Fusión de encuestas mediante árboles de regresión y/o clasificación: aplicación a la Encuesta de Presupuestos de Tiempo

María Jesús BARCENA y Fernando TUSELL
Departamento de Estadística y Econometría.
Facultad de CC.EE. y Empresariales, Avda. del Lehendakari Aguirre, 83, 48015 BILBAO.
E-mail: etptupaf@bs.ehu.es

ABSTRACT

We address the problem of completing two files with records containing a common subset of variables. The technique investigated involves the use of regression and/or classification trees.

Keywords: file matching; survey grafting; imputation; regression trees.

1.INTRODUCCION

El punto de partida en un problema de fusión de encuestas son dos ficheros, A y B, con un total de $N = N_A + N_B$ observaciones, que contienen datos de dos encuestas diferentes realizadas en una misma población. Ambas encuestas tienen un grupo de variables en común, $X_1,...,X_p$, cuyos valores son conocidos para los N individuos. Cada encuesta tiene además otras variables específicas sobre el tema de la encuesta que sólo son medidas para los individuos de esa encuesta: $Y_1,...,Y_q$ y $Z_1,...,Z_r$ según nuestra notación.

Encuesta A

$X_1,...,X_p$	$Y_1,...,Y_q$	**No observado**

Encuesta B

$X_1,...,X_p$	**No observado**	$Z_1,...,Z_r$

El objetivo de la fusión es completar los ficheros A y B, utilizando los valores de las variables comunes X; es decir, imputar los valores de las variables específicas de una encuesta para los individuos de la otra encuesta. Esto nos permitiría realizar 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.

Ser capaces de fusionar encuestas con suficiente calidad supondría un gran ahorro a la hora de recoger información: una encuesta larga ---con muchas variables--- podría sustituirse por encuestas más pequeñas que son menos costosas. Además permitiría

AGRADECIMIENTOS: Este trabajo ha sido realizado como parte integrante de la tesis doctoral del primer autor. Agradecemos la financiación recibida del MEC (proyecto PB95-0346) y de la UPV/EHU(proyecto PB95-0346).

aprovechar mejor la información ya existente en otras encuestas realizadas sobre la misma población. Por ello es un tema interesante, al cual se han aplicado diversas técnicas(regresión, algoritmo EM, métodos de reemplazamiento, técnicas factoriales multivariantes...). En este trabajo desarrollaremos una nueva técnica de fusión que emplea árboles de regresión y/o clasificación y la aplicaremos 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 del tiempo es diferente en días de distinto tipo. Por ello, hemos separado los datos de 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 fusionar los ficheros 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.

Este trabajo se organiza de la siguiente manera: en la Sección 2 proponemos un método de fusión de encuestas que emplea árboles de regresión y/o clasificación. En la Sección 3 aplicaremos este método a los datos de EPT-93.

2. ENLACE DE ENCUESTAS MEDIANTE ARBOLES DE REGRESION Y/O CLASIFICACION

Proponemos el uso de árboles de regresión y/o clasificación como método de imputación en un problema de fusión de encuestas. 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 la fusión y permite realizar fácilmente imputación múltiple (que consiste en buscar varios valores posibles con los que imputar cada valor desconocido, lo cual da idea de la imprecisión introducida en los resultados por el propio procedimiento de imputación; véase Rubin, 1986). 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.

En Bárcena y Tusell, 1997, estudiamos la fusión mediante árboles en el caso más sencillo en que los grupos de variables específicas Y y Z son univariantes. En el presente trabajo se aborda el caso más general, cuando Y y Z son multivariantes.

Trataremos de 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-Lejeune, 1995, a este respecto).
Un problema con el que nos encontramos, a la hora de realizar esta imputación conjunta mediante árboles, es que la metodología existente sobre árboles de regresión y/o clasificación sólo contempla variables respuesta univariantes. Por ello, hemos diseñado un método que permite imputar simultáneamente los valores desconocidos para cada individuo utilizando los árboles construidos para cada variable específica en función de las variables comunes X.
A continuación, se describe el procedimiento para el caso de imputar las variables Y; se procede de modo análogo para las Z.

Con los datos disponibles de (X,Y) se construyen los árboles para cada una de las variables que forman el grupo Y en función de las variables comunes X.
Sea ahora el caso i con $i \in \{N_A+1,...,N\}$ (es decir, un individuo de la segunda encuesta) para el que deseamos imputar el vector de valores de las variables Y, Y . Imaginemos que al clasificar dicho caso con ayuda de los q árboles construidos para las variables Y, finaliza en las hojas o nodos terminales y ,...,y . La idea es entonces imputar Y como función de los vectores correspondientes a casos en la encuesta donante (encuesta A) que al ser procesados por cada uno de los árboles finalizan precisamente en las mismas hojas que el caso a imputar *i*; a este conjunto de casos le denominamos intersección y denotaremos por C . Hay varias opciones: imputar mediante un vector tomado al azar de C , escoger varios dentro de C y realizar imputación múltiple, mediante la media de todos o algunos, etc.

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 C de las hojas en las que cae dicho caso *i* sea vacía. Una solución a este problema es, partiendo de las hojas y ,...,y donde ha caído el individuo *i*, ir “trepando”, esto es, sustituyendo sucesivamente los nodos por sus “padres”. Treparíamos por los árboles 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.

Una descripción más detallada de este método puede verse en Bárcena y Tusell, 1998.

