ENLACE DE ENCUESTAS

2014

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA
www.eustat.eus

Organismo Autónomo del
EUSKO JAURLARITZA
GOBIERNO VASCO

Elaboración:
EUSTAT
Euskal Estatistika Erakundea
Instituto Vasco de Estadística

Edición:
EUSTAT
Euskal Estatistila Erakundea
Instituto Vasco de Estadística
Donostia-San Sebastián 1
01010 Vitoria-Gasteiz

Primera Edición
I/2015

Impresión y Encuadernación:
Servicio de Imprenta y Reprografía del Gobierno Vasco-Eusko Jaurlaritza

ISBN: 978-84-7749-482-9

Depósito Legal: VI-794/2014

ENLACE DE ENCUESTAS

Ines Garmendia Navarro

inesgarmendia@gmail.com

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.eus
www.eustat.eus

Presentación

Desde la estadística oficial se están realizando grandes esfuerzos en el estudio de técnicas de integración datos, a fin de proporcionar a sus usuarios información de la mayor calidad posible. En este sentido, incluso se organizó el XXVI Seminario Internacional de Estadística bajo el lema "Statistical Matching: Methodological issues and practice with R-StatMatch" ("*Enlace de encuestas: Cuestiones metodológicas y práctica con R-StatMatch*").

La presente publicación pretende dar a conocer el trabajo de investigación realizado por una becaria en este campo. Este cuaderno está dividido en cuatro apartados principales: el primero está dedicado a la metodología y el segundo describe las técnicas en el entorno R. El tercero se dedica a exponer un caso práctico de enlace con dos encuestas independientes de Eustat, la Encuesta de Condiciones de Vida y la Encuesta de Población en Relación con la Actividad. Finalmente, el cuarto y último apartado describe el desarrollo de un paquete propio de R.

Vitoria-Gasteiz, Diciembre de 2014

Josu Iradi Arrieta

Director General de EUSTAT

Índice

Capítulo

1

Introducción

El contenido recogido en este Cuaderno Técnico es fruto del trabajo realizado durante el disfrute de la beca de formación e investigación en metodologías estadístico-matemáticas, para el tema de *Enlace de Encuestas*, concedida en el año 2012 por el Instituto Vasco de Estadística - Euskal Estatistika Erakundea.

El enlace de encuestas[1] es una metodología que permite elaborar estadísticas integradas e indicadores combinados partiendo de información de encuestas independientes referidas a una misma población de interés. La principal ventaja es que se explota de forma más eficiente la información proveniente de encuestas distintas, y que reside en ficheros de datos separados.

Esta metodología cubre un amplio abanico de técnicas, tales como la imputación datos *missing* o faltantes, la cuantificación de la incertidumbre, y la teoría del muestreo complejo. Estas técnicas se encuentran en continuo desarrollo, y muchas de ellas se distribuyen a través del entorno de software libre R, plataforma cada vez más impuesta en el mundo académico, en la industria y, gradualmente, también en la estadística oficial.

El cuaderno está organizado como sigue: tras el primer capítulo que contiene esta introducción, el segundo capítulo se dedica a la metodología y presenta las principales técnicas así como diversas recomendaciones para llevarlas a la práctica. El tercer capítulo se dedica a las posibilidades que ofrece el entorno R en cuanto a la implementación de estas técnicas. El cuarto capítulo expone un caso real de enlace entre dos encuestas independientes de Eustat, la Encuesta de Población en Relación con la Actividad y la Encuesta de Condiciones de Vida, y sirve para ilustrar las principales fases que debe comprender un enlace, así como el tipo de resultados que se obtienen. El quinto capítulo expone el desarrollo de un paquete propio de R, y el sexto y último capítulo se dedica a las conclusiones. Finalmente, se incluye un anexo que contiene una serie de tablas con resultados numéricos.

De forma paralela a la elaboración de este cuaderno técnico se desarrollado un paquete propio de R para facilitar la implementación de la metodología aquí presentada. Con este paquete propio, `micromatch`, se quiere ofrecer a usuario un entorno específicamente diseñado para enlazar encuestas, y que integra, en un contexto único, métodos implementados en otros paquetes, ya probados y contrastados (como `StatMatch` y `mice`). El paquete se presentó durante las VI Jornadas de Usuarios de R

[1] En inglés, la terminología es variada: se utilizan equivalentemente, y dependiendo de la fuente y del contexto, *statistical matching*, *data fusion*, *file merging*, *survey linking* y *synthetic matching*.

en Santiago de Compostela (el 23 y 24 de Octubre de 2014), también está disponible a través de la web de Eustat.

Quisiera aprovechar esta introducción para dar las gracias a todos los componentes del Área de Metodología, Innovación e I+D de Eustat, por su apoyo incondicional; en especial, a Elena Goñi, Anjeles Iztueta y Cristina Prado, que tan decisivamente me han ayudado a explorar el amplio mundo del enlace de encuestas, y a distinguir lo importante lo superficial. En general, agradezco muy sinceramente la amabilidad y el apoyo de todo el personal de Eustat.

Asimismo quisiera dar las gracias a Fernando Tusell, por sus consejos a la hora de programar el paquete de R; a Marcello d'Orazio (ISTAT), por tener la amabilidad de responder a nuestras dudas durante el seminario de Eustat dedicado a este tema; y, por último, a David Kaplan, profesor y Director del departamento de Psicología Educativa de la Universidad de Wisconsin-Madison (Estados Unidos), por ser tan amable de responder a una serie de cuestiones metodológicas sobre enlace de encuestas.

Por último agradezco el apoyo de toda mi familia. A mis hijos Ion y Miguel, que han sufrido no pocas veces la ausencia de su madre; y a Gaspar, que ha tenido la inmensa paciencia de acompañarme en todo este camino. Y, muy especialmente, a Rosa y Antonio, y a mis padres, Koro y Jesús, por habernos apoyado en todo momento.

PALABRAS CLAVE: Enlace de encuestas, fusión de datos, imputación de valores faltantes, R

Capítulo

2

Metodología

Fundamentos del enlace de encuestas

El enlace de encuestas comprende una serie de técnicas orientadas a obtener estadísticas integradas de indicadores o variables recogidas a través de diversas fuentes, generalmente, encuestas muestrales realizadas sobre una misma población.

