Enlace de encuestas: una propuesta metodológica y aplicación a la Encuesta de Presupuestos de Tiempo

María Jesús BARCENA* y Fernando TUSELL*

Resumen

Tratamos el problema de completar dos ficheros con registros conteniendo un subconjunto común de variables. La técnica investigada utiliza árboles de regresión y/o clasificación. Se propone y estudia una extensión para variables respuesta multivariantes, ilustrando su empleo sobre la Encuesta de Presupuestos de Tiempo (EPT-93).

Palabras clave: enlace de ficheros; inserción de encuestas; imputación; árboles de regresión y clasificación.

Calsificación AMS: 62G05, 62D05.

1 Introducción

Nuestro punto de partida son dos ficheros, A y B con un total de $N = N_A + N_B$ observaciones. Cada fichero tiene la siguiente estructura:

Fichero A

$X_1, \ldots, X_p$	$Y_1, \ldots, Y_q$	desconocido

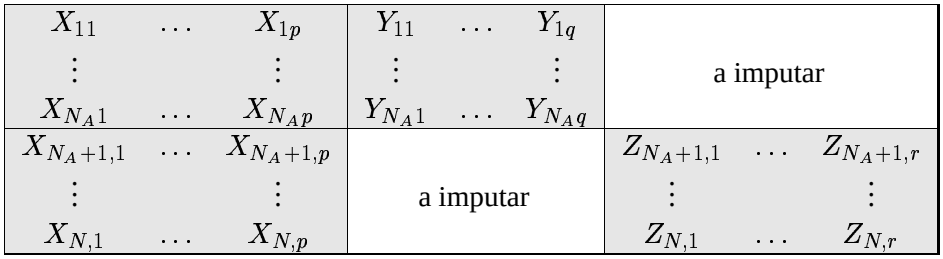
Fichero B

$X_1, \ldots, X_p$	desconocido	$Z_1, \ldots, Z_r$

Estos ficheros contienen, en nuestra aplicación, datos de dos encuestas diferentes con un grupo de variables en común, $X_1, \ldots, X_p$, cuyos valores son conocidos para los N individuos. Cada encuesta tiene además otras variables específicas que sólo son medidas para los individuos de esa encuesta: $Y_1, \ldots, Y_q$ y $Z_1, \ldots, Z_r$ respectivamente, en nuestra notación.

*Departamento de Estadística y Econometría. Facultad de CC.EE. y Empresariales, Avda. del Lehendakari Aguirre, 83, 48015 BILBAO. E-mail: etptupaf@bs.ehu.es. Este trabajo ha sido realizado como parte integrante de la tesis doctoral del primer autor. Agradecemos la financiación recibida del MEC y de la UPV/EHU (proyecto PB95-0346), los comentarios de asistentes a un seminario del Departamento, particularmente Vicente Nuñez, Eva Ferreira y Karmele Fernández, y de un evaluador. Los errores, carencias o inexactitudes que subsistan son nuestros.

Figura 1: Matriz de observaciones disponibles en un problema de enlace de encuestas



En este trabajo estudiamos cómo completar los ficheros A y B, utilizando los valores de las variables comunes X: tratamos de imputar las áreas sin sombreado en la Figura 1. El objetivo es permitir un análisis más rico de la información al poner en relación las variables específicas de ambas encuestas, es decir, aproximarnos al caso de tener datos completos de (X, Y, Z) para los N individuos. Sin embargo, hay una limitación importante: sólo podemos aspirar a reconstruir la parte de la relación entre Y y Z que puede transmitirse a través de X ya que, con los datos disponibles, es imposible conocer la relación entre Y y Z cuando X es constante. Por tanto, los resultados obtenidos con datos completados se aproximarán más a los que se obtendrían con datos completos, es decir, el enlace será mejor:

1. Cuanto menor sea la relación entre Y y Z cuando X es constante.

2. Si tenemos información adicional sobre la relación entre Y y Z cuando X es constante, y realizamos el enlace teniendo en cuenta esa información.

En lo que sigue veremos una breve descripción de algunas técnicas de enlace de encuestas, desarrollaremos una nueva que emplea árboles de regresión y/o clasificación y aplicaremos ésta última a los datos de la Encuesta de Presupuestos de Tiempo, en lo sucesivo designada EPT-93.

La EPT-93 fue realizada por el EUSTAT (Instituto Vasco de Estadística) en otoño de 1992 y primavera de 1993, entrevistando a 5040 individuos. Para cada individuo se recogen sus características personales y el tiempo en minutos que dedica a diferentes actividades (dormir, higiene, trabajo, etc.). La información es recogida mediante un diario que cada encuestado rellena en un sólo día previamente establecido. Como se indica en EUSTAT (1997), pág XII, el diario semanal es una alternativa mucho mejor, pero

> " . . .su calidad disminuye conforme avanza la semana, los diarios incompletos se multiplican y, además, es imposible de realizar sin un incentivo económico alto."

Por ello se optó por utilizar cuestionarios de un sólo día, pero distribuyéndolos aleatoriamente en cuatro bloques semanales (de lunes a jueves, viernes, sábados y domingos). Así, para cada individuo se conocen sus características personales, el tiempo dedicado a las actividades que realiza en un día dado y en cuál de esos bloques semanales ha respondido a la encuesta.

Es necesario analizar por separado las respuestas en días laborables de las obtenidas en días festivos, ya que la utilización de tiempo es diferente en días de distinto tipo. Por ello, hemos separado los datos de la EPT-93 en dos ficheros: `trabajo.dat` (con las observaciones para los 2521 individuos que responden un día de lunes a viernes) y `fiesta.dat` (con los datos para los 2519 que responden un día sábado o domingo).

El objetivo de este trabajo es presentar la metodología empleada para enlazar la información de `trabajo.dat` y `fiesta.dat`, e ilustrar su comportamiento. De esta forma, tratamos de aproximarnos a los resultados que se habrían obtenido con un cuestionario semanal. Esto puede verse como un caso particular del problema general descrito más arriba, considerando `trabajo.dat` y `fiesta.dat` como ficheros A y B.

