Reconstrucción de series temporales en ciencias ambientales

Manuel Benítez-Gilabert* y Miguel Álvarez-Cobelas

CSIC-Instituto de Recursos Naturales
Serrano 115 dpdo, E-28006-Madrid, España

Recibido 24 Septiembre 2008, Revisado 23 Octubre 2008, Aceptado 28 Octubre 2008

Missing data recovery in environmental time series

Abstract

As many environmental data are increasingly recorded on a long term basis, it is unfortunately frequent that they show missing data (MD). In addition to information losses, MD also prevent the use of time series analysis and present the researcher the dilemma of either apply sophisticated methods of analysis or attempt to fill those MD gaps in order to apply conventional methods. In any case, further statistical treatment usually needs complete time series and hence MD must be estimated. The main statistical methods to tackle this problem are briefly outlined here, and available software is reported as well. A case of time series reconstruction of Spanish rainfall and water quality to exemplify these methods is also described, using the maximum likelihood approach of the Expectation-Maximization-Bayesian (EMB) algorithm and the AMELIA-II free software.

Keywords: Missing data, conventional recovery methods, EMB algorithm, Amelia-II software.

Resumen

El registro cada vez más frecuente de datos ambientales a largo plazo tiene, a menudo como consecuencia no deseada, huecos en la serie temporal, debidos a las causas más variadas. La ausencia de datos, además de implicar pérdida de información, imposibilita la aplicación de los métodos clásicos de análisis, poniendo al investigador en el dilema de aplicar métodos avanzados de análisis o de tratar de rellenar los datos ausentes para aplicar métodos estadísticos clásicos. El tratamiento estadístico de dichas series exige el relleno de los datos ausentes. En este trabajo exponemos brevemente las principales metodologías estadísticas que se están aplicando al problema, inventariamos el software disponible (parte del cual es de distribución gratuita) y ofrecemos un ejemplo de aplicación de estas técnicas, basado en series españolas de meteorología y calidad de aguas. El método empleado es el del algoritmo EMB de máxima verosimilitud y el "software" usado es el AMELIA II, de carácter gratuito.

Palabras clave: Datos ausentes, métodos clásicos, métodos de máxima verosimilitud, imputación múltiple, algoritmo EMB, Amelia II.

Introducción

Uno de los problemas que con más frecuencia se presentan en el campo de las Ciencias Ambientales, sobre todo cuando se trabaja con series históricas, es el derivado de los datos ausentes. La ausencia de datos implica no solamente una pérdida de información, sino también en muchas ocasiones la imposibilidad de utilizar métodos estadísticos convencionales del tratamiento de los mismos y, por lo tanto, de la consecución de los objetivos previstos en el trabajo de investigación. Este problema se ha ido acentuando a medida que, con registro automatizado (como ocurre en meteorología u oceanografía) o medidos por el observador con gran esfuerzo personal, se van acumulando series largas de datos para su análisis ulterior, con vistas a responder a cuestiones relativas al cambio global, a la predicción meteorológica o a las tendencias y fluctuaciones en cualquier proceso ambiental de interés. No deja de ser paradójico, sin embargo, que la mayor parte de las aportaciones en el campo del tratamiento de los datos ausentes (DA) provengan del ámbito de las

* Autor para correspondencia
E-mail: ebvbg77@ccma.csic.es; Fax: +34-91-5640800.

Ciencias Sociales (Allison, 2000; Kim y Curry, 1977; Little y Rubin, 1987; Van Buuren y Oudshoorn, 1999), siendo escasas las referencias a sus aplicaciones en el ámbito de las Ciencias Ambientales (Benítez-Gilabert y Alvarez-Cobelas, 2008; Plaia y Bondi 2006).

En el presente trabajo nos proponemos realizar una exposición de los diferentes tipos de datos ausentes que se pueden presentar en el trabajo ambiental, las distintas aproximaciones que se han propuesto para abordar su problemática, y expondremos nuestras experiencias con éstas, así como el método usado por nosotros para una satisfactoria resolución del problema. También, describiremos un caso práctico de aplicación y validación del método propuesto. No entraremos, sin embargo, en la exposición de los métodos basados en los modelos de ecuaciones estructurales, ya que su consideración rebasaría las dimensiones de este artículo y merecería un tratamiento independiente, aparte de la necesidad de estar muy familiarizados con ese tipo de modelos estadísticos.

A pesar de que la necesidad de metodologías de reconstrucción de series temporales con DA es cada vez mayor, apenas hay bibliografía en español sobre el tema, si exceptuamos la de Reales (2007), de distribución muy restringida. Esta contribución, por tanto, pretende incentivar el uso de las metodologías disponibles, presentándolas en español y limitando en lo posible los aspectos estadísticos más áridos.

