilizado este test para identificar series muy inhomogéneas y eliminarlas de los posteriores análisis, i.e., si la serie presenta demasiados cambios bruscos no se deberá tener en cuenta a la hora de realizar el estudio de extremos.

3.2 Test de control de calidad ClimDex

Mediante este test se identifican y corrigen errores lógicos que se deducirían de una inspección visual de las series temporales. Por ejemplo, que la temperatura máxima fuese inferior a la mínima, que la precipitación fuese negativa o bien que ocurriesen valores excesivamente altos o bajos, es decir que superen un múltiplo determinado de desviaciones estándar de los datos.
El software se proporciona en la página: http://cccma.seos.uvic.ca/ETCCDMI/software.shtml. Hay versiones en lenguaje R y Fortran, desarrollado y mantenido por Xuebin Zhang y FENA Yang del Departamento de Investigación Climática del Servicio Meteorológico de Canadá.

El objetivo principal del test es construir índices de extremos climáticos para usarlos en el estudio y detección de cambios climáticos, pero incluye un procedimiento simple de control de calidad de datos antes de calcular los índices, que es el que nosotros aprovechamos para comprobar las posibles “incoherencias” en los datos.

El control de calidad de ClimDex desarrolla el siguiente procedimiento: a) Reemplaza todos los datos faltantes en un formato interno que reconoce el programa. b) Remplaza todos los valores no razonables por NA (not a number). Estos valores incluyen cantidades de precipitación diarias menores que cero y temperatura máxima diaria menor que temperatura mínima diaria.

El control de calidad de los datos consiste en primer lugar en determinar que la temperatura máxima sea menor que la temperatura mínima, que la precipitación sea menor que 0, que existan días en los que las tres variables tengan un valor 0, o que falten datos.

Adicionalmente, el control de calidad también identifica valores extremos (outliers) en temperaturas diarias máximas y mínimas. Los valores extremos son valores diarios que se encuentran fuera de una región definida por el usuario. Actualmente esta región se define como n (entrada del usuario) veces la desviación estándar (std) del valor del día, esto es, [media-n*std, media+n*std]. Hemos considerado n = 3.

El programa proporciona los gráficos de las series temporales e indica los valores faltantes. Si una serie presenta demasiados errores, o demasiados datos faltantes, nosotros la eliminamos para los posteriores estudios de extremos climáticos.

4. RESULTADOS

4.1 Ejemplo del test de homogeneidad Htest

Se utiliza este test para identificar los cambios en las componentes de tendencia de la serie. Entre los archivos de salida del Htest hay uno que nos indica dónde pueden estar los cambios en los valores medios, así como la serie modificada. Por ejemplo en la serie de temperatura máxima de Salamanca (Figura 2) obtendríamos dos cambios: uno para el 3 de diciembre de 1950 y otro para el 28 de enero de 1969:

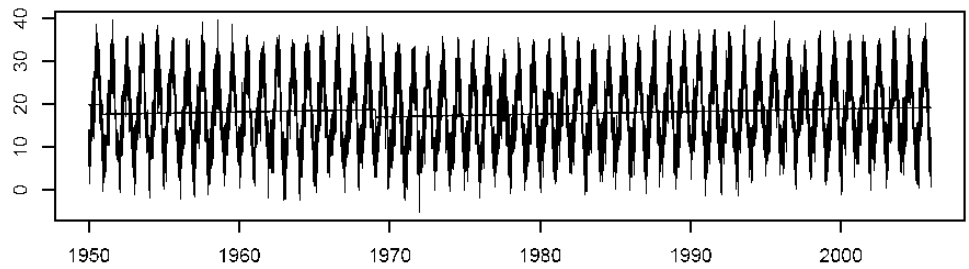

Fig. 2: Cambio en los valores medios en la serie de temperatura máxima de Salamanca

En la Figura 3 vemos la representación de la serie de temperatura máxima corregida, sin puntos de cambios en la tendencia.

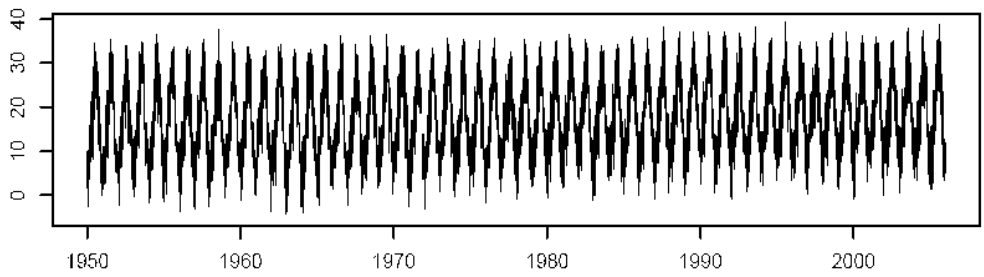

Fig. 3: Serie de temperatura máxima de Salamanca modificada

Las series observadas INM-ECA presentan discontinuidades en precipitación (9%), en temperatura mínima (71%) y en temperatura máxima (76%).