El resto de este artículo se organiza de la siguiente manera: la Sección 2 contiene una breve descripción de algunos de los métodos empleados para enlazar encuestas. En la Sección 3 proponemos un método de enlace de encuestas que emplea árboles de regresión y/o clasificación; examinaremos sus ventajas e inconvenientes frente a los métodos señalados en la Sección 2 en el caso más simple, y los problemas que plantea su extensión multivariante. En la Sección 4 mostraremos un procedimiento para solventar dichos problemas y en la Sección 5 aplicaremos el algoritmo resultante a datos reales (de la EPT-93).

2 Técnicas de enlace de encuestas

Una idea que surge de modo natural es utilizar los valores disponibles de (X, Y) y (X, Z) que aparecen en la Figura 1 para regresar las variables Y y Z sobre las X. Podríamos luego imputar los valores desconocidos por los valores ajustados $\hat{Y}$ o $\hat{Z}$, mediante las regresiones obtenidas. Esta idea se remonta al menos a 1960, (véase Buck (1960)), y posteriormente ha habido numerosas contribuciones en esta línea. También, si puede especificarse la función de verosimilitud, puede emplearse el algoritmo EM en la estimación de los valores perdidos en la Figura 1; véase Dempster et al. (1976).

En lugar de emplear un modelo paramétrico para el ajuste, pueden emplearse métodos de reemplazamiento, sustituyendo los valores de las variables desconocidas para un individuo mediante los valores que esas variables toman en otro individuo "próximo" a él, según una métrica predefinida en el espacio de las variables comunes X. Nos encontramos entonces con diferentes versiones de la idea de vecinos más próximos, entendiendo la proximidad en el espacio $\mathcal{X}$ de las variables X y basándonos en alguna noción útil de distancia. A este método se le denomina

"hot-deck imputation"; ver, por ejemplo, Little y Rubin (1987), p. 60.

Se han propuesto también métodos para desvelar la relación entre Y y Z en casos como el que nos ocupa. En Aluja y Rius (1994), Aluja et al. (1995), se estudia cómo proyectar la información de una encuesta en los planos factoriales obtenidos del análisis de la otra (la encuesta de referencia), mediante técnicas de análisis factorial (análisis de correspondencias múltiples, análisis de componentes principales, etc.); ver también Bárcena et al. (1997) sobre el mismo tema. Métodos similares y estrechamente relacionados son el método de componentes principales de Dear y el método de descomposición en valores singulares de Krzanowski: en Bello (1993) se da una breve descripción de cada uno y un estudio de simulación para comparar su aplicación.

Las redes neuronales tienen la ventaja de su extrema flexibilidad para modelizar las dependencias más diversas, y no requerir la especificación de un modelo *a priori*. Ejemplos recientes del uso de redes neuronales en problemas de imputación son Villagarcía y Muñoz (1997) y Nordbotten (1996).

La técnica conocida como imputación múltiple es un complemento interesante a lo presentado hasta ahora. En Little y Rubin (1987) se presenta convincentemente su fundamento y sus beneficios; ver también Rubin (1986).

La imputación múltiple permite tener una idea de la variabilidad de las estimaciones entre imputaciones. Tenemos así no sólo un resultado, sino también una idea de la imprecisión introducida en el mismo por el proceso de imputación.

3 Enlace de encuestas mediante árboles de regresión y/o clasificación

Proponemos el uso de árboles de regresión y/o clasificación como método de imputación. Desde nuestro punto de vista, utilizar árboles soluciona varios problemas: proporciona un tratamiento unificado tanto para variables cualitativas como cuantitativas, de su empleo se derivan resultados que pueden utilizarse para valorar el enlace y permite realizar fácilmente imputación múltiple. Además los árboles tienen en ésta como en otras aplicaciones ventajas bien conocidas: flexibilidad, escasez de supuestos, resistencia a *outliers*, etc. El trabajo seminal Breiman et al. (1984) describe bien estas ventajas.

3.1 Grupos de variables específicas, Y y Z, univariantes

Estudiamos aquí el caso más sencillo en que tenemos p variables comunes X y $q = r = 1$, es decir, una única variable específica a imputar en cada encuesta (nos referimos a la Figura 1). Relegamos el caso $q > 1$, $r > 1$ a la Sección 3.2, que muestra los problemas que plantea la imputación multivariante. Un método para llevarla a cabo se presenta en la Sección 4.

Denotaremos $\mathcal{X}$ el espacio formado por todos los posibles valores de X. La idea es construir dos particiones de $\mathcal{X}$, tales que en cada clase de la primera (se-

gunda) partición tengamos valores lo más parecidos posible de Y (respectivamente, Z). Dado que no imponemos restricciones sobre el tipo de variables, la metodología CART de Breiman et al. (1984) es adecuada. Remitimos al lector a dicha fuente para una presentación de dicha metodología, y especialmente de las técnicas para construir y podar árboles.

Nuestra primera aproximación viene recogida en el Algoritmo 1. Nótese que

---

Algoritmo 1 – Imputación univariante por árbol.

---

1. Construir un árbol $\mathcal{Y}_X$ "regresando" Y sobre las X, usando validación cruzada y las observaciones $i = 1, \ldots, N_A$. Denotamos las hojas por $\mathcal{Y}_1, \ldots, \mathcal{Y}_a$. Forman la partición $\mathcal{Y}$.

2. Construir un árbol $\mathcal{Z}_X$ "regresando" Z sobre las X, usando validación cruzada y las observaciones $i = N_A + 1, \ldots, N$. Denotamos las hojas por $\mathcal{Z}_1, \ldots, \mathcal{Z}_b$. Forman la partición $\mathcal{Z}$.

3. Para imputar un valor de Z correspondiente a un caso $i = 1, \ldots, N_A$, lo dejamos caer por el árbol $\mathcal{Z}_X$. Si cae en la hoja $\mathcal{Z}_{\delta(i)}$, imputamos en función de los casos de la muestra de entrenamiento que han caído en dicha hoja.

   Procedemos del mismo modo para imputar valores de Y para los casos en que falta dicha variable $(i = N_A + 1, \ldots, N)$.

---

el método descrito en el mismo se presta bien a la imputación múltiple, ya que cada individuo cae en un nodo terminal que normalmente contiene casos con diferentes valores de Z (o de Y). Ello permite tomar uno o varios al azar. Podemos también imputar mediante la media, mediana, etc. de los valores que encontremos, o siguiendo la regla de la mayoría en el caso de que la variable a imputar sea cualitativa.