Por último, conviene tener en cuenta que las metodologías propuestas por los diversos autores están orientadas al tratamiento de series multivariantes, y que lo que se pretende con su tratamiento es obtener, a partir de la información existente, la estructura de datos subyacente que sea más verosímil, es decir, como veremos más adelante, se trata de "escoger como estimadores aquellos que, de ser ciertos, maximizarían la probabilidad de observar lo que en realidad ha sido observado" (Allison, 2001).

Tipos de datos ausentes

Son varias las tipologías en las que se suelen clasificar los DA, pero en todas ellas se hace referencia a los tres tipos fundamentales propuestos por Little y Rubin (1987): Pérdida completamente aleatoria o MCAR (*missing completely at random*), pérdida aleatoria o MAR (*missing at random*) y pérdida no aleatoria o MNAR (*missing not at random*). Estos autores matizan *a priori* dichas categorías añadiendo las de pérdida ignorable o NM (*negligible missing*) y pérdida no ignorable o NNM (*non-negligible missing*), en función de los mecanismos que hayan generado DA. Se dice que los datos faltan completamente al azar cuando su ausencia -o presencia- es independiente de cualquier característica de los casos y, por lo tanto, resulta impredecible; sin embargo, esto ocurre muy raramente (Coenders et al., 2005). Los datos faltan al azar cuando la ausencia o presencia de un dato es independiente del valor de la propia variable en la que falta, pero podría depender de las otras variables, sin mantener con ellas una relación estricta. Los datos faltan no al azar cuando la ausencia o presencia de un dato depende del valor de la propia variable; este mecanismo de pérdida nunca puede pasarse por alto.

Sin embargo, como apunta Reales (2007), aunque esta tipología resulte útil para la clasificación de los DA utilizando exclusivamente el mecanismo de generación de dichos datos, puede dar como resultado tanto una comprensión incompleta como fallos en la comunicación y la identificación de los mismos. Para este autor, el sistema de Little y Rubin se centra en la estructura de los DA en vez de hacerlo sobre su función. Por otra parte, los DA no pertenecen necesariamente a tres categorías definidas de forma rígida y mecánica, sino que a menudo reflejan los tres mecanismos al mismo tiempo (Schafer, 1997). Lógicamente, el ignorarlos o no supone tener en cuenta la importancia que la ausencia de una parte más o menos considerable de los datos tiene en el enfoque global de los mismos, en su mecanismo de generación y en la influencia que pueda tener el mecanismo de la pérdida de los datos sobre su tratamiento. El que su relevancia haga o no viable su estudio y tratamiento estadístico dependerá de la cantidad y distribución de éstos. Obviamente, una vez considerado que los DA son necesarios (es decir, no "*neglibibles*", valga el barbarismo), se hará necesario utilizar algún tipo de aproximación numérica que permita su aproximación y posterior análisis.

Hay que tener en cuenta que los DA casi siempre son necesarios porque los métodos estadísticos multivariantes usuales suponen que los datos están completos, de ahí, la necesidad de utilizar una aproximación numérica para analizarlos con métodos que requieren datos completos.
Por otra parte, el concepto de ignorable ("*negligible*") se refiere a que si los datos se ignoran, las inferencias no se alteran. Para ilustrar este comentario, supóngase que los datos numéricos más grandes tuvieran una probabilidad mayor de perderse; en este caso, si sólo se eliminaran los datos ausentes, las inferencias se modificarían porque se estarían eliminando sistemáticamente datos con valores grandes, con lo cual sólo se estarían analizando datos numéricos pequeños. Por esta razón, cuando la ausencia de un dato depende de su valor numérico, la ausencia nunca es ignorable.

Métodos clásicos utilizados para el manejo de los datos ausentes

Son varias las aproximaciones que se han efectuado para el tratamiento de los DA, basadas fundamentalmente en la eliminación de casos, en su sustitución por estimaciones indirectas o, posteriormente, basadas en la Máxima Verosimilitud. Salvo este último caso, los demás son definidos por Allison (2001) como clásicos o convencionales y, según este mismo autor, ninguno de ellos ha demostrado ser claramente más eficaz que la eliminación por pares (*pairwise*).
A continuación haremos una pequeña revisión de los métodos clásicos:

Eliminación por lista ó *Listwise*.- Consiste en la eliminación sistemática de todos los casos en los que falten datos en alguna de las variables del modelo. Con este método puede llegar a despreciarse una gran cantidad de información, sobre todo cuando se trabaje con modelos que contemplen un número elevado de variables (Coenders et al., 2005). Según estos mismos autores, es un método que conduce a contrastes poco potentes y a errores estándar elevados.