En el caso más general, se parte de dos encuestas muestrales independientes referidas a una misma población (por ejemplo, los residentes de la C.A. de Euskadi en el año 2014), cada una de las cuales mide una serie de dimensiones o indicadores de forma separada (estilos de vida, situación laboral, ingresos...). A la hora de plantear el enlace, se requiere que las encuestas compartan una serie de variables o medidas comunes, habitualmente, variables sociodemográficas básicas tales como la edad, el sexo o el nivel de estudios.

A la hora de "unir" los datos recogidos por dos encuestas independientes[2] surge una situación como la de la Figura 1. El bloque Z simboliza las variables comunes (sociodemográficas y otras) recogidas por ambas encuestas. Los otros dos bloques, X e Y, simbolizan los ítems propios de cada cuestionario. Si bien la información del bloque común Z estará disponible para todos los registros (ésta es la premisa básica del enlace), en los dos bloques específicos (X e Y), solamente los registros de cada encuesta tendrán valores informados. Surgen así dos grandes bloques de valores "faltantes" o no-observados (partes más claras en la Figura 1).

[2] Por simplicidad, en este cuaderno sólo se ha considerado el enlace de dos encuestas, pero la metodología se puede aplicar igualmente a n>2 encuestas.

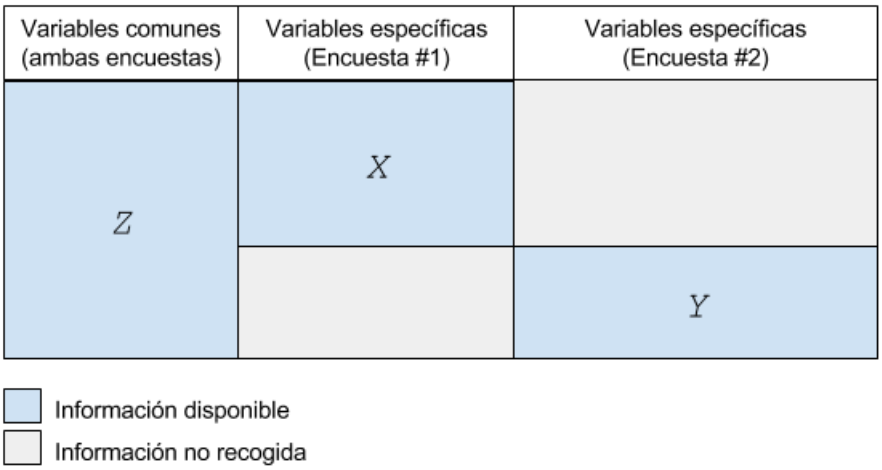

Figura 1. *Esquema que surge al "unir" los registros de dos encuestas que comparten un bloque Z de variables comunes. Los bloques X e Y simbolizan los ítems específicos o no compartidos entre las encuestas. A este proceso se le llama concatenación.*

Punto de partida:

* Existe un bloque Z de variables comunes, i.e. compartidas por ambos ficheros
* Se tienen dos bloques, X e Y, de variables específicas que no observadas conjuntamente: X figura sólo figura en el fichero #1, e Y sólo en el fichero #2.
* La probabilidad de que una unidad de la población figure en las dos muestras es próxima a cero y la podemos ignorar.

El enlace de encuestas plantea una situación sustancialmente diferente al de otras técnicas como la **fusión de registros**, en las que el objetivo es identificar unidades idénticas entre ficheros (por ejemplo, entre un censo y un fichero administrativo). El enlace de encuestas es distinto porque parte de encuestas muestrales independientes. Así la situación es en cierto modo la contraria: de partida se *sabe* que las unidades son distintas, pero se *busca* hallar parejas "similares" con el objetivo de relacionar *no las unidades*, sino las *variables* de las encuestas.

En resumen, el enlace de encuestas persigue relacionar variables específicas de encuestas muestrales independientes referidas a la misma población de interés, empleando como "puente" la información compartida entre ellas. A continuación veremos las principales aproximaciones a este problema así como las soluciones que se han adoptado a lo largo del tiempo.

Aproximaciones y métodos

A la hora de abordar un enlace de encuestas generalmente se adopta una de las siguientes aproximaciones (M. D'Orazio, M. Di zio & M. Scanu, 2008):

- La **aproximación macro** busca elaborar estimaciones directas de ciertas características de las variables específicas, por ejemplo, un coeficiente de correlación entre una variable X y otra variable Y, o una distribución marginal conjunta.

- La **aproximación micro** tiene como objetivo crear un fichero sintético con información completa de todas las variables específicas de los ficheros, y para todos los registros. Este fichero sintético posteriormente se emplea para realizar análisis conjuntos referentes a variables que inicialmente se encuentran en ficheros separados.

Hoy día existen múltiples métodos estadísticos para enlazar encuestas. En efecto, desde los inicios de esta metodología en los los años 60 del pasado siglo se han desarrollado múltiples soluciones, tanto en el área de investigación de mercados (en Europa) como en el área de la estadística oficial (en Estados Unidos y Canadá, desde los años 70). (Para el lector interesado, el capítulo tercero del libro de S. Rässler (2002) ofrece una breve historia sobre el enlace de encuestas hasta 2002).

A continuación revisaremos los principales métodos dentro de las aproximaciones micro y macro. Asimismo, dada su relevancia para la estadística oficial, revisaremos el tratamiento de encuestas con diseños muestrales complejos.

Métodos micro

En la aproximación micro el objetivo es generar un fichero sintético con información completa para todas las variables específicas consideradas de interés, para todos los registros. A. Leulescu y M. Agafitei (2013) destacan cuatro grandes grupos de métodos micro, que pasamos a describir a continuación.

Métodos hot-deck

A lo largo de la trayectoria del enlace de encuestas, la familia de métodos hot-deck ha sido la más utilizada con diferencia. Se trata de un conjunto de métodos no-paramétricos, es decir, que no presuponen ninguna distribución estadística inicial para las variables.