El método propuesto. en esencia, puede verse como un método de regresión en el que se emplea un árbol en lugar de un modelo paramétrico; o como un método "hot deck", en que la selección del individuo donante (o una pluralidad de ellos, si se recurre a imputación múltiple) se hace con ayuda de un árbol.

A la vista del Algoritmo 1 cabe preguntarse: ¿Podemos mejorarlo? ¿Podemos aprovechar para predecir, por ejemplo, Z, que conocemos no sólo los valores de X sino también los de Y? La respuesta es: no. Con la muestra disponible (ver Figura 1), no hay información en la Y que pueda ayudarnos a predecir Z salvo la ya contenida en las X (ver Bárcena y Tusell (1997) a este respecto). Sólo si tuviésemos información adicional sobre la relación entre Y y Z dada X podríamos aprovechar que conocemos los valores de Y además de los de X; esta información adicional podría obtenerse mediante una muestra de observaciones completas de los tres grupos de variables (X, Y, Z) o mediante información extra-muestral (por ejemplo, conocimiento experto). El uso de árboles puede orientar sobre cómo tomar una muestra adicional de observaciones completas de (X, Y, Z), si ello fuera

posible: Clases o nodos terminales con una variabilidad intra-clase alta indican estratos de población donde conviene tomar una muestra completa (X, Y, Z).

Como conclusión, creemos que el uso de árboles a la hora de enlazar encuestas tiene las siguientes ventajas:

1. Su uso es sencillo y los resultados fáciles de interpretar.
2. No es necesario que las variables sigan una determinada distribución ni que la relación entre la variable a explicar y los "regresores" tenga una forma funcional determinada.
3. Cada caso a imputar cae en un nodo terminal en el que normalmente hay más de un caso de la encuesta donante. Esto permite realizar de forma sencilla imputación múltiple.
4. Las hojas o clases con elevada dispersión sugieren estratos de población donde sería más conveniente tomar una muestra adicional de observaciones completas de (X, Y, Z), si existiera la posibilidad de hacerlo.
5. El empleo de variables suplentes *(surrogate splitting)* permite emplear el árbol incluso con valores perdidos entre las X.

3.2 Generalización: Y y Z multivariantes

Al tratar de generalizar el método a situaciones en que Y y Z son multivariantes (q>1 y r>1) nos encontramos con un problema: la metodología existente sobre árboles de regresión y/o clasificación sólo contempla variables respuesta univariantes. La idea más inmediata sería construir árboles para variables respuesta multivariantes que proporcionasen una partición de $\mathcal{X}$ en cada una de cuyas regiones estuviesen los individuos con valores "semejantes" del vector respuesta. No hay una manera única de definir semejanza en este contexto. Como alternativa, pensamos en varias opciones que se describen a continuación. Por brevedad, nos centramos en el caso de imputar valores de las variables Y; para imputar Z se procede de modo análogo.

Opción 1: Imputar cada variable $Y_1, \ldots, Y_q$ por separado, construyendo un árbol para cada una de ellas sobre las variables comunes X. Se emplean para ello los datos completos en (X, Y) en la encuesta A.

Esto supone convertir un problema multivariante en varios univariantes, prescindiendo de las relaciones entre las variables Y. Por tanto, esta opción parece apropiada si las variables a imputar están próximas a la independencia. Si éste no es el caso y entre las variables a imputar existe una estructura de correlación que interesa preservar, nos planteamos las alternativas siguientes.

Opción 2: Efectuar un cambio de variables, reemplazando $Y_1, \ldots, Y_q$ por sus componentes principales $U_1, \ldots, U_q$. Se imputan entonces las componentes principales y, hecho esto, se deshace el cambio, convirtiendo los valores imputados de las componentes a valores imputados de las variables originales.

Esta segunda opción tiene la ventaja (respecto a la Opción 1) de que parte de la relación entre las variables a imputar (aquélla de que dan cuenta las componentes principales) queda recogida. Tiene el inconveniente de que los árboles construidos son más difíciles de interpretar.

Al imputar cada variable o componente por separado, los valores imputados pueden pertenecer a individuos diferentes y, por tanto, no hay garantía de que esos valores imputados sean coherentes entre sí, en el sentido de verificar una relación lógica o matemática que necesariamente debería darse entre los valores imputados. Por ejemplo, en el caso de la encuesta EPT-93, si imputásemos el tiempo dedicado a cada actividad por separado nada impediría que la suma de tiempos fuese diferente de 24 horas (algo que necesariamente debe ocurrir).

Opción 3. Imputar a cada caso i de una encuesta todo el vector de variables desconocidas de una vez, tomando los valores del vector homólogo en un individuo "semejante" de la otra encuesta. Este modo de operar parece ser comúnmente aceptado (véase Lejeune (1995), pág. 140 y Lebart y Lejeune (1995) a este respecto).

Así aseguramos que los valores imputados son coherentes; los toma un individuo de esa población (estamos suponiendo que las encuestas se han realizado sobre la misma población). Por ello, consideramos esta opción la más adecuada. La Sección 4 describe un método para implementarla.

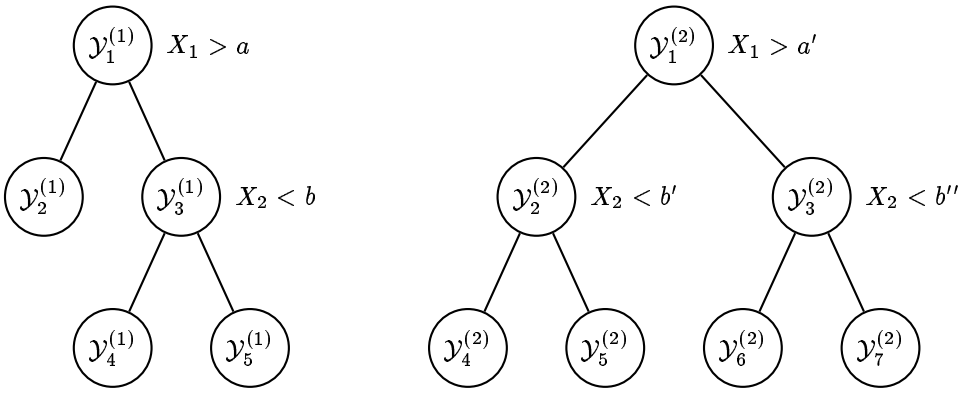
4 De la imputación simple a la conjunta: un nuevo método

Para imputar simultáneamente $\mathbf{Y_i} = (Y_{i1}, \ldots, Y_{iq})$ utilizamos los resultados de los árboles $\mathcal{Y}_X^{(j)}$ construidos para cada una de las variables Y_j, $j = 1, \ldots, q$, del modo que se explica a continuación y se formaliza en el Algoritmo 2. Como anteriormente, discutimos sólo el caso de imputar las variables Y; se procede de modo análogo para las Z.

