ESTIMACIÓN DE PARÁMETROS DE LA DISTRIBUCIÓN ESTADÍSTICA DE TEMPERATURAS MEDIAS MENSUALES A PARTIR DE FICHEROS DE DATOS INCOMPLETOS

Enric Terradellas
Agencia Estatal de Meteorología, carrer de l'Arquitecte Sert, 1, 08005 Barcelona,
enric@inm.es

Resumen

La falta de continuidad en las observaciones climatológicas complica notablemente el tratamiento de las series climáticas. Operaciones tan simples como la aplicación de pruebas de homogeneidad o la estimación de parámetros de la distribución estadística de los datos se convierten en problemas complejos. En este trabajo se describe un método completo de tratamiento de datos que parte de los registros mensuales de los observatorios y concluye con la obtención de los parámetros estadísticos de las series climáticas. El método incluye control de calidad de los datos, imputación de valores ausentes, estimación de parámetros de la distribución estadística y homogeneización de las series. El control de calidad consiste en una serie de filtros basados en la dispersión de los datos que permiten identificar tanto series como datos individuales sospechosos de ser erróneos. La imputación de valores ausentes y la estimación de los parámetros de la distribución se realizan conjuntamente mediante un método basado en la maximización de la esperanza. Finalmente, la homogeneización de las series se realiza a partir de los resultados obtenidos con el *standard normal homogeneity test*. El método descrito se aplica a las series de valores medios mensuales de las temperaturas extremas diarias registradas en Cataluña durante el periodo 1971-2000.

1. Introducción

En los últimos años, a excepción de un reducido número de observatorios, la gran mayoría de estaciones meteorológicas están constituidas por sistemas automáticos de observación o son atendidas por personal voluntario. Esta situación posibilita el establecimiento de redes de observación relativamente densas. Sin embargo, las series climáticas obtenidas en ellas carecen, muy a menudo, de la continuidad deseable.

Esta circunstancia complica notablemente el tratamiento de las series climáticas, dificultando operaciones simples como la estimación de los parámetros de la distribución estadística. En particular, una estimación de la matriz de covarianza de distintas series obtenida usando sólo los datos disponibles y sin considerar los términos con datos ausentes puede no ser simétrica y definida positiva, condiciones necesarias para la aplicación de técnicas estadísticas multivariantes tales como el análisis de componentes principales, el análisis discriminante o el análisis de conglomerados (Wilks, 1995).

La forma más elemental de evitar las dificultades que presentan los ficheros de datos incompletos es la eliminación de todos los registros en los que falta algún dato. Desde este punto de vista, al abordar estimaciones de parámetros estadísticos de series climáticas, deberían ser excluidas del análisis todas las series en las que faltara algún dato. Ello es totalmente inviable en el caso de ficheros con un elevado porcentaje de datos ausentes y, por tanto, es necesario el desarrollo de métodos que permitan estimar los parámetros de la distribución estadística en series climáticas a partir de ficheros de datos incompletos. Estos métodos suelen basarse en la estimación previa de estos valores ausentes. Una buena recopilación de las técnicas que se han empleado para la estimación de valores ausentes puede leerse en Ramos-Calzado et al. (2008).

La estimación de los parámetros estadísticos de la serie y la imputación de valores ausentes pueden considerarse como dos problemas íntimamente relacionados. En efecto, el valor de los datos ausentes puede ser estimado una vez definidos los parámetros estadísticos de la serie (Buck, 1960). Y, por otra parte, los parámetros estadísticos de la serie pueden ser calculados una vez conocidos los valores estimados de los datos ausentes y la matriz de covarianza del error de la estimación (Little y Rubin, 1987). Desde este punto de vista, parece lógico abordar el problema de forma iterativa. Schneider (2001) lo hace utilizando el algoritmo de maximización de la esperanza (ME) (Dempster et al., 1977).