El procedimiento es el siguiente: para cada registro en uno de los ficheros (denominado *fichero receptor*), se buscan uno o varios registros en el otro fichero (denominado *fichero donante*), lo más parecidos en cuanto a las variables comunes (la edad, el sexo, el nivel de estudios…). Los valores referidos al registro donante hallado se *imputan* en el registro receptor (véase la Figura 2).

La característica principal del procedimiento hot-deck es que los valores imputados son siempre valores reales, es decir, corresponden a valores realmente observados, y recogidos en el fichero donante.

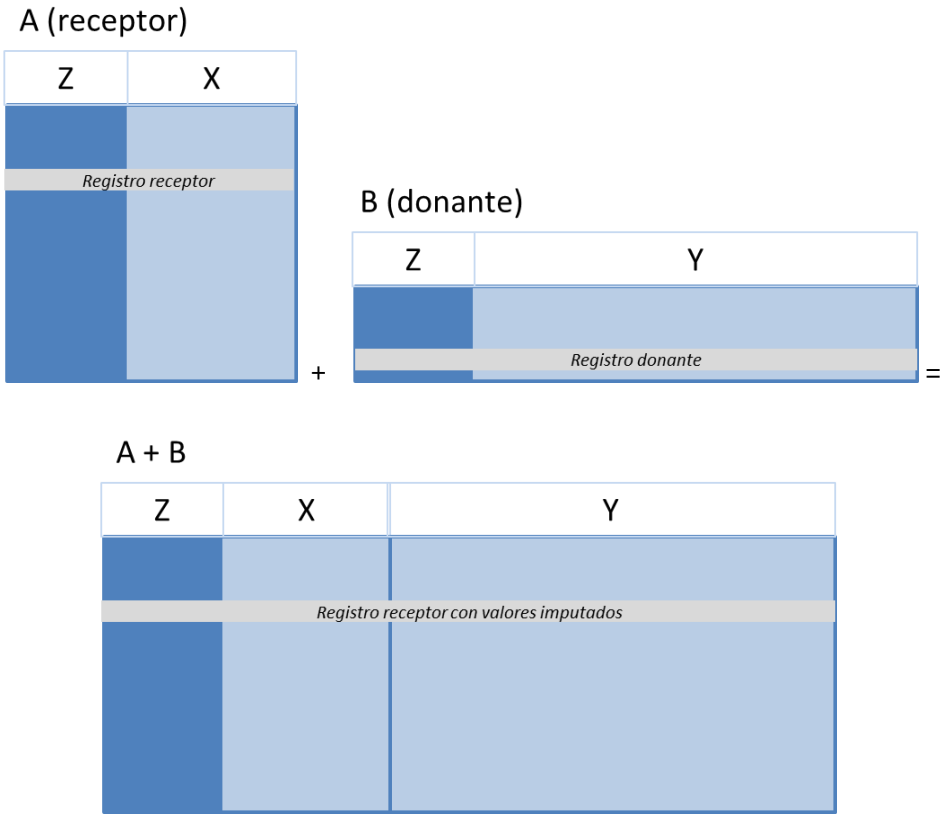

Figura 2. *Idea esquemática de la imputación por hot-deck. Para cadaregistro receptor se busca el donante que más se le parece, y se le "imputa" el valor correspondiente.*

A menudo, y dependiendo de los datos, no resulta posible hallar un donante idéntico en todas las variables para cada receptor. Por ello lo habitual es definir una *distancia* en base a las variables comunes para poder buscar parejas de registros similares. La distancia tendrá una formulación matemática distinta en función de la naturaleza de las variables (categóricas o numéricas) y dependiendo otras consideraciones; véase M. D'Orazio et al. (2008), Anexo C.

Tal como se indica en el Cuadro 1, el algoritmo de imputación hot-deck admite diversas variantes en función de su implementación: definición de estratos, restricciones para no repetir el uso de registros donantes, etc. Parametrizando convenientemente estas variantes, un buen algoritmo hot-deck reproducirá correctamente, en el fichero receptor imputado, las distribuciones observadas en el fichero donante.

MÉTODOS HOT-DECK

Clasificación

- Método micro.
- No paramétrico.
- Frecuentista.

Algoritmo

- Para cada registro receptor en A se busca un donante en B, lo más similar en las variables comunes Z. Los valores Y del registro donante se le imputan al registro receptor.
    - Cuando no existe ningún registro donante en B idéntico al receptor se introduce una distancia para poder hallar el más similar. Existen varias funciones de distancia: la distancia de Manhattan, la Euclídea, la de Gower (para el caso mixto de variables categóricas y numéricas)...

Variantes

- Definir estratos: se emplea un hot-deck separado para determinados niveles de las variables comunes. Por ejemplo, es habitual realizar un hot-deck para cada sexo, o para cada cruce de edad y sexo: los registros donantes se buscan dentro de cada estrato. Se evitan así resultados incoherentes y el cálculo de distancias se simplifica considerablemente.

- Restringir el uso de donantes: En principio, los registros de B pueden ser utilizados más de una vez como donantes, lo que introduce el riesgo de alterar las distribuciones originales. Para evitarlo se han desarrollado métodos *hot-deck restringidos*, en los que se introduce la restricción de no emplear los donantes más de una vez.

- En lugar de emplear un solo donante para cada receptor existe la posibilidad de emplear los "*K*" registros más similares. El valor imputado en estos casos es una combinación de dichos K valores. Otra variante es tomar los registros que se encuentran a una distancia determinada "*d*".

Cálculo

- R > StatMatch > NND.hotdeck(), RANDwNND.hotdeck(), rankNND.hotdeck()

Referencias

[1] M. D'Orazio et al. (2008). Anexo C: Selección de distancias (p. 34-45).
[2] A. Leulescu et al. (2013).