Supondremos los nodos de cada árbol, terminales o no, numerados. Sea $\mathcal{Y}_k^{(j)}$ el k-ésimo nodo (terminal o no) del árbol $\mathcal{Y}_X^{(j)}$. Designaremos con el mismo símbolo $\mathcal{Y}_k^{(j)}$ el nodo del árbol, la región de $\mathcal{X}$ con valores de X que pueden dar lugar a que un caso transite por, o finalice en, dicho nodo, y el subconjunto de casos que lo hacen en la encuesta donante.

Por ejemplo, supongamos $q = 2$ (hay por tanto dos variables Y_1 e Y_2 a imputar en la encuesta A, para las que hemos construido dos árboles $\mathcal{Y}_X^{(1)}$ e $\mathcal{Y}_X^{(2)}$). Imaginemos que los árboles construidos tienen una forma tan simple como la recogida

Figura 2: Arboles $\mathcal{Y}_X^{(1)}$ e $\mathcal{Y}_X^{(2)}$. Junto a cada nodo no terminal aparece la condición que, de verificarse, da lugar a clasificar en el hijo derecho.



en la Figura 2.

Las particiones del espacio $\mathcal{X}$ realizadas por ambos árboles pueden entonces verse una junto a otra en la Figura 3. En cada caso se han rotulado las regiones de $\mathcal{X}$ con los nombres de la hojas correspondientes.

Para cualquier q-tupla $(\alpha_1, \ldots, \alpha_q)$ tal que α_j ($j \in \{1, \ldots, q\}$) sea etiqueta de un nodo en el árbol j, definimos

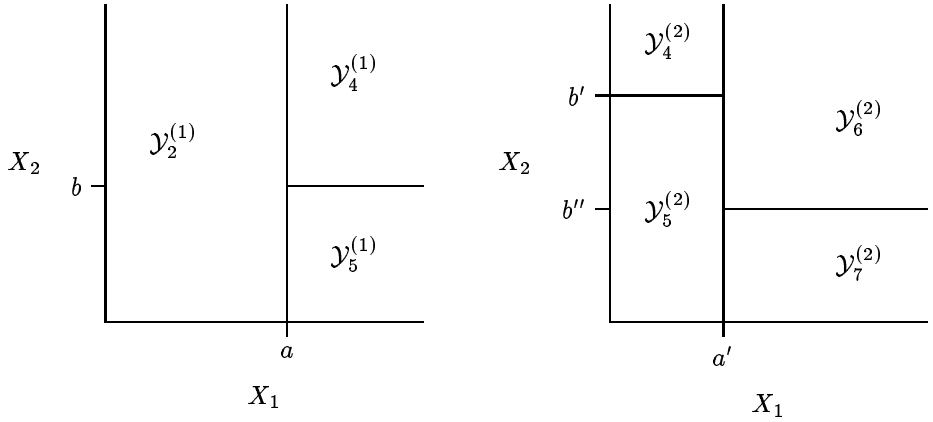
$$\mathcal{C}_{\alpha_1,\ldots,\alpha_q} \quad = \quad \mathcal{Y}_{\alpha_1}^{(1)} \cap \mathcal{Y}_{\alpha_2}^{(2)} \cap \ldots \cap \mathcal{Y}_{\alpha_q}^{(q)}. \tag{1}$$

Consecuentes con la definición anterior, $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$ designará a un tiempo un subconjunto de casos de la muestra de entrenamiento —el de aquéllos que al ser procesados por los árboles $\mathcal{Y}^{(1)}, \ldots, \mathcal{Y}^{(q)}$ respectivamente transitan por, o finalizan en, los nodos $\mathcal{Y}_{\alpha_1}, \ldots, \mathcal{Y}_{\alpha_q}$— y una región de $\mathcal{X}$ —la formada por los valores de las variables comunes X que dan lugar a que un caso esté en la clase $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$. Designaremos, finalmente, por $\mathcal{Y}_{(\uparrow\alpha_k)}^{(j)}$ al nodo "padre" del $\mathcal{Y}_{\alpha_k}^{(j)}$ en el árbol $\mathcal{Y}_X^{(j)}$. Cuando no exista duda del árbol al que nos referimos, designaremos el nodo simplemente por α_k y a su padre por $(\uparrow \alpha_k)$.

Sea ahora el caso i con $i \in \{N_A+1, \ldots, N\}$, para el que deseamos imputar $\mathbf{Y_i}$. Imaginemos que al clasificar dicho caso con ayuda de los árboles construidos para las variables Y, finaliza en las hojas $\mathcal{Y}_{i_1}^{(1)}, \ldots, \mathcal{Y}_{i_q}^{(q)}$. La idea es entonces imputar $\mathbf{Y_i}$ como función de los vectores $\mathbf{Y}$ correspondientes a casos en la encuesta donante pertenecientes a $\mathcal{C}_{i_1,\ldots,i_q}$, es decir, que al ser procesados por cada uno de los árboles finalizan precisamente en las mismas hojas que el caso a imputar i. De nuevo, al igual que en el Algoritmo 1, las opciones son ahora muchas: imputar mediante un vector tomado al azar de $\mathcal{C}_{i_1,\ldots,i_q}$, mediante la media de todos o algunos, etc.

En el ejemplo anterior, supongamos un caso con $a' < X_1 < a$ y $X_2 < b''$; terminará en las hojas $\mathcal{Y}_2^{(1)}$ e $\mathcal{Y}_7^{(2)}$ al dejarlo caer respectivamente por los árboles

Figura 3: Particiones del espacio $\mathcal{X}$ realizadas respectivamente por los árboles $\mathcal{Y}_X^{(1)}$ e $\mathcal{Y}_X^{(2)}$.