En el método que acabamos de describir, pueden utilizarse los árboles construídos para las variables originales (Y, Z) o para sus componentes principales. Una ventaja de utilizar árboles sobre las componentes principales es que el número de componentes relevantes puede ser inferior al número de variables originales. Si es así, se simplifica la busqueda de intersecciones y el proceso de trepa, ya que hay que tener en cuenta un número menor de árboles. Además, las componentes principales recogen relaciones lineales entre las variables a imputar, que pueden o no tener relación con las variables comunes X. Si se utilizan árboles construídos para las componentes, cabe la posibilidad de recoger estas relaciones. En la aplicación a la encuesta EPT-93 se han utilizado los árboles construídos para las componentes principales por este motivo.

3. 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 de los ficheros hay dos tipos de variables:

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

Se trata de un caso particular de fusión de encuestas en que TRABAJO.DAT y FIESTA.DAT son respectivamente los ficheros A y B. 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. Los cálculos necesarios se han realizado mediante varias funciones programadas en el lenguaje específico del paquete estadístico S-Plus.

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. En esta particular aplicación optamos por conservar todas las componentes principales dando lugar a árboles con más de un nodo, 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 C 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 Z (1.91%). Una vez imputados los valores de las componentes principales para todos los individuos se reconstruyen, a partir de ellos, los valores de las variables originales (según la fórmula de reconstrucción propia de un ACP tipificado).

El resultado final es una matriz de datos de los tres grupos de variables (X,Y,Z) completada para los N individuos.
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-Lejeune, 1995) comparar las distribuciones marginales de las variables específicas 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. También hemos comparado las matrices de correlación para valores disponibles e imputados. Los resultados obtenidos para datos reales e imputados son similares.

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 sobre la relación entre las Y y las Z es en la aplicación realizada escasa, confirmando la necesidad de un conjunto de variables comunes X suficientemente rico y descriptivo (cf. Lejeune,1995).

ANEXO1

Variables comunes (de caracterización) junto a sus modalidades

Variable	Descripción	Identificador	Modalidad
X1	Edad	ED1 ED2 ED3	Hasta 34 años Entre 35 y 59 años 60 años o más
X2	Sexo	VAR MUJ	Varón Mujer
X3	Estado civil	SOL CAS RES	Soltero Casado Resto
X4	Nivel de instrucción	PRI MED SUP	Primarios Medios Superiores
X5	Relación con la actividad	SRM OCU PAR JUB EST LAH OTR	Servicio militar Ocupados Parados Jubilados Estudiantes Labores del hogar Otros

ANEXO 2

Variables específicas de uso del tiempo. Las variables Y1,...,Y24 corresponden a usos del tiempo en días laborables, y Z1,...,Z24 son las variables homólogas para los días festivos.

Variables	Descripción
Y1,Z1	Dormir
Y2,Z2	Higiene y cuidado personal
Y3,Z3	Comer
Y4,Z4	Actividades privadas y actividades no descritas
Y5,Z5	Trabajo
Y6,Z6	Formación
Y7,Z7	Tareas domésticas (cocinar, fregar, limpiar la casa arreglo y cuidado ropa y trabajos diversos)
Y8,Z8	Compras
Y9,Z9	Gestiones
Y10,Z10	Actividades de semiocio (punto, costura, pintura, escultura, reparaciones, bricolaje, jardinería, cuidado de animales..)
Y11,Z11	Cuidado de niños y adultos
Y12,Z12	Reuniones de tipo familiar (comidas, defunciones, bodas, visitas hospitalarias..)
Y13,Z13	Reuniones con amigos, fiestas, ir de potes o copas...
Y14,Z14	Participación religiosa o política
Y15,Z15	Gimnasia y deporte
Y16,Z16	Excursiones y paseos
Y17,Z17	Ocio en el hogar (TV, vídeo, música, radio...)
Y18,Z18	Ocio fuera del hogar (cine, teatro, conciertos, museos y exposiciones, espectáculos deportivos..)
Y19,Z19	Otras actividades de ocio (micro informática, fotografía, cartas, juegos, crucigramas...)
Y20,Z20	Trayecto al trabajo o formación
Y21,Z21	Acompañar a otros
Y22,Z22	Esperas en el trabajo o formación
Y23,Z23	Esperas en cuidados médicos y gestiones administrativas
Y24,Z24	Otras esperas

Referencias

Bárcena, M. J. and Tusell, F. (1997). Linking surveys using reciprocal classification trees. *Analyses Multidimensionelles des Données.* IV-éme Congrès International NGUS'97 . K. Fernández-Aguirre et A. Morineau (Eds.), p. 133-148, CISIA-CERESTA.

Bárcena, M. J. and Tusell, F. (1998). Enlace de encuestas: una propuesta metodológica y aplicación a la Encuesta de Presupuestos de Tiempo. *BILTOKI DT 98.07 , documentos de trabajo. Dpto de Estadística y Econometría Facultad de CC EE y Empresariales. Bilbao*

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, 1984.

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

Lebart, L., and Lejeune, M. (1995). Assessment of Data Fusion and Injection. *Encuentro Internacional AIMC sobre Investigación de Medios*, p. 1-18, Madrid, 1995

Lejeune, M. (1995). De L'usage des Fusions de Données dans les Etudes de Marché. IASS *Proceedings – Beijing 1995*, p. 137-149

Rubin, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics, 4 (1):87-94,1986.*