Cuadro 1. *Familia de métodos hot-deck.*

Al enlazar encuestas independientes vía hot-deck, tal como se ha demostrado repetidas veces en la literatura, es importante tener en cuenta que implícitamente se estará asumiendo una situación denominada *hipótesis de independencia condicionada.*

Bajo esta hipótesis, toda la relación existente (pero no observada) entre las variables específicas X e Y viene recogida por las relaciones parciales (observadas) entre las variables (Z ,X), y (Z ,Y). Siempre que la búsqueda de parejas donante-receptor sea exacta (i.e. si para cada registro receptor el algoritmo es capaz de encontrar un donante exacto en todas las variables comunes), entonces bajo dicha hipótesis, el método hot-deck reproducirá la relación "real" (inobservada) entre las variables específicas (X e Y) de forma correcta. En esta situación ideal, la información proporcionada por las variables Z es suficiente para reproducir la relación entre X y Z.

Sin embargo, esta hipótesis es restrictiva y no se cumple siempre. Pongamos un ejemplo: una encuesta "A" mide los ingresos individuales y otra encuesta "B" registra la situación laboral ("parado", "ocupado" o "inactivo"). Se emplea la variable edad para enlazar estas encuestas vía hot-deck, con lo que el fichero A (receptor) se rellena con la variable "situación laboral" de B (donante). Entonces, en el fichero A imputado, la distribución de ingresos condicionada a una determinada situación laboral (por ejemplo, inactivos) solamente reflejará su dependencia con la edad. Dicho de otro modo: el fichero imputado de esta forma (i.e. usando solamente la variable edad) no reflejará toda la dependencia de los ingresos con la situación laboral de inactivo, más que en la medida en que esté determinada por la edad.

En la práctica hay que tratar de acercarse a esta hipótesis, utilizando para ello toda la información disponible. Como veremos a lo largo de este cuaderno técnico, el mayor reto del enlace de encuestas es buscar la mejor estrategia para acercarnos a la situación ideal de independencia condicionada.

Métodos basados en la regresión

Al contrario que los hot-deck, los métodos basados en la regresión lineal son puramente paramétricos, es decir, asumen un modelo estadístico específico para las variables. En este caso la hipótesis de independencia condicionada se traduce en el hecho de que la función de distribución conjunta es el producto de las funciones de distribución marginales, esto es:

$$f(x,y,z) = f(x|z) \times f(y|z) \times f(z)$$

(En otras palabras: se asume que los datos en los ficheros parciales son suficientes para construir un fichero completo con todas las variables).

La imputación por regresión emplea el modelo de regresión para obtener valores predichos para las observaciones "faltantes". De este modo, no se imputa un valor real (observado), sino uno estimado en base a la información común. Pero el procedimiento simple desventajas: sensibilidad ante una inadecuación del modelo especificado,

regresión hacia la media y subestimación de la varianza (por tomar valores situados en la recta de regresión).

Una solución es *la imputación por regresión estocástica*, la cual consiste en introducir un valor aleatorio residual para reflejar más adecuadamente la varianza. Este tipo de variantes, como seguidamente veremos, sirven de base para desarrollar métodos mixtos, más sofisticados.

Métodos mixtos

Los métodos mixtos surgen al combinar las ventajas de las dos aproximaciones anteriores: los métodos no-paramétricos de la familia hot-deck, que son robustos ya que no especifican ningún modelo explícito a priori; y los métodos paramétricos basados en la regresión, más parsimoniosos puesto que no dependen críticamente de las variables elegidas para calcular distancias.

Dentro de esta familia hay que destacar el método de imputación conocido como *predictive mean matching,* introducido por Rubin (1986). En este procedimiento los valores "faltantes" en el fichero receptor se imputan en base a valores predichos por una regresión. Más concretamente: primero se calcula una regresión de X sobre Z en el fichero donante B. Con esta ecuación, se calcula un valor predicho intermedio, $\hat{X}$, en el fichero receptor A. A continuación se emplea un método hot-deck basado en una distancia $d(X,\hat{X})$ para buscar parejas donante-receptor. Finalmente, se imputan los valores observados. En resumen, el *predictive mean matching* consiste en una imputación hot-deck, pero basándose en valores intermedios obtenidos por modelos de regresión.

Otro método mixto que merece ser destacado es el *propensity score* (S. Rässler (2002), A. Leulescu et al. (2013)). Ambos ficheros, donante y receptor, se "extienden" con una variable adicional que toma valor 1 para todos los registros del fichero A (donante) y el valor 0 para todos los registros del fichero B (receptor). Juntando todos los registros en un único fichero, se estima un modelo logit o probit tomando como variable dependiente la variable adicional añadida, y como independientes las variables comunes entre los ficheros. El propensity score se define como la probabilidad condicionada estimada de una unidad de pertenecer a uno de los ficheros. El enlace finalmente se realiza seleccionando los donantes más parecidos en base a los valores del propensity score.

Métodos basados en la imputación múltiple

La *imputación múltiple* fue introducida por Rubin en la década de 1970 en el campo de los valores faltantes y se ha utilizado a menudo en el contexto del enlace de encuestas. La idea es extraer *m>1* valores plausibles para cada valor faltante (no observado) en lugar de un solo valor, reflejando así la incertidumbre acerca de dicho valor. Con los *m* valores extraídos se efectúa un *pooling* (o combinación) para dar lugar a un único valor, así como una estimación de la incertidumbre (o *varianza intra-imputación*) asociada al mismo (A. Leulescu et al., 2013).

En el contexto del enlace de encuestas, la imputación múltiple ha sido generalmente empleada para obtener ficheros de datos completos. Pero la imputación múltiple se puede emplear en contextos más complejos. Tal es el caso de la imputación múltiple por regresión secuencial (*sequential regression multiple imputation*), en la que se utilizan modelos independientes para imputar cada una de las variables a lo largo de una serie de iteraciones.

Aunque estos métodos avanzados pueden ser muy costosos computacionalmente, potencialmente ofrecen una gran flexibilidad. Por ejemplo, sería posible imputar muchas variables a la vez. Hoy día es posible emplearlos con relativa facilidad gracias a que se encuentran implementados en paquetes del software libre R tales como `mice`. (Para más información consúltese el apartado tercero: Software).