Por otro lado, cualquier análisis de datos climáticos, muy especialmente si se dirige a estudiar la dinámica del clima, debe basarse en series de datos homogéneas. Una definición muy

simple de serie homogénea es la que dieron Conrad y Pollack (1950): una serie cuyas variaciones son causadas sólo por las variaciones en el tiempo y el clima. Sin embargo, es bastante frecuente que las series climáticas sufran variaciones por otras causas, que pueden ser cambios en el emplazamiento del observatorio o las condiciones físicas de su entorno, el instrumental, los métodos de observación, etc. (Karl y Williams, 1987).

Las técnicas empleadas para detectar y corregir inhomogeneidades de origen no climático en las series son muy diversas. Algunas han sido diseñadas para identificar inhomogeneidades originadas por un determinado factor y, por tanto, son útiles solamente para cierta variable climatológica o para determinados ámbitos geoclimáticos, o aplicables únicamente en redes de estaciones más o menos densas o con series que reúnan determinadas características de longitud o disponibilidad de metadatos. Una buena revisión de los distintos métodos que se han empleado para tratar las inhomogeneidades en las series climáticas puede leerse en Petterson et al. (1998). La elección de la mejor técnica dependerá de los datos disponibles pero también del objetivo del análisis que se esté realizando. Distintas técnicas de homogeneización producen resultados distintos pero normalmente las series homogeneizadas siguiendo distintos métodos se parecen mucho más entre sí que con la serie original (Petterson et al., 1998).

La técnica de homogeneización más universalmente empleada es la basada en el *standard normal homogeneity test* (SNHT) inicialmente desarrollado por Alexandersson (1986), aunque con modificaciones posteriores introducidas para tener en cuenta la posible presencia de más de una discontinuidad o para permitir la existencia de una tendencia en la serie (Alexandersson and Moberg, 1997).

La aplicación de la mayor parte de las técnicas de homogeneización presenta problemas en caso de que las series sean incompletas. Podría optarse por estimar los valores ausentes antes de proceder a la homogeneización de las series. No obstante, cuando el porcentaje de valores ausentes es muy grande, las estimaciones de estos valores ausentes pueden alterar los resultados de las pruebas de homogeneidad

En este trabajo se aborda el tratamiento de las series de valores medios mensuales de las temperaturas extremas diarias registradas en Cataluña entre los años 1971 y 2000. Se analizan los registros de 215 estaciones, que presentan una ausencia media del 38% de los datos. Los datos usados para el análisis se describen en la sección 2, el método de tratamiento de las series en la sección 3 y los resultados obtenidos en la sección 4. Finalmente la sección 5 contiene las conclusiones y algunas indicaciones sobre futuras líneas de trabajo.

2. Datos

Para abordar el estudio de los valores medios mensuales de las temperaturas extremas diarias en Cataluña durante el periodo 1971-2000 se han analizado los datos registrados en 215 estaciones pertenecientes a la red climatológica de la Agencia Estatal de Meteorología (AEMET). Se han considerado aquellas estaciones de las que se disponía de un mínimo de 10 años de datos. Tras el proceso de control de calidad que se describe en el apartado 3.1 se han descartado los datos de 10 estaciones, con lo que se ha elaborado el estudio utilizando las series de las 205 estaciones restantes. De ellas, 23 se encuentran situadas fuera del territorio catalán (22 en Aragón y 1 en la Comunidad Valenciana) y han sido incluidas para permitir que futuros estudios de distribución espacial de temperaturas extremas diarias tengan suficiente cobertura de datos en las zonas fronterizas. La figura 1 muestra la distribución geográfica de los observatorios cuyos datos han sido utilizados para el presente estudio. Hay que destacar que la base de datos de valores mensuales contiene un 38% de datos ausentes.

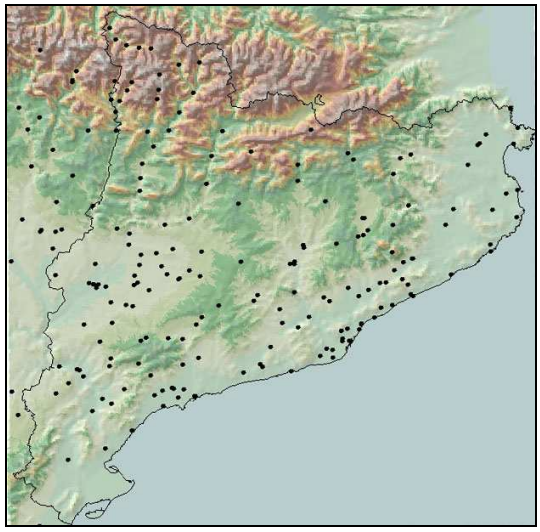

Fig. 1. Distribución geográfica de los observatorios cuyas series climáticas se han empleado en el estudio.