$\mathcal{Y}_X^{(1)}$ e $\mathcal{Y}_X^{(2)}$. La intersección de dichas dos hojas,

$$\mathcal{C}_{2,7} = \mathcal{Y}_2^{(1)} \cap \mathcal{Y}_7^{(2)}, \tag{2}$$

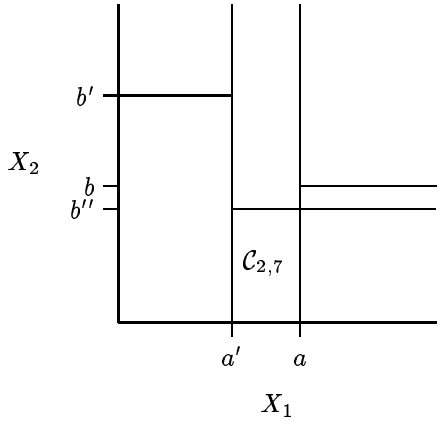
es el conjunto de casos en la muestra de entrenamiento que caen en los mismos nodos que el caso analizado o, alternativamente, que tienen valores de (X_1, X_2) en la región señalada como $\mathcal{C}_{2,7}$ de la Figura 4.

Si entre las variables Y_1 e Y_2 existe relación, tiene sentido imputarlas conjuntamente en función de los valores que esas variables toman en los casos de la muestra de entrenamiento que caen en la intersección. Buscar intersecciones supone afinar las particiones de $\mathcal{X}$ que determinan cada uno de los árboles $\mathcal{Y}_X^{(1)}$ e $\mathcal{Y}_X^{(2)}$ por separado, considerando "semejantes" al caso a imputar aquéllos que han caído en las mismas hojas tanto de $\mathcal{Y}_X^{(1)}$ como de $\mathcal{Y}_X^{(2)}$.

Un problema que puede presentarse en el método propuesto es que no exista ningún individuo de la muestra de entrenamiento que caiga en los mismos nodos que el caso a imputar i; es decir, que la intersección $\mathcal{C}_{i_1,\ldots,i_q}$ de las hojas en las que cae dicho caso i sea vacía. Una solución a este problema es, partiendo de las hojas $\mathcal{Y}_{i_1}^{(1)}, \ldots, \mathcal{Y}_{i_q}^{(q)}$ donde ha caído el individuo i, ir "trepando", esto es, sustituyendo sucesivamente los nodos por sus "padres". Treparíamos por los árboles $\mathcal{Y}_X^{(1)}, \ldots, \mathcal{Y}_X^{(q)}$ hasta encontrar una intersección no vacía. La idea es eliminar paulatinamente divisiones en algunos árboles para conseguir una intersección no vacía, procurando al hacerlo que se pierda lo mínimo posible en calidad de imputación.

Veamos cómo hacerlo. En lo que sigue, como hasta ahora, nos referimos exclusivamente a la imputación de las variables Y; con las Z se procedería de modo

Figura 4: Particiones superpuestas del espacio $\mathcal{X}$ realizadas respectivamente por los árboles $\mathcal{Y}_X^{(1)}$ e $\mathcal{Y}_X^{(2)}$, mostrando $\mathcal{C}_{2,7}$.



análogo. Por concreción supondremos también que las Y son continuas, y estamos ante árboles de regresión, pero la idea es generalizable a variables Y cualitativas y árboles de clasificación.

Como sabemos (ver, por ejemplo, Breiman et al. (1984), Cap. 3), en el proceso de construcción de un árbol de regresión se subdividen los nodos en tanto ello reporte alguna mejora en términos de desviación *(deviance)* —por lo general, en árboles de regresión, suma de cuadrados residual—. En el contexto de árboles de regresión, llamamos $R(t)$ a la desviación en el nodo t, y $R(T)$ a la desviación total del árbol T, definidas respectivamente así:

$$R(t) = \sum_{i \in t} (y_i - \overline{y}_t)^2 \tag{3}$$

$$R(T) = \sum_{t \in \tilde{T}} R(t), \tag{4}$$

en que $\tilde{T}$ denota el conjunto de hojas o nodos terminales del árbol T e $\overline{y}_t$ es la media aritmética de los valores que la variable respuesta toma en los casos caídos en el nodo t.

El coste de trepar en un árbol cualquiera $\mathcal{Y}_X^{(j)}$, $j = 1, \ldots, q$, del nodo hijo t_h al nodo padre t_p puede evaluarse de la siguiente forma:

$$c^{(j)}(t_h) = \frac{\sum_{i=1}^{N_p} (y_{ij} - \overline{y}_{j,t_p})^2}{N_p} - \frac{\sum_{i=1}^{N_h} (y_{ij} - \overline{y}_{j,t_h})^2}{N_h} \tag{5}$$

$$= \hat{R}(t_p)/N_p - \hat{R}(t_h)/N_h \qquad \forall j = 1, .., q, \tag{6}$$

en que $\hat{R}(t_h)$ y $\hat{R}(t_p)$ son estimaciones por resustitución de la desviación en el nodo "hijo" t_h (es decir, el nodo desde el que queremos trepar) y en su nodo "padre" t_p respectivamente, N_p y N_h son el número de casos en los nodos t_p y t_h, e $\overline{y}_{j,t_p}, \overline{y}_{j,t_h}$ las medias respectivas de la variable Y_j en dichos nodos.

Con la notación anterior, estamos ya en situación de especificar un algoritmo de imputación; es el recogido como Algoritmo 2. Nótese que el modo en que se hace la imputación a partir de una intersección $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$ queda sin especificar: cualquiera de las opciones mencionadas anteriormente (media, mediana, imputación múltiple, etc.) es utilizable. En la medida en que el procedimiento selecciona grupos de casos de la encuesta donante similares (en lo que se refiere a los valores de las variables X) al que tratamos de imputar, la imputación múltiple es fácil de hacer. Nótese también que es posible hacer la imputación sobre las variables originales o transformaciones de las mismas, como componentes principales. Esto último se ha hecho en la aplicación que mostramos en la Sección 5.

Algoritmo 2 – Imputación multivariante mediante árboles.