Métodos macro

En los métodos macro el objetivo es obtener una estimación directa de algún parámetro de interés relacionado con las variables específicas X e Y. Para ilustrar este tipo de procedimiento vamos a posicionarnos en la situación hipotética en la que tuviéramos dos ficheros separados referidos a una misma población:

- A: conteniendo una muestra aleatoria simple con n_A observaciones de las variables Z y X
- B: conteniendo una muestra aleatoria simple con n_B observaciones de las variables Z e Y

En el caso más simplificado, vamos a suponer que la distribución trivariante (inobservada) (X,Y,Z) es una distribución normal con parámetros:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_Z \\ \mu_X \\ \mu_Y \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_Z^2 & \sigma_{ZX} & \sigma_{ZY} \\ \sigma_{XZ} & \sigma_X^2 & \sigma_{XY} \\ \sigma_{YZ} & \sigma_{YX} & \sigma_Y^2 \end{pmatrix}$$

A fin de caracterizar la distribución conjunta podemos emplear el fichero A para estimar σ_{ZX}y el fichero B para estimar σ_{ZY}. Para estimar σ_Z^2, podemos utilizar los registros de A o los de B; o, posiblemente mejor, los $n_A + n_B$ registros del fichero concatenado, $A \cup B$.

Si el objetivo es obtener una estimación para σ_{XY} –y en ausencia de un fichero adicional, C, con observaciones de la distribución conjunta–, es necesario introducir alguna

hipótesis adicional, tal como puede ser la hipótesis de independencia condicionada. Bajo esta hipótesis, para calcular σ_{XY} serían suficientes las covarianzas parciales:

$$\sigma_{XY} = \frac{\sigma_{ZX}\,\sigma_{ZY}}{\sigma_X^2}$$

En el paquete `StatMatch` los métodos macro se hallan implementados en las funciones `mixed.mtc()` y `comb.samples()`. Esta última permite tener en cuenta el diseño muestral en el caso de muestras complejas (véase el siguiente apartado).

Métodos específicos para diseños muestrales complejos

Frecuentemente, los ficheros A y B a enlazar corresponden a muestreos complejos, es decir, provienen de encuestas que no efectúan un muestreo aleatorio simple de unidades de la población.

Existen varias formas de introducir el diseño muestral en los procedimientos de enlace de encuestas. Aquí destacamos el **procedimiento de Renssen**, implementado en la función `comb.samples()` del paquete `StatMatch` de R. Este procedimiento consiste en una serie de calibraciones sucesivas de los pesos asociados a los registros de cada uno de los ficheros, y que reflejan el diseño muestral. (En la terminología de muestreos complejas, la calibración es básicamente el recálculo de los pesos, de tal manera que se obtienen valores lo más similares posible a los teóricos del diseño, pero cumpliendo una serie de condiciones, por ejemplo, reproducir en las principales variables los totales de referencia conocidos sobre la población).

El procedimiento de Renssen generalmente se emplea con variables categóricas y con un objetivo macro (i.e. estimar la tabla de contingencia $X \times Y$, de las variables no observadas conjuntamente). Durante todo el procedimiento, que consiste en dos fases, los ficheros A, B y el auxiliar C (en su caso) se mantienen separados.

En la primera fase, los pesos en A y en B (w_A y w_B, respectivamente) se recalculan para obtener los totales para las variables comunes X (bien conocidos, o estimados por medio de los mismos ficheros A y B) . En la segunda fase se consideran dos casos:

- Si se tiene un fichero auxiliar C con información completa, los pesos en éste, w_C, se calibran para alinearse con los totales de A y B (tras sus respectivas calibraciones en el paso 1), tras lo cual se calcula una estimación de $X \times Y$
- Si no se tiene C se emplea la hipótesis de independencia condicionada para obtener una estimación

Para el lector interesado, la referencia "*Old and new approaches in statistical matching when samples are drawn with complex survey designs*" (D'Orazio et al., 2010) compara procedimiento de calibración de Renssen con otros métodos similares.

Fases de un enlace de encuestas

Independientemente del método de adoptado , el enlace de encuestas implica seguir una serie de fases estrechamente relacionadas con el desarrollo de una encuesta muestral. Es importante tener presente que el método de enlace es solamente una de estas fases, y a menudo, no la más importante (A. Leulescu et al., 2013).