3. Método de tratamiento de datos

El tratamiento de los datos se inicia con controles que permiten descartar las series que presentan un deficiente nivel de calidad e identificar datos individuales cuya validez se considera dudosa. Los procedimientos empleados en este control de calidad se detallan en el apartado 3.1. A

continuación se procede a estimar el valor de los datos ausentes y de los parámetros estadísticos de las series usando el procedimiento descrito en Schneider (2001) y que se detalla en el apartado 3.2. Después, una vez completadas las series con valores estimados, se procede a una primera homogeneización de las mismas, utilizando el SNHT (apartado 3.3). Puesto que el proceso de homogeneización altera algunos de los valores de las series originales, deberá procederse a una nueva estimación de los valores ausentes y de los parámetros estadísticos. Se entra, pues, es un proceso iterativo en el cual se impone a las series un nivel de homogeneidad cada vez más alto.

3.1. Control de calidad

En primer lugar, se aplica a todos los datos un filtro para identificar aquellos valores que se alejan más del valor medio de la serie correspondiente. Para ello, se calculan los valores medios y las desviaciones típicas de las series originales (incompletas). Se consideran sospechosos de ser erróneos aquellos valores que se alejan del valor medio en más de cuatro veces la desviación típica. Este mismo límite había sido considerado previamente por Aguilar et al. (1999). En cualquier caso, antes de descartar cualquier dato se han consultado los registros originales y se ha comparado con los registros correspondientes a observatorios cercanos. De todas formas, hay que mencionar que todos los datos que quedaban fuera del rango considerado han sido finalmente descartados, aunque algunos han podido ser recuperados al acudir a los registros originales, de los que habían sido transcritos incorrectamente

Posteriormente, se establece un segundo sistema de control que permite identificar tanto series de calidad dudosa como valores individuales sospechosos de ser erróneos. El método parte de la idea de que una varianza excesivamente alta en la serie es un indicio de falta de homogeneidad de la misma o de la presencia de valores erróneos. Se podría pensar en establecer un límite en el valor de la varianza y rechazar, o investigar detenidamente, las series que superaran ese umbral. Sin embargo, se puede constatar fácilmente que existen áreas geoclimáticas donde la variabilidad -y por tanto la varianza de las series- es muy superior que en otras. Por ello, ha parecido más razonable realizar previamente una modelización de la varianza de las series en función de variables fisiográficas y establecer el límite en el residuo de la varianza, es decir, la diferencia entre la varianza calculada de la serie y la estimada por el modelo.

El modelo de varianza se ha construido a partir de una regresión lineal múltiple entre la varianza de las distintas series mensuales y los valores de posición (coordenadas U.T.M.), altitud, distancia al mar y diferencia entre la altitud de la estación y la altitud media del territorio en radios de 10 y 20 km. Los coeficientes de correlación no son muy elevados, oscilando entre 0.49 en diciembre y 0.62 en octubre para las temperaturas máximas y entre 0.17 en julio y 0.52 en febrero para las temperaturas mínimas. En cualquier caso, es evidente que imponer un límite a los valores de los residuos ofrece mejores resultados que imponerlo directamente a los valores de la varianza La aparición en una estación de un residuo mensual anormalmente alto es síntoma de la presencia de un valor erróneo en la serie correspondiente, mientras que la aparición de varios residuos mensuales anormalmente altos es síntoma de inhomogeneidad de la serie. En el presente estudio, el límite para investigar las series en profundidad se ha establecido en $3ºC^2$. Además, también se han investigado algunos casos en que todos los residuos son inferiores al umbral, pero uno de ellos es muy superior a los demás.