Eliminación por pares (método de los casos disponibles) ó *Pairwise*.- Este método permite utilizar todos los casos en los que las variables estén disponibles al calcular las covarianzas que posteriormente se emplearán en los análisis estadísticos. Según Allison (2001), este procedimiento puede ser utilizado en numerosos modelos lineales, incluyendo la regresión lineal, el análisis factorial y modelos más complejos de ecuaciones estructurales. Sin embargo, para Coenders et al. (2005), a pesar de ser un método más eficiente que el anterior por no descartar información útil, puede adolecer de problemas de cálculo. El motivo es que cada covarianza se calcula sobre un conjunto diferente de casos, pudiendo haber "incoherencias" en la matriz, lo que podría dar lugar a que ésta dejara de ser una matriz definida positiva, imposibilitando así los cálculos.
Aunque es usual el término, parece más correcto utilizar el término de matriz positiva definida, porque también los números de las matrices pueden ser positivos o negativos (aunque debido al incremento de complejidad en las matrices también haya matrices indefinidas).

Variable ficticia.- Se emplea principalmente en los problemas de regresión (Cohen y Cohen, 1985). Cuando en estos modelos una o más variables independientes D presenten DA, es posible crear una variable ficticia X, en la que los casos tomen el valor 1 cuando falta el dato, ó el valor 0 en los casos que contienen datos Al mismo tiempo, se crea otra variable X* tal que

$$X^* = X \text{ cuando no hay DA}$$
$$X^* = C \text{ cuando faltan los datos}$$

donde C es cualquier constante. A continuación, se efectúa la regresión de Y en X*, D y las demás variables que entren a formar parte del modelo. La principal ventaja de este método es que permite utilizar toda la información disponible, pero proporciona estimaciones sesgadas de los coeficientes de regresión (Jones, 1996).

Sustitución por el promedio.- En ella se reemplazan los DA por los valores medios de los

valores presentes. Este método sesga los valores de las varianzas y las covarianzas y, por lo tanto, no debiera utilizarse nunca (Graham et al., 1994). En el método de regresión se ajusta un modelo de regresión para cada variable que presenta DA y se sustituyen los DA por sus estimados mediante el modelo. Hay que tener en cuenta que este procedimiento tiende a sobreestimar las correlaciones entre variables y, por eso, debe utilizarse con cautela a la hora de interpretar los resultados de su posterior tratamiento estadístico.

Donación.- Llamado *hot deck* en la bibliografía anglosajona (Ford, 1983; Rao and Shao, 1992), en él los DA se sustituyen por los de otro caso en el cual los valores presentes sean similares.

Métodos basados en Máxima Verosimilitud

Máxima verosimilitud (MV).- Como indica Allison (2001), su principio básico reside en escoger como estimadores a aquellos valores que, de ser ciertos, maximizarían la probabilidad de observar lo que en realidad ha sido observado. Para conseguir esto, primero necesitamos enunciar una fórmula que exprese la probabilidad de los datos como una función de éstos y de los parámetros desconocidos que especifiquen la función de distribución de probabilidades del modelo a considerar. Cuando las observaciones son independientes, la verosimilitud total (probabilidad) para la muestra es igual al producto de todas las verosimilitudes individuales de cada uno de los casos, es decir

$$V\ (\theta\ ) = \prod_{i=1}^{n} f\left\langle y_i \middle| \theta \right\rangle$$

Donde $V(\theta)$ es la función de verosimilitud y $f\langle y_i|\theta\rangle$ es la función de densidad para datos continuos, la cual debe ser especificada para cada modelo en particular. Una vez calculada la función de verosimilitud, existe una gran variedad de técnicas que permiten maximizar el valor de V(θ). Cuando nos encontremos con DA y el mecanismo de ausencia de datos sea ignorable, podremos obtener la función de verosimilitud "multiplicando", agregando las verosimilitudes de los DA. En el caso de variables continuas, integraríamos la función de verosimilitud, obteniendo la expresión

$$V(\theta) = \prod_{i=1}^{n} f\langle x_i, y_i|\theta\rangle \prod_{j=1}^{r} g\langle y_j|\theta\rangle$$

En esta ecuación, si únicamente una variable presentase DA, el patrón de comportamiento sería necesariamente monotónico, con lo cual podríamos maximizar independientemente los dos términos de la ecuación anterior, debidamente transformada.

MV directa.- Al igual que en el caso anterior, en el método de máxima verosimilitud directa (Finkbeiner, 1979) también se supone que los datos faltan al azar, pero la estimación se hace en una sola etapa y los contrastes de hipótesis, así como los errores estándar, son correctos. Suele ser el método implementado en la mayoría de los paquetes estadísticos utilizados para el análisis y modelización de ecuaciones estructurales. El método es consistente y proporciona inferencias correctas, pero es más exigente con respecto a la normalidad de distribución de los datos (Enders, 2001).