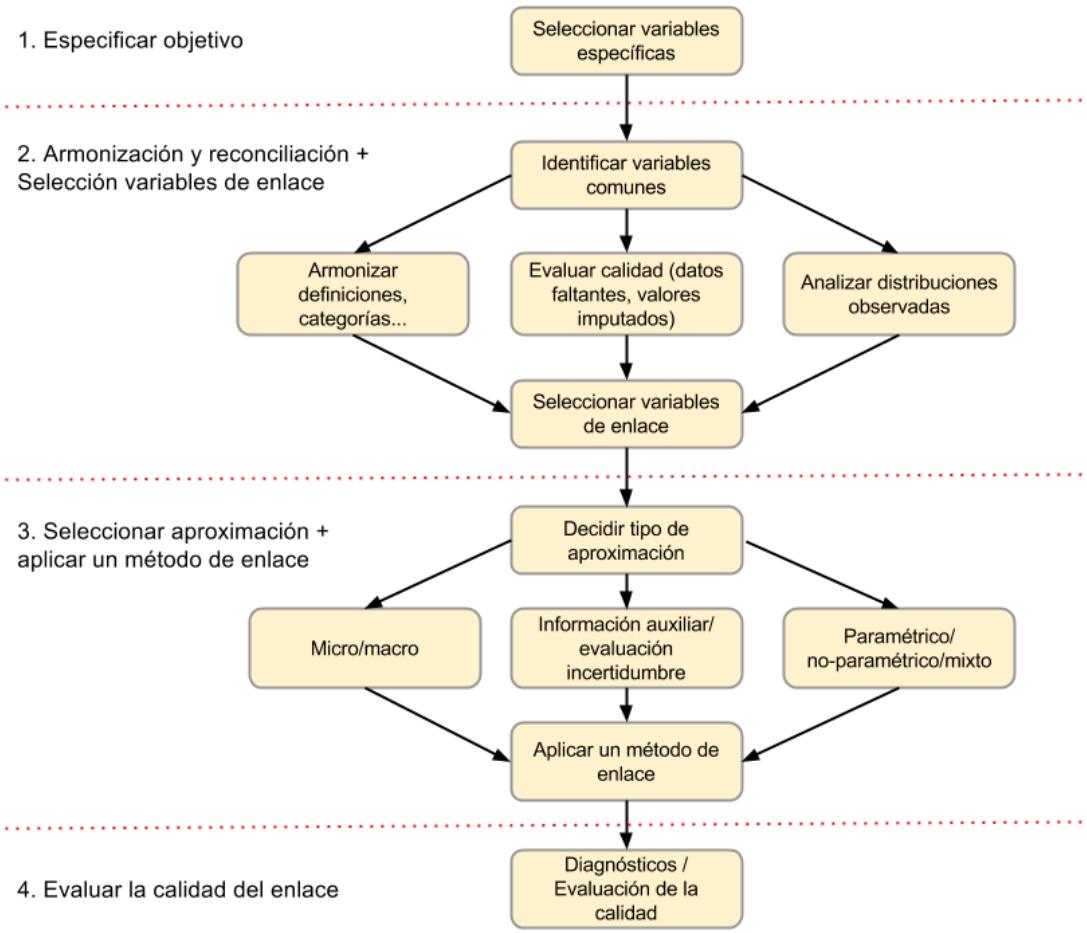

Figura 3. *Fases de un enlace de encuestas. "Report WP2 ESS-net, Statistical Methodology Project on Integration of Surveys and Administrative Data", pág. 34. Adaptado para este cuaderno.*

En la **primera fase** es necesario fijar el objetivo del enlace: (i) hay que concretar si éste va a ser micro o macro; y (ii) fijar de antemano las variables específicas a enlazar. Estas decisiones son críticas puesto que determinarán todas las fases posteriores.

La **segunda fase** requiere abordar dos tareas principales. Primero, se debe estudiar la coherencia entre los datos de las dos muestras, lo que supone realizar un análisis

exhaustivo del grado de armonización y de reconciliación entre las fuentes, incluyendo, como mínimo, el estudio de los siguientes aspectos (M. D'Orazio et al., 2008):

- La concordancia en la definición de unidades y en el periodo de referencia.
- La concordancia de las variables o medidas comunes y sus clasificaciones (en el caso de variables categóricas)
- La no-respuesta total y parcial: tratamiento de los valores faltantes o *missing*
- El cálculo de errores (sesgo y precisión de las muestras)
- El (posible) procesamiento de las variables originales en indicadores sintéticos

Habitualmente surgirán discrepancias en uno o varios de estos puntos, principalmente en lo que se refiere a la recolección de datos (variables medidas con categorías distintas…). También es esperable que el tratamiento posterior a la recolección (calibraciones de pesos, cálculo de indicadores sintéticos) sea distinto en cada muestra. Teniendo en cuenta todos estos aspectos, tras identificar una lista inicial de variables equivalentes o con una definición comparable entre las dos fuentes, se realizará un estudio comparativo de las distribuciones empíricas observadas en los dos ficheros, descartando definitivamente aquellas variables que no se puedan armonizar.

Seguidamente se debe evaluar hasta qué punto las variables comunes candidatas seleccionadas aportan información relevante para poder predecir variables específicas objetivo. Pongamos un ejemplo concreto: si quisiéramos relacionar la autopercepción de la salud (medida en la primera encuesta) con el nivel de ingresos (medida en la segunda), tendríamos que analizar qué variables comunes candidatas (la edad, el sexo, el nivel de estudios…) están estrechamente relacionadas (i.e. son predictivas) con dichas variables a la vez.

Idealmente, se debería seleccionar una lista de variables comunes con un *alto grado de concordancia* entre las muestras, tales *que sean predictivas* para las variables específicas de nuestro objetivo. Adicionalmente, es recomendable no introducir variables redundantes en la selección (por ejemplo: la edad categorizada y sin categorizar).

En ocasiones, a fin de hacer uso de toda la información disponible será conveniente generar variables derivadas de las originales.

El éxito del enlace dependerá en gran medida de la calidad en la selección de variables comunes. Por ello, esta fase resulta especialmente crítica. Existen varias herramientas estadísticas que pueden servir de ayuda para la selección óptima de las variables de enlace: véase un listado en el Cuadro 2.

HERRAMIENTAS PARA LA SELECCIÓN DE VARIABLES

- Z: variable común
- X: variable específica fichero A
- Y: variable específica fichero B

1. Evaluación del grado de concordancia

a) Variables Z categóricas

- Índice de similitud
- Índice de superposición
- Coeficiente de Bhattacharyya
- Distancia de Hellinger
- Análisis gráfico: gráficos de barras, sectores…

Cálculo: R > StatMatch > comp.prop()

b) Variables Z numéricas

- Estadísticos descriptivos: mínimo, máximo, media, desviación estándar, coeficiente de variación, percentiles
- Análisis gráfico: qqplots, histogramas, funciones de densidad estimadas superpuestas

Cálculo: Múltiples paquetes de R, por ejemplo: Hmisc > describe()

2. Evaluación del valor predictivo (relevancia)

a) X o Y continua o categórica ordinal, Z continua o categórica ordinal

- Coeficiente de correlación de Spearman corregido

b) X o Y continua o categórica ordinal, Z categórica nominal

- Coeficiente de determinación correspondiente al Eta-cuadrado (relacionado con el test de Kruskal-Wallis)

c) X o Y categórica nominal, Z categórica nominal u ordinal

- Medidas de asociación basadas en el estadístico Chi-cuadrado tales como la V de Cramer
- Medidas basadas en la reducción de la varianza (reducción proporcionan de la varianza) o la disminución de la entropía ()

Cálculo: R > Hmisc > spearman2() o también StatMatch > pw.assoc()

3. Evaluación de la redundancia

a. Análisis de redundancia para descartar predictores que aportan información similar

b. Métodos exploratorios basados en el clústering de variables

Cálculo: R > Hmisc > redun(), varclus()

4. Métodos multivariantes

a. Métodos estadísticos genéricos: análisis de regresión para variables continuas, árboles de clasificación y regresión (CART) y *random forests* para tipologías mixtas. A utilizar con cautela.