Una vez localizadas las series que presentan una excesiva varianza, se comparan con las correspondientes a observatorios cercanos. Ello permite identificar de forma subjetiva series y datos individuales sospechosos de ser erróneos. Durante este proceso, se han eliminado o corregido algunos valores mensuales y se han descartado definitivamente 10 series.

3.2. Estimación de valores ausentes y de los parámetros estadísticos de las series

La estimación de valores ausentes y de los parámetros estadísticos de las series se realiza conjuntamente, de forma iterativa, siguiendo el método descrito en Schneider (2001) que, como se ha mencionado en el capítulo 1, utiliza el algoritmo ME.

En primer lugar, es importante mencionar una hipótesis que se considera de forma implícita siempre que se aborda la estimación de valores ausentes por métodos estadísticos. Consiste en que la probabilidad de que un dato esté ausente no depende del valor de la variable (Rubin, 1976). La hipótesis no se verificaría, por ejemplo, en caso de que valores muy extremos de temperatura dificultaran la realización de las medidas, aumentando así la probabilidad de no disponer del dato mensual. En el presente trabajo no se han hallado indicios de que no se verifique la hipótesis.

En series que presentan una distribución gaussiana, el algoritmo ME comienza con una estimación de los valores medios y de la matriz de covarianza y sigue, en pasos alternos, con una estimación de los valores ausentes y una reestimación de los parámetros estadísticos.

Siguiendo la notación de Schneider (2001), llamamos μ' a la estimación del vector de valores medios y Σ' a la estimación de la matriz de covarianza obtenidos en un determinado paso de la iteración. Para un cierto registro con valores ausentes, descomponemos la matriz Σ' en cuatro submatrices: Σ'_{pp}, Σ'_{pa}, Σ'_{ap} y Σ'_{aa}, correspondiendo el subíndice 'a' a variables cuyo valor está ausente en el registro considerado y el subíndice 'p' a variables cuyo valor está presente en el mismo. Análogamente, descomponemos el vector μ' en dos subvectores: μ'_p y μ'_a. Para estimar los valores ausentes, se establece una regresión lineal múltiple entre las series con valor ausente en el referido registro y el resto de series. Los coeficientes B' de esta regresión serán:

$$B' = \Sigma'^{-1}_{pp} \Sigma'_{pa} \qquad (1).$$

Con ello, los valores x_a ausentes en el registro se podrán estimar fácilmente a partir de los presentes x_p mediante la expresión:

$$x'_a = \mu'_a + (x_p - \mu'_p) B' \qquad (2).$$

Y también se podrá estimar la matriz de covarianza del error de la imputación:

$$C' = \Sigma'_{aa} - \Sigma'_{ap} \Sigma'^{-1}_{pp} \Sigma'_{pa} \qquad (3).$$

Una vez completadas las series con los valores estimados, se podrán obtener fácilmente nuevas estimaciones de μ y Σ. Para estimar la nueva matriz de covarianza es muy importante tener en cuenta el error en la imputación:

$$\Sigma' = \frac{1}{n-1} \sum_{i=1}^{n} (x_p^T x_p + x_p^T x'_a + x'^T_a x_p + x'^T_a x'_a + C' - \mu'^T \mu) \qquad (4)$$

donde el sumatorio se extiende sobre todos los registros de las series (i=1 … n).

Como se demuestra en Little y Rubin (1987), el algoritmo ME converge siempre, aunque de forma lineal, por lo que suele ser necesario realizar un elevado número de iteraciones.

El algoritmo ME habitualmente debe ser modificado para su aplicación práctica en estudios climatológicos. Como se ha visto, la estimación de los valores ausentes en la serie de datos de un observatorio se realiza a partir de una regresión lineal múltiple de la propia serie con las correspondientes al resto de observatorios. Y como el número de estaciones suele ser mucho mayor que el número de datos contenidos en cada serie, los parámetros de las regresiones quedan indeterminados. Schneider (2001) recurre al método de regularización de Tijonov (Tijonov y Arsenin, 1977) en el cual se reduce progresivamente la amplitud de los componentes de pequeña escala en los coeficientes de la regresión.