Algoritmo EM.- Es uno de los métodos más ampliamente utilizados para maximizar el cálculo de la MV cuando hay DA (Dempster et al., 1977; McLachland y Krihnan, 1997). Sus siglas se deben a que su aplicación requiere dos pasos: *expectation* y *maximization*. Estos dos pasos son ejecutados en un proceso iterativo que busca la convergencia con el valor de Máxima Verosimilitud. La utilización de este algoritmo en el tratamiento de los DA supone la obtención de la matriz de covarianzas mediante MV y utilizando todos los datos. En principio, se considera que dicha matriz es consistente si los datos faltan al azar y están normalmente distribuidos. No obstante, la mayoría de los métodos que utilizan esta aproximación algorítmica mantienen su robustez en caso de no normalidad de los datos (Graham y Schafer, 1999). Y también será consistente la estimación de los parámetros del modelo, aunque sus errores estándar

sean desconocidos.

Imputación múltiple.- Aunque los métodos basados en MV con imputación simple (un solo valor por cada dato ausente) representen un gran avance con respecto a los métodos clásicos, también tienen sus limitaciones, sobre todo cuando se trabaja con modelos no lineales. Por suerte, existe un método alternativo –la Imputación Múltiple (IM)– que proporciona las mismas ventajas que los métodos de MV con imputación simple y elimina sus limitaciones. La IM, cuando se utiliza correctamente y los datos faltan al azar, proporciona estimaciones que son consistentes, asintóticamente eficientes y asintóticamente normales y puede ser virtualmente usada con cualquier tipo de datos y cualquier tipo de modelo.
No obstante, el método de IM tiene sus desventajas. Su implementación puede ser engorrosa, lo cual puede inducir a errores si no se mecaniza su utilización utilizando el *software* apropiado. Pero quizá lo más relevante es que la IM proporciona resultados ligeramente distintos cada vez que es utilizada. Esto puede dar lugar a que diferentes investigadores obtengan resultados ligeramente distintos al utilizar el mismo método. Sin embargo, no debe olvidarse que el objetivo de los diferentes procedimientos de tratamiento de los DA es obtener el valor más probable de los mismos, sin aspirar a obtener el valor exacto del dato ausente, puesto que ello es imposible.
Conviene resaltar que, en los algoritmos de imputación múltiple en dos etapas, siempre y cuando el modelo de imputación contenga al menos tanta información como las variables comprendidas en el modelo analizado, no se producirán sesgos en los estimadores, aunque el modelo no sea normal o lineal (Meng, 1994).

Algoritmo EMB.- Aparece como una alternativa al algoritmo EM. A la potencia de éste, permite unirle una mayor flexibilidad de computación y la utilización de métodos de *bootstrapping* en el cálculo de los estimadores de los modelos de imputación múltiple (King et al., 2001). Estos autores utilizan un método mixto en el que se combinan la aplicación del análisis *Bayesiano* en el proceso de imputación y un algoritmo de *bootstrapping* en el proceso de determinación de los estimadores, a partir de la función de densidad de probabilidades *a posteriori*. La utilización de un algoritmo de *bootstrapping* tiene la ventaja de proporcionar convergencias asintóticas de orden menor que las obtenidas por la aproximación paramétrica que se obtiene mediante el algoritmo EM.

Software más utilizado y sus características

Debido a lo tedioso de los cálculos necesarios para efectuar el tratamiento de los DA y dadas las características iterativas de su proceso de cálculo, es lógico que se hayan implementado diversos procedimientos informáticos para realizar esta labor. Casi todos los paquetes estadísticos poseen procedimientos específicos para la eliminación de casos, en cualquiera de sus modalidades. Un número reducido de paquetes estadísticos posee subrutinas para la aplicación de modelos de MV. Y también existe *software* específico para el tratamiento de los DA.

***Software* no específico:** Algunos paquetes estadísticos han incluido en sus rutinas procedimientos para el tratamiento de los DA, basándose en los métodos de máxima verosimilitud. Es el caso de BMDP (Program 5V) y de SAS (PROC MIXED), pero no son específicos para este propósito y su capacidad y flexibilidad resultan limitadas.

***Software* específico:**
***Solas*.-** Comercializado por Statistical Solutions (http://statsolusa.com) y desarrollado en colaboración con D.B. Rubin, coautor de una monografía sobre el tema (Little y Rubin, 1987), se basa en su tipología para efectuar imputaciones simples y múltiples utilizando el algoritmo EM.