1: *(opcionalmente)* Calcular las componentes principales del grupo de variables Y a imputar utilizando la muestra de entrenamiento (la encuesta donante).
2: Construir los árboles $\mathcal{Y}_X^{(1)}, \ldots, \mathcal{Y}_X^{(q)}$.
3: **for** $i \in$ {Casos a imputar} **do**
4: Dejarlo caer por cada árbol, y a partir de las hojas $\mathcal{Y}_{\alpha_1}^{(1)}, \ldots, \mathcal{Y}_{\alpha_q}^{(q)}$ en que cae determinar $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$.
5: **if** $\mathcal{C}_{\alpha_1,\ldots,\alpha_q} \neq \emptyset$ **then**
6: **break**
7: **else**
8: **while** $\mathcal{C}_{\alpha_1,\ldots,\alpha_q} = \emptyset$ **do**
9: Computar los costes $c^{(1)}(\alpha_1), \ldots, c^{(q)}(\alpha_q)$ de trepar desde los nodos actuales.
10: Seleccionar k tal que la trepa desde el nodo α_k sea de mínimo coste.
11: $\alpha_k \leftarrow \quad (\uparrow \alpha_k)$; sustituir el nodo α_k por su padre.
12: **end while**
13: **end if**
14: Imputar el caso i a partir de $\mathcal{C}_{\alpha_1,\ldots,\alpha_q}$.
15: **end for**
16: *(si procede)* Reconstruir las variables originales a partir de los valores imputados de las componentes principales.

5 Aplicación a la Encuesta de Presupuestos de Tiempo (EPT-93)

A continuación ilustraremos el funcionamiento del método enlazando los ficheros `trabajo.dat` y `fiesta.dat`, mencionados en la Introducción. En cada uno

Tabla 1: Variables comunes (o de caracterización) junto a sus modalidades.

Variable	**Descripción**	**Código**	**Modalidad**
X_1	Edad	EDA1	Hasta 34 años.
		EDA2	Entre 35 y 59 años.
		EDA3	60 años o más.
X_2	Sexo	VARO	Varón.
		MUJE	Mujer.
X_3	Estado civil	SOLT	Soltero.
		CASD	Casado.
		REST	Resto.
X_4	Nivel de instrucción	PRIM	Primarios.
		MEDI	Medios.
		SUPE	Superiores.
X_5	Relación con la actividad	SRMI	Servicio militar.
		OCUP	Ocupados.
		PARA	Parados.
		JUBI	Jubilados.
		ESTD	Estudiantes.
		LAHO	Labores del hogar.
		OTRS	Otros.

de los ficheros hay dos tipos de variables:

- Variables comunes o de caracterización X cuya descripción y modalidades aparecen en la Tabla 1
- Variables específicas: tiempo en minutos al día dedicados a las actividades que aparecen en la Tabla 2.

Ambos ficheros proceden de una misma operación de muestreo, y pueden verse como muestras independientes procedentes de la misma población; pueden verse detalles sobre el mismo problema en Bárcena y de Agirre (1998). Se trata de un caso particular del problema de enlace de encuestas descrito en la Sección 1 , en que `trabajo.dat` y `fiesta.dat` son respectivamente los ficheros A y B. Si colocamos los datos de ambos ficheros formando una matriz incompleta (como la que aparece en la Figura 1), tenemos en este caso: p=5, $q = r$ =24, $N_A = 2521$ y $N_B = 2519$.

Hemos enlazado los ficheros `trabajo.dat` y `fiesta.dat` siguiendo el método explicado en la sección anterior (Algoritmo 2). Los cálculos necesarios se han realizado mediante varias funciones programadas en el lenguaje específico del paquete estadístico S-PLUS. Una descripción de dicho paquete puede verse en Becker et al. (1988) y Chambers y Hastie (1992).

Primero se realizó un Análisis en Componentes Principales (ACP) tipificado de los valores disponibles de las variables Y y Z; el análisis se realiza tipificado ya que las variables Y y Z tienen muy diferente dispersión.

A continuación se construyó el árbol de regresión sobre las variables comunes X de cada componente principal. Para cada componente se obtiene una sucesión de árboles $T_1 \succ T_2 \succ ..... \succ \{t_1\}$ ordenados de más a menos frondoso: $\{t_1\}$ es el árbol degenerado que consiste sólo en el nodo raíz. Se estima mediante validación cruzada la desviación $R(T)$ para cada árbol de la sucesión y se selecciona como árbol T^* de tamaño óptimo el que tenga menor $\hat{R}^{vc}(T)$ (desviación estimada). La validación cruzada se realizó dividiendo aleatoriamente la muestra en 10 bloques, y utilizando por turno las observaciones de cada bloque para estimar la tasa de error de los árboles construidos con los nueve bloques restantes.

Una vez construidos los árboles para cada una de las componentes principales, se ejecuta el bucle principal del Algoritmo 2. En esta particular aplicación optamos por conservar todas las componentes principales dando lugar a árboles con más de un nodo (todas menos dos), dada la casi total falta de correlación entre las variables originales, que no aconsejaba reducir su dimensionalidad. Optamos también por imputar al azar dentro de cada clase de equivalencia: cuando se encuentra una intersección $\mathcal{C}_{\alpha_1,...\alpha_q}$ no vacía se escoge aleatoriamente un sólo caso de la misma y se asignan los valores de sus componentes principales al caso a imputar.

El número de individuos que han requerido trepar para encontrar una intersección es de 33 para las variables Y (tan sólo un 1.31% de los casos) y de 48 para las Z (1.91%). Hemos comprobado que para los individuos para los que es necesario trepar, se trepa principalmente por los árboles de las últimas componentes principales (como era de esperar, ya que son los árboles de las componentes con menos dispersión los que cabe imaginar menos costosos de trepar). En varios casos es necesario trepar hasta el nodo raíz de algunos árboles.

El resultado final es una matriz de datos como la representada en la Figura 1 pero completada.

Con el fin de estudiar la calidad del enlace comparamos, desde un punto de vista descriptivo, las distribuciones de las variables Y y Z para valores imputados con las obtenidas para valores disponibles. Un buen enlace de encuestas no debe alterar la distribución de Y ni de Z, y es práctica habitual (véase Lebart y Lejeune (1995)) comparar las distribuciones marginales de las variables en la encuesta donante con las de las mismas variables imputadas en la encuesta receptora.

Así lo hemos hecho nosotros comparando media, mediana, cuartiles y valores extremos de los valores disponibles de las variables específicas (Y y Z) y los obtenidos por imputación. Los resultados para Y se muestran en la Tabla 3 (fueron análogos para Z, y no se reproducen en interés de la brevedad.) Las principales diferencias se dan en la variable Y_7 ("Tareas domésticas") e Y_{13} ("Reuniones con amigos").