De esta manera, en cada iteración y para cada registro con valores ausentes, la matriz inversa Σ'^{-1}_{pp} que aparece en la ecuación (1) se sustituye por una matriz inversa regularizada:

$$(\Sigma'_{pp} + h^2 D')^{-1} \rightarrow \Sigma'^{-1}_{pp} \qquad (5)$$

donde D' es la matriz diagonal de Σ'_{pp} y h es el parámetro de regularización.

El parámetro de regularización debe seleccionarse de manera que se minimice el error en la estimación de los valores ausentes. Cuanto más pequeño es el valor de h, tanto más afectados por ruido están las estimaciones de estos valores ausentes. Por el contrario, cuanto mayor es el valor de h, tanto más cerca se encuentran los valores estimados de los valores medios de las respectivas series. En el presente estudio, el parámetro se ha escogido siguiendo el método de validación cruzada generalizada descrito en Golub et al. (1979).

La introducción de la regularización en el algoritmo ME lleva implícita una subestimación del error en la imputación. En efecto, la incertidumbre sobre la idoneidad tanto del modelo de regresión como del propio parámetro de regularización contribuyen al error en la imputación pero no se tienen en cuenta en su estimación, de modo que el error en la imputación obtenido debería considerarse únicamente como un límite inferior del error real. La infravaloración de este error dependerá de factores como la relación entre el número de series y su longitud, las varianzas de las series y, sobre todo, el porcentaje de valores ausentes.

3.3. Homogeneización de las series

Se procede, en primer lugar, a construir, para cada observatorio, la serie de todos los valores mensuales consecutivos (Xi). Estas series se normalizan (Zi), restando a cada valor el correspondiente valor medio mensual y dividiendo por la desviación típica. A

continuación, se calcula la matriz de coeficientes de correlación entre las distintas series, para cada estación se identifican las 12 series mejor correlacionadas con la propia (Zi:j) y se construyen las series diferenciales:

$$Zij = Zi - Zi:j \qquad (6).$$

El SNHT se aplica a las series Zi para comprobar la homogeneidad absoluta y a las series Zij para comprobar la homogeneidad relativa. Con el fin de poder detectar múltiples puntos de ruptura, cada una de las series Zi y Zij, que contienen 360 elementos, se descompone en 12 subseries parcialmente superpuestas de 60 elementos, de modo que el SNHT se aplica finalmente a 205 x 13 x 12 = 31980 subseries En estas subseries se consideran los puntos de ruptura que tengan una significación superior al 99%.

En un primer proceso de homogeneización se corrigen las series que presentan simultáneamente una inhomogeneidad absoluta y varias inhomogeneidades relativas con una ruptura en el mismo punto o en un punto muy cercano. La corrección se basa en la diferencia de medias entre los 30 elementos anteriores y los 30 elementos posteriores de la propia serie y de las series de diferencias que presentan la ruptura. Las diferencias de medias se ponderan según el coeficiente de correlación. En este primer paso no se corrigen las series con un elevado número de valores imputados alrededor del punto de ruptura.

Una vez finalizado este primer proceso de homogeneización, se eliminan los valores imputados de las series y se procede de nuevo a estimar los valores ausentes y los parámetros de la distribución estadística.

En nuevos y sucesivos pasos de iteración se corrigen series que presentan al menos 8, 6 y 5 inhomogeneidades relativas con la ruptura en el mismo punto o en uno cercano y se corrigen de acuerdo al procedimiento descrito anteriormente.

Se ha establecido un criterio para valorar los resultados obtenidos después de cada uno de los procesos de homogeneización y posterior estimación de valores ausentes y parámetros de la distribución estadística Se trata de establecer una regresión lineal múltiple entre los valores medios obtenidos y los valores fisiográficos mencionados en el apartado 3.1. Se ha considerado que una mejora en el coeficiente de correlación era indicativa de una mayor calidad en las series. El argumento se basa en la idea de que las inhomogeneidades en las series constituyen uno de los factores que hacen descender la correlación entre factores fisiográficos y parámetros climáticos.