***SPSS Missing Value Analysis*.-** Como el programa anterior, es un programa comercial basado en la clasificación de Little y Rubin con las mismas características que aquél y utilizando el mismo algoritmo. Se puede encontrar en http://www.spss.com/stores/1/SPSS_Missing_Value_Analysis_tr_P7978C2.cfm.

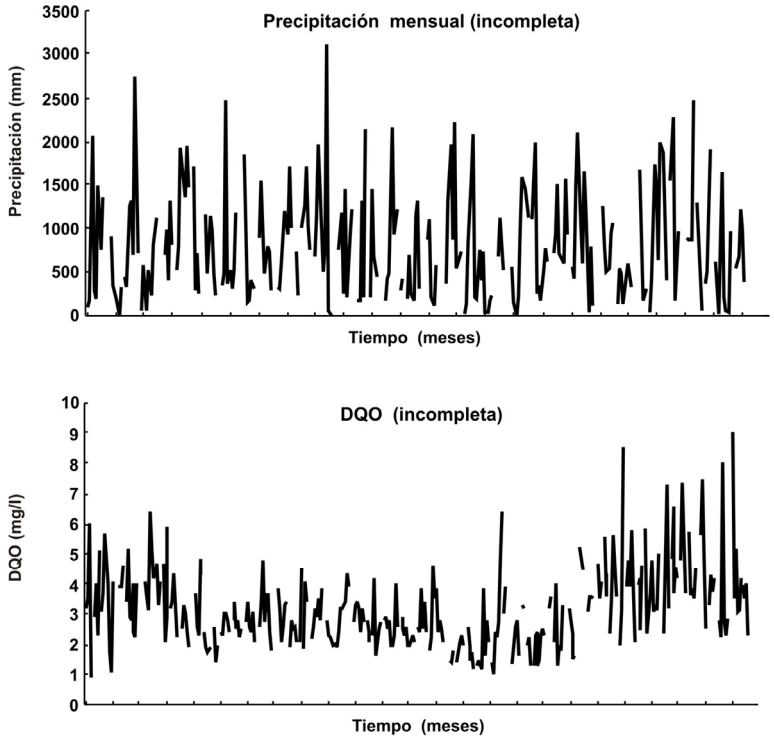

Figura 1. Series temporales de pluviosidad y Demanda Química de Oxígeno con datos ausentes, obtenidas eliminando aleatoriamente un tercio de los datos intermedios. Las series originales han sido registradas por la Agencia Nacional de Meteorología de España y el Ministerio español de Medio Ambiente, Medio Rural y Marino en San Esteban de Gormaz, Cuenca Hidrográfica del Duero (41º 34' N, 0º 28' E).

***Norm*.-** Gratuito, sirve para procedimientos de imputación múltiple de datos continuos multivariantes con distribución normal. Está en http://www.stat.psu.edu/~jls/misoftwa.html (Schafer, 1997).

***CAT*.-** Gratuito, es para imputación múltiple de datos categóricos multivariantes con distribución loglineal (http://www.stat.psu.edu/~jls/misoftwa.html; Schafer, 1997).

***Amelia II*.-** Gratuito y desarrollado por Homaker et al. (2006; http://gking.harvard.edu/amelia/), está basado en el algoritmo EMB. Es un programa de fácil utilización, bien documentado y con capacidades específicas para el tratamiento de datos multivariantes, series históricas, datos longitudinales, etc. Por su sencillez y potencia es, para nosotros, el más recomendable.

Caso práctico

Para ilustrar las posibilidades de utilización del tratamiento de datos ausentes mediante la metodología de imputación múltiple utilizando el algoritmo EMB, expondremos el procedimiento y los resultados obtenidos con el paquete estadístico Amelia II. Los datos utilizados en este análisis corresponden a datos meteorológicos y de calidad de aguas fluviales, recogidos por la Agencia Española de Meteorología y el Ministerio español