Para los datos del NNR, sólo se presentan 2 discontinuidades en los valores de temperatura mínima una el 31 de agosto de 1969 y otra el 19 de noviembre de 1990 para 23 de los 48 lugares (48%).

4.2 Ejemplo de aplicación del test de control de calidad de los datos ClimDex

El orden de aplicación de los procedimientos de control es importante. En primer lugar realizamos un control absoluto "rápido" que analiza la evolución de las medidas, se identifica la incoherencia en algunos datos y la ausencia de medición en días concretos. Estos indicadores junto con la representación gráfica, facilitan la inspección de la serie inicial.

El programa incorpora herramientas de visualización en la que muestra los datos faltantes. Un ejemplo correspondiente a la serie de la temperatura máxima diaria en Albacete se presenta en la Figura 4, en la que los valores faltantes aparecen como un círculo en la línea inferior de la gráfica.

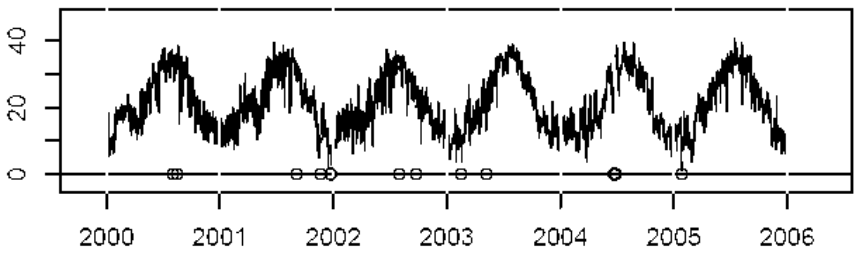

Fig. 4: Representación de Temperatura máxima diaria de Albacete de 2000 a 2006

En segundo lugar la serie pasa por el control de calidad de los datos que indica los valores extremadamente altos o bajos (hemos elegido un desviación estándar igual a 3); si las tres variables aparecen con un valor 0; si hay precipitaciones negativas; y si las temperaturas máximas son inferiores en un día concreto a las temperaturas mínimas de ese día.

El 50% de las estaciones del INM-ECA presentan alguna discontinuidad de este tipo, sin embargo las series de datos del NNR pasan el control de calidad.

En la Figura 5, se muestra un ejemplo de serie en la que existen algunos días con la temperatura mínima mayor que la temperatura máxima, lo que evidentemente es un error, probablemente de transcripción. Según lo que interese para el estudio que se esté haciendo se pueden dejar los valores que el programa da por defecto (-99.9), sustituir esos valores por el valor medio de ese día, por una interpolación considerando los valores del entorno, mediante series de referencia, etc. En concreto se muestra la serie de datos de temperaturas extremas (máxima es la línea discontinua y mínima la línea continua) en agosto de 1992 en Cádiz, del INM-ECA. Como se puede observar en la parte izquierda (Figura 5 a) existen algunos días en los que la temperatura mínima supera la máxima, sobre todo a partir del día 12, en la parte derecha (Figura 5 b), se han sustituido los valores erróneos por la media diaria del mes de agosto en el día correspondiente. Por ejemplo el día 12, los valores cedidos por el INM-ECA

eran Tx=19.8ºC y Tn=20.9ºC, y han sido sustituidos por Tx=25.3ºC y Tn=18.6ºC (valores medios del 12 de agosto de las temperaturas máxima y mínima respectivamente en la serie de Cádiz).

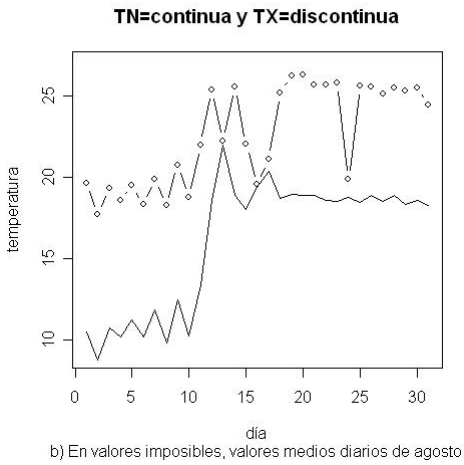

Fig. 5: Temperaturas extremas en agosto de 1992 en Cádiz.