4. Resultados

Mediante el método descrito en el capítulo anterior se han estimado los parámetros estadísticos de las series de valores medios mensuales de temperaturas extremas diarias en 182 observatorios catalanes y 23 ubicados en comarcas limítrofes. También se han estimado los valores de los datos ausentes en las series.

Un ejemplo de los resultados obtenidos puede verse en la Figura 2, donde se muestra la distribución espacial de los valores medios de las temperaturas máximas diarias en abril. La representación se ha realizado siguiendo el método descrito en Téllez et al. (2008) que consiste en la combinación de un modelo de regresión lineal múltiple entre los valores medios de la variable climatológica y distintos factores fisiográficos. y una interpolación de los residuos de este modelo mediante técnicas de *kriging*.

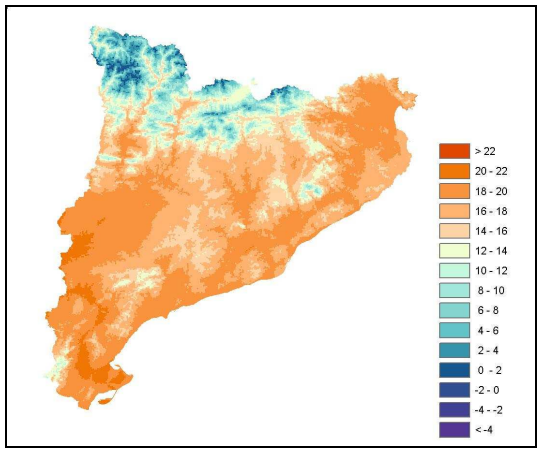

Fig. 2. Distribución geográfica del valor medio de la temperatura máxima diaria (en ºC) durante el mes de abril en Cataluña.

Es interesante mencionar los resultados obtenidos para el coeficiente de correlación entre los valores medios de temperatura y las variables fisiográficas para distintos niveles de homogeneización de las series. Tanto para temperaturas máximas como para temperaturas mínimas diarias, casi todos los meses se sigue el patrón de variación que puede verse en la tabla 1. Tras una primera homogeneización de las series (columna 1 de la tabla) se obtiene un coeficiente de correlación inferior al obtenido con las series sin homogeneizar (columna 0). No obstante, posteriores procesos de homogeneización consiguen aumentar el coeficiente de correlación hasta un valor máximo (columna 3). No se ha encontrado una explicación para este comportamiento, y en particular para la

disminución de la correlación tras la primera homogeneización. En cualquier caso, se han considerado como definitivos los resultados obtenidos tras el tercer proceso de homogeneización.

	0	1	2	3	4
Ene	0.897	0.893	0.894	0.897	0.895
Abr	0.899	0.896	0.898	0.902	0.901
Jul	0.770	0.765	0.764	0.770	0.769
Oct	0.935	0.931	0.933	0.937	0.936

Tabla 1. Coeficientes de correlación entre los valores medios de las temperaturas máximas diarias y distintos parámetros fisiográficos para distintos niveles de homogeneización de las series.

Una vez finalizado el análisis, se ha realizado una estimación del error real en la imputación de valores ausentes utilizando un método de Monte Carlo. Para ello se han eliminado pequeños porcentajes de datos escogidos aleatoriamente y se han estimado sus valores mediante el algoritmo ME regularizado. A continuación se han comparado los valores obtenidos con los originales. Se ha llegado a la conclusión de que los errores reales en la imputación son, en nuestro caso, aproximadamente el doble de los estimados.

5. Conclusiones

En primer lugar, hay que destacar que se describe un método completo de tratamiento de datos que parte de los registros mensuales de los distintos observatorios y concluye con la obtención de parámetros estadísticos de las series, incluyendo control de calidad de los datos, homogeneización de las series, imputación de valores ausentes y estimación de parámetros estadísticos.

La modelización de la varianza de las series originales en función de variables fisiográficas de los observatorios parece una técnica interesante para el control de calidad de los datos. Hay que destacar que el método permite identificar tanto datos individuales sospechosos de ser erróneos como series completas de deficiente calidad.