También hemos comparado las matrices de correlación para valores disponibles e imputados. De nuevo se representan sólamente (en la Figura 5) las correspondientes a las variables Y; análogos resultados se obtuvieron para las Z.

Tabla 2: Variables específicas de uso del tiempo. Las variables $Y_1, \ldots, Y_{24}$ corresponden a usos del tiempo en días laborables, y las $Z_1, \ldots, Z_{24}$ son las variables homólogas para los días no laborables.

Variables	**Descripción**
Y_1, Z_1	Dormir.
Y_2, Z_2	Higiene y cuidado personal.
Y_3, Z_3	Comer.
Y_4, Z_4	Actividades privadas y actividades no descritas.
Y_5, Z_5	Trabajo.
Y_6, Z_6	Formación.
Y_7, Z_7	Tareas domésticas (cocinar, fregar, limpiar la casa arreglo y cuidado ropa y trabajos diversos).
Y_8, Z_8	Compras.
Y_9, Z_9	Gestiones.
Y_{10}, Z_{10}	Actividades de semi ocio (punto, costura, pintura, escultura, reparaciones, bricolaje, jardinería, cuidado de animales..).
Y_{11}, Z_{11}	Cuidado de niños y adultos.
Y_{12}, Z_{12}	Reuniones de tipo familiar (comidas, defunciones, bodas, visitas hospitalarias..).
Y_{13}, Z_{13}	Reuniones con amigos, fiestas, ir de potes o copas...
Y_{14}, Z_{14}	Participación religiosa o política.
Y_{15}, Z_{15}	Gimnasia y deporte.
Y_{16}, Z_{16}	Excursiones y paseos.
Y_{17}, Z_{17}	Ocio en el hogar (TV, vídeo, música, radio...).
Y_{18}, Z_{18}	Ocio fuera del hogar (cine, teatro, conciertos, museos y exposiciones, espectáculos deportivos..).
Y_{19}, Z_{19}	Otras actividades de ocio (micro informática, fotografía, cartas, juegos, crucigramas...).
Y_{20}, Z_{20}	Trayecto al trabajo o formación.
Y_{21}, Z_{21}	Acompañar a otros.
Y_{22}, Z_{22}	Esperas en el trabajo o formación.
Y_{23}, Z_{23}	Esperas en cuidados médicos y gestiones administrativas.
Y_{24}, Z_{24}	Otras esperas.

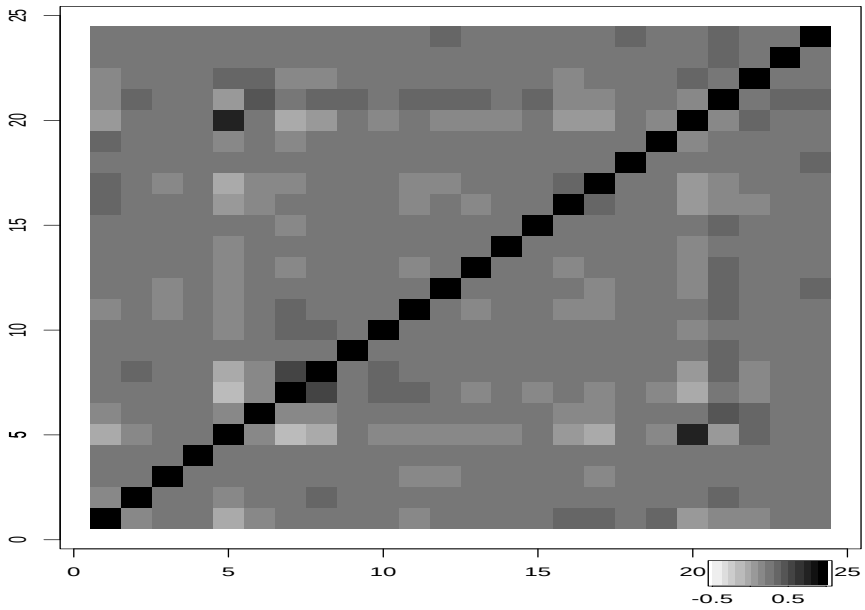
Tabla 3: Imputación de variables de uso del tiempo en días laborables. Estadísticos de posición y dispersión para datos observados e imputados de variables Y. Se han señalado en negrita las variables cuya diferencia de medias entre datos observados y completados excede de dos desviaciones típicas.

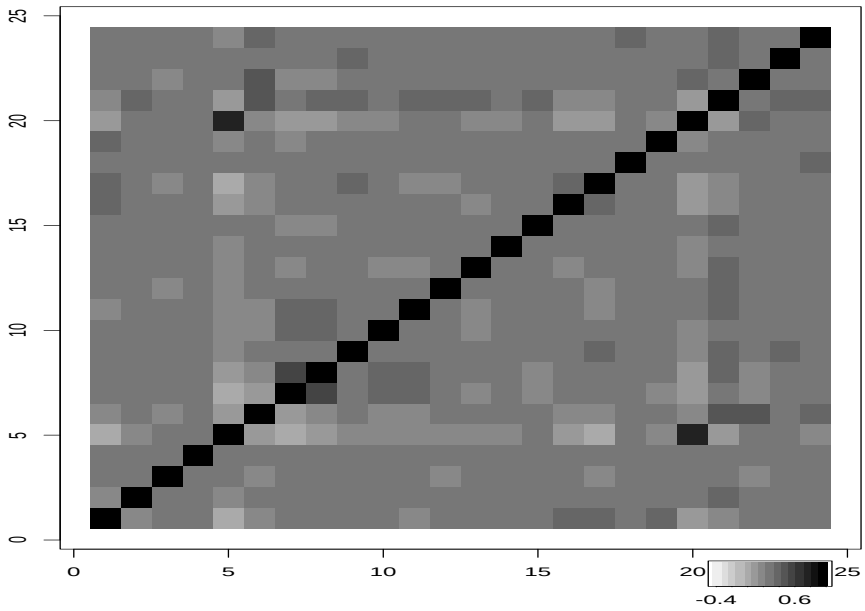
	Valores observados						Valores imputados					
	Min.	q_1	Me	$\overline{x}$	q_3	Max.	Min.	q_1	Me	$\overline{x}$	q_3	Max.
Y_1	179	450	499	511.30	559	1214	179	450	495	508.50	555	975
Y_2	0	20	35	42.23	55	295	0	20	35	41.90	55	295
Y_3	0	75	105	109.90	135	374	0	75	100	109.90	135	374
Y_4	0	0	0	0.33	0	105	0	0	0	0.25	0	75
Y_5	0	0	0	196.50	455	990	0	0	0	194.70	460	929
$\mathbf{Y_6}$	0	0	0	23.84	0	760	0	0	0	56.25	0	760
$\mathbf{Y_7}$	0	0	75	133.10	235	770	0	0	45	107.60	185	630
$\mathbf{Y_8}$	0	0	0	30.22	55	355	0	0	0	25.34	45	330
Y_9	0	0	0	2.73	0	415	0	0	0	3.05	0	240
$\mathbf{Y_{10}}$	0	0	0	16.90	0	634	0	0	0	13.81	0	634
$\mathbf{Y_{11}}$	0	0	0	20.53	0	735	0	0	0	15.71	0	625
$\mathbf{Y_{12}}$	0	0	0	7.32	0	630	0	0	0	8.90	0	630
$\mathbf{Y_{13}}$	0	0	0	40.43	60	510	0	0	5	47.35	75	510
Y_{14}	0	0	0	4.64	0	365	0	0	0	5.12	0	350
$\mathbf{Y_{15}}$	0	0	0	6.91	0	315	0	0	0	8.39	0	230
Y_{16}	0	0	0	54.83	90	600	0	0	0	53.92	90	600
$\mathbf{Y_{17}}$	0	74	140	163.00	225	964	0	65	135	157.70	224	730
$\mathbf{Y_{18}}$	0	0	0	1.48	0	239	0	0	0	2.09	0	239
Y_{19}	0	0	0	10.76	0	405	0	0	0	9.59	0	350
Y_{20}	0	0	0	20.48	30	340	0	0	0	21.14	30	340
$\mathbf{Y_{21}}$	0	0	10	33.32	50	355	0	0	15	39.25	60	355
Y_{22}	0	0	0	1.49	0	165	0	0	0	1.81	0	165
Y_{23}	0	0	0	0.50	0	165	0	0	0	0.56	0	165
Y_{24}	0	0	0	0.79	0	75	0	0	0	0.94	0	75