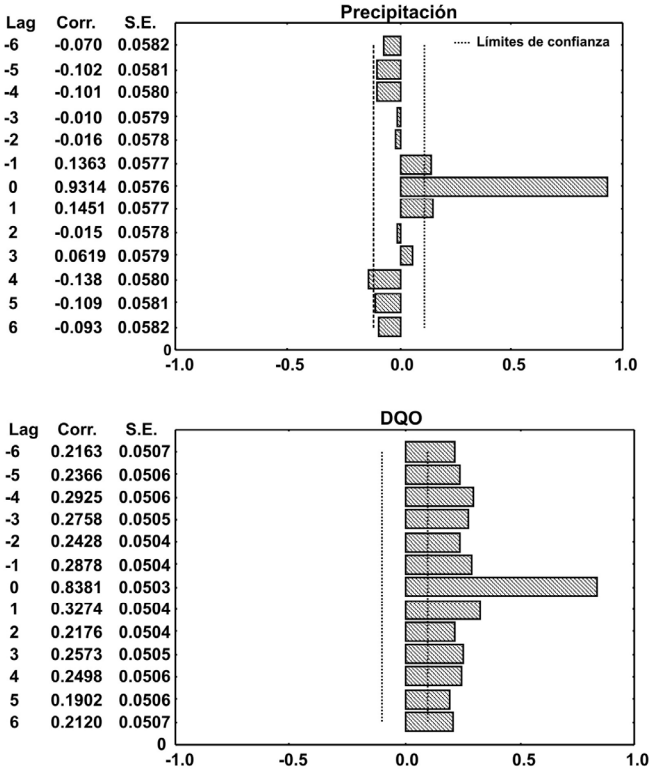

Figura 2. Correlaciones cruzadas entre las series de pluviosidad y Demanda Química de Oxígeno con datos ausentes y la misma serie reconstruida.

de Medio Ambiente, Medio Rural y Marino entre los años 1973 y 2005 con periodicidad mensual. Hemos trabajado con las variables de pluviosidad y de Demanda Química de Oxígeno (DQO).

Para obtener dos series incompletas que nos sirvieran de ejemplo, se procedió a eliminar aleatoriamente algunos de los datos (aproximadamente un 30% de los mismos; Fig. 1). Este procedimiento nos proporcionó dos series que pueden considerarse de tipo MAR, de acuerdo con la clasificación de Little y Rubin (1987).

Para el tratamiento de las series incompletas se utilizó el paquete estadístico Amelia II, tanto por las ventajas que presenta el algoritmo EMB frente al EM como por su facilidad de obtención (es gratuito) y manejo. Se usó la opción de series temporales sin la utilización de "cortes transversales" (*cross-sectional data*). Amelia II permite la utilización de este tipo de datos en algunos supuestos, utilizando la aproximación polinómica

$$f(t) = \beta_0 + \beta_1 * t + \beta_2 * t^2 + ...\beta_k * t^k$$

lo cual, si se incluye un orden del término de tiempo (k) lo suficientemente elevado, permite la estimación del patrón de comportamiento de $f(t)$ y esos patrones de comportamiento quedan acotados por los cortes transversales. Dichos cortes transversales son los valores sincrónicos de todas las variables, observadas a intervalos definidos. Amelia II no efectúa ningún suavizado (*smoothing*)

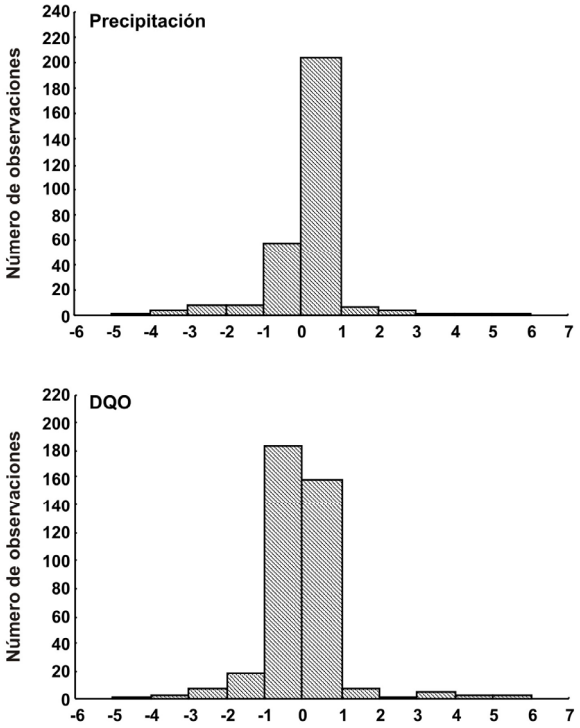

Figura 3. Distribuciones estadísticas de los residuos de las correlaciones cruzadas de las series de pluviosidad y Demanda Química de Oxígeno con datos ausentes y la misma serie reconstruida.

de las series originales y, para imputar los datos ausentes, sólo utiliza la aproximación funcional y una componente de incertidumbre derivada de las demás variables. Los resultados obtenidos tras la imputación múltiple pueden ser tratados posteriormente con cualquier paquete estadístico.

El programa Amelia II puede descargarse libremente de la página "web" http://gking.harvard.edu/amelia/. Los datos que utiliza deben estar en formato "csv", es decir, en formato de texto delimitado por comas y exportado directamente desde una hoja de cálculo, los cuales son importados directamente por Amelia II. Las especificaciones necesarias para efectuar los análisis se proporcionan a través de menús desplegables, aunque permita la posibilidad de trabajar con líneas de comandos. Los resultados obtenidos en la imputación de los DA (los datos ya existentes no los modifica), también salen en formato "csv", lo cual facilita exportarlos a una hoja de cálculo.