La dificultad de realizar procesos de homogeneización en series incompletas o el problema de estimar valores ausentes en series sin homogeneizar pueden quedar solventados mediante un método iterativo que realiza alternativamente la estimación de parámetros estadísticos y valores ausentes, y la homogeneización de las series.

No se conoce la razón por la cual tras el primer proceso de homogeneización de las series, el coeficiente de correlación entre los valores medios de las series climáticas y las variables fisiográficas sufre un descenso. Ello se produce casi todos los meses tanto en el caso de temperaturas máximas como mínimas diarias.

Cualquier estimación de valores ausentes necesita conocer los parámetros de la distribución estadística de las series y, al mismo tiempo, para estimar estos parámetros se necesita realizar hipótesis sobre el valor de los datos ausentes. Por tanto, el cálculo simultáneo de ambos mediante un proceso iterativo como el algoritmo ME regularizado parece una solución ideal. En posteriores estudios será conveniente determinar con cierta precisión la relación entre el error real y el estimado en la imputación de valores ausentes. La corrección del valor de este error en los distintos pasos de la iteración puede mejorar los resultados del método.

Referencias

Aguilar, E., J.M. López, M. Brunet, O. Saladié, J. Sigró y D. López, 1999. *Control de calidad y proceso de homogeneización de series térmicas catalanas*, en La climatología española en los umbrales del siglo XXI, Publicaciones de la A.E.C., 15-23.

Alexandersson, H., 1986. *A homogeneity test applied to precipitation data*, Int. J. Climatol. **6**, 661-675.

Alexandersson, H. y A. Moberg, 1997. *Homogeneization of swedish temperature data. Part I: Homogeneity test for linear trends*, Int. J. Climatol. **17**, 25-34.

Buck, S.F., 1960. *A method of estimation of missing values in multivariate data suitable for use with an electronic computer*, J. Roy. Stat. Soc. B **22**, 302-306.

Conrad, V. and L.W. Pollack, 1950. *Methods in climatology*, Harvard University Press, 459 pp.

Dempster, A.P., N.M. Laird y D.B. Rubin. *Maximum likelihood estimation from incomplete data via the EM algorithm*, J. Roy. Stat. Soc. B **39**, 1-38.

Golub, G.H., M. Heath y G. Wabba, 1979. *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, **21**, 215-223.

Karl, T.R. y C.N. Williams Jr., 1987. *An approach to adjusting climatological time series for discontinuous inhomogeneities*, J. Climate Appl. Meteorol. **26**, 1744-1763.

Little, J.A. y D.B. Rubin, *Statistical analysis with missing data*, Series in Probability and Mathematical Statistics, John Wiley and sons Inc., 278 pp.

Petterson, T.C., D.R. Easterling, T.R. Karl, P. Groisman, N. Nicholls, N. Plummer, S. Torok, I. Auer, R. Boehm, D. Gullett, L. Vincent, R. Heino, H. Tuomenvirta, O. Mestre, T. Szentimrey, J. Salinger, E.J. Forland, I. Hanssen-Bauer, H. Alexandersson, P. Jones y D. Parker, 1998. *Homogeneity adjustments of in situ atmospheric climate data: A review*, Int. J. Climatol. **18**, 1493-1517.

Ramos-Calzado, P., J. Gómez-Camacho, F. Pérez-Bernal y M.F. Pita-López, 2008. *A novel approach to precipitation series completion in climatological datasets: application to Andalusia*, Int. J. Climatol., en imprenta.

Rubin, D.B., 1976. *Inference and missing data*, Biometrika, **63**, 581-592.

Schneider T., 2001. *Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values*, J. Climate, **14**, 853-871.

Téllez, B., T. Cernocky y E. Terradellas, 2008. *Calculation of climatic referente values and their use for automatic outlier detection in meteorological datasets*, Advances in Science and Research, en imprenta.

Tijonov, A.N. y V.Y. Arsenin, 1977. *Solutions of ill-posed problems*, John Wiley and sons Inc., 258 pp.

Wilks, D.S., 1995. *Statistical Methods in the Atmospheric Science: an Introduction*, International Geophysics Series, vol. 95, Academic Press, 464 pp.