Se observa su gran similitud aparente: la máxima diferencia en valor absoluto es de 0.1124 (que corresponde a la correlación entre Y_5 e Y_6).

Como conclusión de esta aplicación observamos que el método de enlace descrito es factible, reproduce aceptablemente las distribuciones marginales de las variables Y y Z y sus respectivas estructuras de correlación. La información recuperada sobre la relación entre las Y y las Z (a la que hemos omitido en lo que antecede) es escasa, confirmando la necesidad de un conjunto de variables comunes X suficientemente rico y descriptivo (cf. Lejeune (1995)). Detalles adicionales sobre la aplicación pueden obtenerse de Bárcena y Tusell (1998).

Figura 5: Representación gráfica de las matrices de correlación de las Y observadas (arriba) e imputadas (abajo). Nótese la disposición de las variables en los ejes y la gran similitud entre ambas.





Referencias

Aluja, T., R. Nonell, R. Rius, y M. Martínez (1995). File grafting. En Mola y Morineau (1997), págs. 23–32.

Aluja, T. y R. Rius (1994). Inserción de datos de encuesta mediante análisis de componentes principales. *Presented at the XXI Congreso nacional de Estadística e Investigación Operativa (SEIO). Calella.*

Becker, R., J. Chambers, y A. Wilks (1988). *The New S Language. A Programming Environment for Data Analysis and Graphics.* Wadsworth & Brooks/Cole, Pacific Grove, California.

Bello, A. (1993). Choosing among imputation techniques for incomplete multivariate data: a simulation study. *Communications in Statistics - Theory and Methods*, vol. 22(3), págs. 853–877.

Bárcena, M. y K. F. de Agirre (1998). Visualization Techniques Related to Survey Linking: Two Alternative Ways and Bootstrap Validation. BILTOKI DT98.14, Universidad del País Vasco.

Bárcena, M., K. Fdez. Aguirre, y F. Tusell (1997). Survey grafting with common structural base: Time use Survey and Health Survey. En Mola y Morineau (1997).

Bárcena, M. y F. Tusell (1997). Linking surveys using reciprocal classification trees. En *Analyses Multidimensionelles des Donées* (eds. K. Fernández-Aguiree y A. Morineau). CISIA.

Bárcena, M. y F. Tusell (1998). Enlace de encuestas: una propuesta metodológica y aplicación a la Encuesta de Presupuestos de Tiempo. Inf. Téc. 98.07, Departamento de Econometría y Estadística.

Breiman, L., J. Friedman, R. Olshen, y C. Stone (1984). *Classification and Regression Trees.* Wadsworth, Belmont, California.

Buck, S. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Ser. B*, vol. 22, págs. 302–306.

Chambers, J. y T. Hastie (1992). *Statistical Models in S.* Wadsworth & Brooks/Cole, Pacific Grove, Ca.

Dempster, A., N. Laird, y D. Rubin (1976). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, vol. 39, págs. 1–38.

EUSTAT (1997). *Análisis de Tipologías de Jornadas Laborales.* Instituto Vasco de Estadística, (EUSTAT), Vitoria/Gazteiz.

Lebart, L. y M. Lejeune (1995). Assessment of Data Fusions and Injections. En *Encuentro Internacional AIMC sobre Investigación de Medios*, págs. 1–18. Madrid.

Lejeune, M. (1995). De l'usage des fusions de données dans les études de marché. En *Proceedings of the IASS Meeting, Beijing*.

Little, R. y D. Rubin (1987). *Statistical Analysis with Missing Data*. John Wiley and Sons, New York.

Mola, F. y A. Morineau (eds.) (1997). *Analyses Multidimensionnelles des Données*, 1 avenue Herbillon, 94160 SAINT-MANDE (France). III Congrès International NGUS'95 - Naples 1995., CISIA.

Nordbotten, S. (1996). Neural Network Imputation Applied to the Norwegian 1990 Population Census Data. *Journal of Official Statistics*, vol. 12(4), págs. 385–401.

Rubin, D. (1986). Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations. *Journal of Business and Economic Statistics*, vol. 4(1), págs. 87–94.

Villagarcía, T. y A. Muñoz (1997). Imputación de datos censurados mediante redes neuronales: una aplicación a la EPA. *Cuadernos Económicos de I.C.E.*, (63), págs. 193–204.