También hay que tener en cuenta la configuración regional del sistema en el ordenador, puesto que Amelia II trabaja con delimitadores de serie (",") y decimales ("."). Si no se respeta esta configuración, el programa puede dar errores. Teniendo presente ese detalle, el manejo del programa es muy sencillo y cuenta con un manual muy explícito que soluciona cualquier duda que se presente sobre su utilización.

Para comprobar la bondad de la reconstrucción de las series temporales incompletas se ha procedido a realizar un análisis de correlación cruzada entre ambas series, así como a un análisis de los residuos. En el caso de las correlaciones cruzadas puede observarse el alto grado de correlación en el

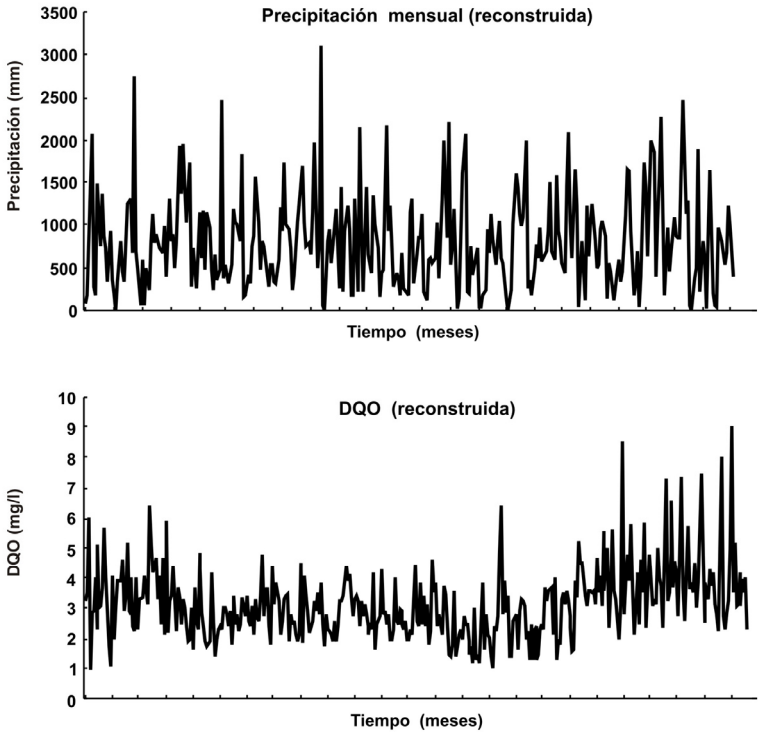

Figura 4. Series reconstruidas de pluviosidad y Demanda Química de Oxígeno, usando la metodología de máxima verosimilitud y el programa AMELIA II. Compárese con la Fig. 1.

desfase 0, es decir, el valor correspondiente a una correlación de Pearson convencional, disminuyendo los valores de correlación de forma significativa para todos los demás retardos, tanto en el caso de la precipitación como en el de la DQO (Fig. 2). En este último caso se observa la periodicidad (estacionalidad) de las series. Esto es lógico, puesto que los datos originales no fueron sometidos a ningún tipo de tratamiento y el análisis posterior para su imputación tampoco efectúa ninguna.

El análisis de los residuos confirma la bondad de las estimaciones en la reconstrucción de las series. Para la precipitación (Fig. 3, residuos normalizados), con una media serial de 787,70 mm se obtiene una población de residuos con media -14,5 y desviación típica 21,2 mm; en el caso de la DQO (Fig. 3), con un valor medio serial de 3,15 mg O_2/L, se obtiene una población de residuos con un valor medio de -0,04 y una desviación típica de 0,07 mg O_2/L. Si consideramos los residuos estándar, la mayor proporción de éstos está comprendida entre los valores de -1 y +1 unidades. Lógicamente, su distribución no es normal, aproximándose más a una distribución de tipo "gamma desfasada", dado que casi todos los valores se encuentran en un reducido intervalo.

Finalmente, la Fig. 4 muestra las dos series reconstruidas por el procedimiento AMELIA II mediante la aplicación del algoritmo EMB. Comparándola con la Fig. 1, puede apreciarse la gran similitud de las estructuras reconstruidas con las pautas temporales de las series originales.