Cálculo: R > randomForest

b. Métodos específicos para el enlace de encuestas: selección de variables que más reducen la incertidumbre en la estimación de parámetros de la distribución conjunta.

Cuando las variables son categóricas se pueden emplear las *bandas de Frèchet*.

Cálculo: R > StatMatch > Fbwidths.by.x()

Referencias

A. Leulescu et al. (2013)
A. Agresti (2014).

Cuadro 2. *Herramientas para la selección de variables.*

En la **tercera fase** se debe elegir un método de enlace apropiado para obtener bien un fichero sintético fusionado (caso micro), bien una estimación de algún parámetro estadístico de interés (caso macro). (Estos métodos se han revisado en el apartado anterior, "Aproximaciones y métodos"). La elección del método dependerá del tipo de información disponible: por ejemplo, si se puede disponer de un fichero auxiliar, C, con información completa (por ejemplo, de una encuesta realizada años atrás sobre una población similar), esta información se puede emplear para mejorar el resultado del enlace.

La **cuarta y última fase** consiste en validar los resultados para garantizar la aplicabilidad del fichero fusionado. Dada su importancia, esta fase se discute con detalle en el siguiente apartado.

Validez de los resultados

A la hora de valorar la validez de un enlace de encuestas (en el sentido de grado de aplicabilidad[3]; última fase de la Figura 2) debemos tener en cuenta todas las fases del proceso, siendo especialmente críticas: la calidad y coherencia de las fuentes originales, los supuestos realizados acerca de las distribuciones condicionadas (i.e. asumir la hipótesis de independencia condicionada), y el método de enlace en sí.

S. Rässler (2002) estableció cuatro niveles para poder evaluar la validez de un enlace de forma sistemática. En adelante, y para facilitar la exposición de estos conceptos, vamos a suponer que hemos obtenido un fichero fusionado (aproximación micro). Estas consideraciones se aplicarían de forma análoga a la aproximación macro.

Niveles de validez

Nivel 1: Preservar valores individuales

Tras el enlace los valores reales (no-observados) de las variables Y se reproducen de manera exacta en el fichero recipiente: $y_i = \hat{y}_i$ para $i = 1,2,\dots,n_A$, donde y_i son los valores reales (inobservados), e $\hat{y}_i$ son los imputados, siendo n_A el tamaño del fichero recipiente A. Se busca calcular el número de veces que el valor imputado coincide con el "real" para poder calcular una "tasa de aciertos".

Nivel 2: Preservar distribuciones conjuntas

Tras el enlace, la distribución conjunta real (inobservada) de los tres conjuntos de variables X, Y, Z se refleja correctamente en el fichero imputado. Esto es: $f(X,Y,Z) = \hat{f}(X,Y,Z)$, donde f denota la distribución conjunta observada y $\hat{f}$ la obtenida en sobre el fichero imputado.

Nivel 3: Presevar la estructura de correlaciones

Tras el enlace, el fichero fusionado preserva la estructura de correlaciones y momentos de orden superior. Esto es: $cov(X,Y,Z) = \widehat{cov}(X,Y,Z)$, donde cov denota la matriz de varianzas-covarianzas observada, y $\widehat{cov}$ la calculada sobre el fichero fusionado.

Nivel 4: Preservar las distribuciones marginales

Tras el enlace, las distribuciones conjuntas marginales observadas en el fichero donante se reproducen correctamente en el fichero imputado. En particular, se cumplen: $f(Y) = \hat{f}(Y)$ y $f(Y|Z) = \hat{f}(Y|Z)$, donde f denota las distribuciones marginales observadas y real y $\hat{f}$ las imputadas.

[3] Se habla aquí de *validez* y no de *eficiencia*: no se emplea aquí ningún criterio similar al del mínimo error cuadrático medio tal como es utilizado en otros campos de la estadística; más bien, aquí el objetivo es evaluar los diferentes niveles de reproducción y preservación de las distribuciones y asociaciones originales.

Discusión

Dado que en general no se dispondrá de los valores "reales" de Y en el fichero imputado (de lo contrario no tendría sentido plantear un enlace), el **primer nivel** de validez no se estudia en general. De hecho, este nivel sólo cobra sentido en el marco de un estudio de simulación: un fichero A se divide en dos artificialmente, tal que en uno de ellos se eliminan una serie de "variables Y" y se mantienen las Z. Después, las variables eliminadas se "recuperan" vía enlace, y se calcula una "tasa de acierto".

En un caso real (i.e. no-simulado), generalmente se querrá obtener un fichero sintético válido para poder efectuar análisis estadísticos combinados, y en este sentido son más relevantes los niveles segundo, tercero y cuarto.

Los **niveles segundo** (preservar la distribución conjunta) **y tercero** (preservar la estructura de correlación o, más en general, los momentos de orden superior) garantizan la validez del fichero sintético en cuanto a su capacidad para reflejar estadísticas adecuadas sobre la relación real de las variables en la población. Pero estos niveles tampoco son directamente contrastables, por la misma razón: por definición se desconoce la forma de la distribución $f(X,Y,Z)$ o de la estructura de correlación.

De hecho, en la práctica, solamente el **cuarto nivel** se podrá contrastar directamente. El cumplimiento de este nivel garantiza que las distribuciones marginales observadas en el fichero donante se reproducen correctamente en el fichero receptor. Siempre que se empleen procedimientos robustos y datos de calidad, es relativamente sencillo alcanzar este nivel; por ello, es el requisito mínimo que se le puede pedir a cualquier ejercicio de enlace (A. Leulescu et al., 2013).

Como ya se ha indicado, la sola creación de un fichero sintético que cumpla con el cuarto nivel de validez no implica automáticamente que éste vaya a reflejar correctamente las relaciones entre las variables no-observadas conjuntamente. Por ello es necesario realizar un esfuerzo adicional, que dependerá de si se dispone de información auxiliar. Si éste es el caso, la información adicional deberá integrarse en el enlace para aumentar la validez del fichero sintético. En caso contrario, será conveniente llevar a cabo un análisis de incertidumbre.