5. CONCLUSIONES

En este trabajo presentamos el estudio realizado sobre las series de datos diarias de precipitación, temperatura máxima y temperatura mínima procedentes del Instituto Nacional de Meteorología y completadas con las de procedencia del European Climate Assessment and Dataset (INM-ECA) y las del reanálisis del NCEP/NCAR (NNR). Se pretende identificar las inhomogeneidades que muestran las series bien por reubicación de los observatorios meteorológicos, por cambios en el instrumento de medición, por cambios bruscos en el entorno inmediato o cualquier otra circunstancia. Sin embargo, el método puede identificar inhomogeneidades derivadas de la propia dinámica de las series climáticas como cuasi-ciclos en diferentes escalas y tendencias climáticas. Nuestro trabajo ha consistido en identificar las variaciones no climáticas para excluir estas series de posteriores análisis sobre el cálculo de los extremos climáticos. El test de in-homogeneidad Htest nos permitió identificar aquellas series que deberíamos descartar. Según este test los datos observados INM-ECA presentan in-homogeneidades del 9% para la precipitación, del 71% para la temperatura mínima y del 76% para la temperatura máxima. Para los datos del NNR, las series de temperatura máxima y precipitación son homogéneas. Sólo se presentan 2 discontinuidades en el 48% de las series de temperatura mínima una el 31 de agosto de 1969 y otra el 19 de noviembre de 1990.

El test de control de calidad ClimDex nos proporcionó los siguientes resultados: el 50% de los datos observados INM-ECA no superan el control de calidad. Sin embargo los datos del reanálisis NNR sí que lo superan.

Los datos observados INM-ECA son reales pero en ellos existen errores lo cual complica los estudios sobre la detección del cambio climático y las relaciones entre la variabilidad climática regional con las causas de las variaciones. Por el contrario, los datos del NNR aunque no son

datos “in situ”, son más homogéneos y están distribuidos en una malla regular, aunque presentan el inconveniente de su baja resolución.

6. AGRADECIMIENTOS

Expresamos nuestro agradecimiento a los proveedores de datos: el Instituto Nacional de Meteorología de España y Portugal, al “European Climate Assessment & Dataset” (ECA) por los datos in situ y al NCEP/NCAR por los datos de reanálisis. Agradecemos a los desarrolladores de los programas ClimDex y Htest utilizados en esta investigación. Este trabajo ha sido financiado con los proyectos de investigación MEC-CGL2005-06600-CO-01/CLI, de la Junta de Castilla y León SA039/A05, cofinanciados con fondos europeos y con el proyecto de ayudas a la investigación “Memoria D. Samuel Solórzano Barruso” de 2008.

7. REFERENCIAS

ABAURREA, J. *et al* (2004). Metodología para el control de calidad y homogeneidad de una base de datos de precipitación diaria. *El clima entre el mar y la montaña.* Publicaciones de la AEC. Serie A, nº 4.

ALEXANDERSSON, J. y ANDERS, M. (1997). Homogenization of Swedish temperature data. Part I: homogeneity test for linear trends. *International Journal of Climatology*, 17: 25-34.

BRUNET, M. y LÓPEZ, D. (2001). *Detecting and modelling regional climate change*. Ed Springer

CANO, R. y GUTIERREZ, J. M. (2004). Relleno de lagunas y homogeneización de series de precipitación en redes densas a escala diaria. *El clima entre el mar y la montaña*. Publicaciones de la AEC. Serie A, nº 4.

GUIJARRO J. A. (2004). Climatol: Software libre para la depuración y homogeneización de datos climatológicos. *El clima entre el mar y la montaña*. Publicaciones de la AEC. Serie A, nº 4.

GUTTMAN, N. B. (1998). *Homogeneity, Data Adjustments and Climatic Normals*.

HYNDMAN, R. J., y FAN Y. (1996): Sample quantiles in statistical packages. *The American Statisticiam,* 50: 361-367.

LÓPEZ, J.A. *et al* (2007). *Métodos y técnicas en Climatología*. La climatología española. Pasado, presente y futuro. Prensa Universitaria de Zaragoza.

STAUDT, M. *et al* (2007). Homogenization of long-term monthly Spanish temperature data. *International Journal of Climatology,* 27: 1809-1823.

WANG, X. L. and FENG Y., (2004): *RHTest User Manual*

ZHANG, X., G. *et al* (2005). Avoinding inhomogeneity in percentile-based indices of temperature extreme. *Journal of Climate,* 18: 1641-1651