Por lo tanto, consideramos estos resultados lo suficientemente satisfactorios como para poder trabajar con las series reconstruidas en posteriores análisis y, así, nos atrevemos a recomendar este

método como una alternativa válida a la hora de tratar con series históricas incompletas que, de otra forma, no podrían ser útiles para nuestro trabajo de investigación. La sencillez del método y la consistencia de los resultados ofrecen la posibilidad de manejar fuentes de información que -por numerosas causas- adolecen de la calidad de datos necesaria para su adecuado tratamiento e interpretación. De todos modos, aunque los resultados obtenidos resultan muy aceptables, convienen recalcar que no tienen por qué ser mejores con otros métodos de reconstrucción no probados aquí. Una comprobación exhaustiva de métodos con las mismas series de datos de partida, aunque deseable, quedaba fuera de los objetivos de este escrito introductorio al tema de los datos ausentes.

Agradecimientos

Los datos empleados aquí fueron facilitados por la Agencia Española de Meteorología y el Ministerio de Medio Ambiente, Medio Rural y Marino. Este trabajo se ha financiado con el Proyecto CGL2006-02346/HID y un convenio del CSIC con la Confederación Hidrográfica del Guadiana. Un revisor anónimo nos hizo valiosas sugerencias a la primera versión del texto.

Bibliografía

Allison, P. D. 2000. Multiple imputation for missing data. Sociological Methods and Research 28: 301-309.

Allison, P. D. 2001. Missing Data. Sage University, Paper 136, London.

Benítez-Gilabert, M. y Alvarez-Cobelas, M. 2008. Stream water quality is influenced by climatic change and teleconnections in Southwestern Europe. Climatic Change (en prensa).

Coenders, G., Batista-Folget, J. M. y Willem, E. S. 2005. Temas avanzados en modelos de ecuaciones estructurales. La Muralla S.A., Madrid.

Cohen, J. y Cohen, P. 1985. Applied multiple Regression and Correlation Analysis for the behavioural Sciences. 2nd edition. Eribaum Press. Hillsdale, New Jersey.

Dempster, A. B., Laird, N. M. y Rubin, D. B. 1997. Maximum likelihood estimation from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B: 39: 1-38.

Enders, C. K. 2001. The impact of non-normality on full information maximum likelihood estimation for structural equation models with missing data. Psychological Methods 6: 352-370.

Finkbeiner, C. 1979. Estimation for the multiple factor model when data are missing. Psychometrika 44: 409-420.

Ford, B. L. 1983. An overview of Hot-Deck procedures. In Madow, W. G., Olkin, I. y Rubin, D. B. (Eds.), Incomplete Data in multiple Surveys, pp 185-207. Academic Press, New York.

Graham, J. W. y Schafer, J. L. 1999. On the performance of multiple imputation for multivariate data with small sample size. In Hoyle, R. H. (Ed.), Statistical Strategies for small Sample Research, pp 1-29. Sage Publications, London.

Graham, J. W, Hofer, S. M. y Piccinin, M.,1994. Analysis with missing data in drug prevention research. In Collins, I. M. y Seitz, L. (Eds.), Advances in Data Analysis for Prevention Intervention Research. National Institute on Drugs Abuse. Monography series 142. Washington DC.

Homaker, J., King, G. y Blackwell, M. 2006. Amelia II: A program for missing data. (http://gking.harvard.edu/amelia/).

Jones, M. P. 1996. Indicator and stratification methods for missing explanatory variables in multiple linear regressions. Journal of the American Statistical Association 91: 222-230.

Kim, J.O. y Curry, J. 1977. The treatment of missing data in multivariate analysis. Sociological Methods and Research 6: 215-240.

King, G. J., Honaker, A. J. y Kenneth, S. 2001. Analyzing incomplete Political Science Data. An alternative Algorithm for multiple imputation. Political Science Review 95: 49-69. (http://gking.harvard,edu/files/abs/evil-abs.shtml).

Little, R. J. A. y Rubin, D. B. 1987. Statistical Analysis with missing Data. Wiley, New York.

McLachlan, G. J. y Krishnan, T. 1997. The EM Algorithm and Extension. Wiley, New York.

Meng, X. 1994. Multiple imputation inference with uncongenial sources of input. Statistical Science 9: 538-573.

Plaia, A. y Bondi, A. L. 2006. Single imputation method of missing values in environmental pollution data sets. Atmospheric Environment 40: 7316–7330.

Rao, J. N. K. y Shao, J. 1992. Jacknife variance estimation with survey data under Hot-Deck imputation. Biometrika 70: 811-822.

Reales, J. M. 2007. Valores perdidos en el análisis de datos en Psicología. Apuntes en preparación. UNED, Facultad de Psicología. Dpto. de Metodología de las Ciencias del Comportamiento. Madrid.

Schafer, J. L. 1997. Analysis of incomplete multivariate Data. Chapman and Hall, London.

Van Buuren, S. y Oudshoorn, K. 1999. Flexible multiple imputation by MICE. Report TNO-PG 99054, TNO Prevention and Health, Leiden. (http://www.multiple-imputation.com).