Dada la relevancia de la información auxiliar a la hora de mejorar los resultados de un enlace, seguidamente le dedicaremos un apartado específico.

Integración de la información auxiliar

Se ha demostrado que el uso de información auxiliar puede mejorar considerablemente los resultados de un enlace. Por ejemplo en M. D'Orazio et al. (2008), los deciles mensuales de ingresos netos se emplean para mejorar ciertas estimaciones más detalladas de ingresos y gastos.

La información auxiliar puede provenir de distintas fuentes:

a. Un fichero de datos adicional, C, con observaciones de (X,Y,Z), posiblemente de años anteriores, o de otras fuentes independientes

b. Información paramétrica auxiliar en forma de una estimación externa independiente

c. Información a priori sobre el fenómeno estudiado. (El caso típico es el de las restricciones lógicas sobre los valores que pueden tomar las variables.)

En `StatMatch` existen varias funciones que permiten introducir información auxiliar. Por ejemplo, la función `comb.samples()` está diseñada para calcular tablas de contingencia para variables categóricas observadas en ficheros separados. La información de un fichero auxiliar, C, con observaciones conjuntas se puede introducir mediante el parámetro `svy.C` para mejorar las estimaciones.

Otra función del mismo paquete es `mixed.mtc()` la cual admite un parámetro `rho.yz`[4] para poder proporcionar una estimación a priori (externa) de la correlación entre las variables no observadas conjuntamente.

Análisis de la incertidumbre

En ausencia de información auxiliar, es aconsejable emplear métodos para estimar la incertidumbre de un enlace.

En el contexto que nos ocupa, la palabra *incertidumbre* hace referencia a la indeterminación debida a que existe un rango posible de valores compatible con los datos observados, para aquellas relaciones entre variables no observadas conjuntamente (lo que también se conoce por *falta de identificación*).

Como hemos visto anteriormente, cuanto más estrecha sea la relación entre las variables específicas y las comunes en cada uno de los ficheros, menor será la incertidumbre tras el enlace. Existen varias alternativas para evaluar la incertidumbre tras un enlace:

- En el caso de variables categóricas es posible calcular las *bandas de Frèchet*, que proporcionan cotas inferiores y superiores para las celdas en tablas de contingencia. El intervalo entre las cotas contiene todos los valores compatibles con los datos observados. En `StatMatch`, este cálculo está implementado en la función `Frechet.bounds.cat()`

- La imputación múltiple es el contexto natural para evaluar la incertidumbre. En efecto, mediante esta herramienta, y especialmente en un contexto Bayesiano[5], es posible analizar la sensibilidad de los resultados hacia diferentes hipótesis iniciales

[4] La notación de este cuaderno no coincide con la de StatMatch: Z denota la variable común; X e Y, las específicas.
[5] El libro S. Rässler (2002) trata extensamente este tema.

acerca de las relaciones condicionadas a Z. Esta vía se puede explorar con ayuda del paquete `mice` (Buuren, S. y Groothuis-Oudshoorn, K., 2011).

Capítulo

3

Software

Muchos de los métodos presentados en este cuaderno se hallan implementados a lo largo de diversos paquetes del entorno de computación estadística R. Algunos de estos paquetes están orientados al mundo de la estadística oficial (caso de `StatMatch`) o al mundo del análisis de datos encuestas (caso de `survey`); otros (tales como `Hmisc`) son genéricos y ofrecen múltiples funciones para el análisis de datos.

El Cuadro 3 recoge una lista no exhaustiva de paquetes disponibles en R con diversas funciones para el enlace de encuestas independientes.

PAQUETES DE R PARA EL ENLACE DE ENCUESTAS INDEPENDIENTES

- `StatMatch`: por Marcello d'Orazio (ISTAT), surgido en parte como resultado de dos proyectos sobre integración de datos llevados a cabo dentro sistema Estadístico Europeo (European Statistical System), véase ESSnet[6]. Este paquete está específicamente orientado al enlace e imputación de datos de encuestas independientes. Proporciona funciones cubriendo varias de las fases del enlace de encuestas, principalmente:
  - métodos no-paramétricos de imputación hot-deck,
  - métodos mixtos basados en predictive mean matching,
  - métodos para tratar con muestras complejas,
  - métodos para explorar la incertidumbre en el contexto de un enlace.

- `survey`: por Thomas Lumley. Contiene una amplia variedad de herramientas para analizar datos de muestras complejas: estadísticos descriptivos, tests, modelos lineales generalizados, modelos de Cox, análisis factoriales y de componentes principales, etc.

- `Hmisc`: *Harrell Miscellaneous*, por Frank E Harrell Jr., con contribuciones de Charles Dupont. Contiene funciones cubriendo diversos aspectos del análisis de datos: gráficos avanzados, creación de tablas, *clustering* de variables, manipulación de vectores de caracteres, recodificación de variables.

- `mice`: *Multiple Imputations via Chained Equations*, por Stef van Buuren. Implementa la imputación múltiple basada en la especificación condicional completa (FCS, *Fully Conditional Specification*) implementada por el algoritmo MICE. La idea es que a cada variable se le asigna su propio modelo de imputación. El paquete proporciona modelos para variables continuas (*predictive mean matching*, normal), variables dicotómicas (regresión logística), variables categóricas sin orden (regresión logística multinomial) y variables categóricas ordinales (*odds* proporcionales). El paquete

[6] Los proyectos son *Data Integration* (12/2009-12/2011) e ISAD: *Integration of Survey and Administrative Data* (12/2006-06/2008), ambos liderados por ISTAT.

también proporciona gráficos de diagnóstico para inspeccionar los resultados de las imputaciones.

- `Amelia`: *Amelia II: A Program for Missing Data*, de James Honaker, Gary King y Matthew Blackwell. Contiene funciones para efectuar imputación múltiple de encuestas implementado en un algoritmo basado en la técnica *bootstrap*, creado por los mismos autores, más avanzado que otras soluciones similares, y que permite manejar varias variables a la vez. Contiene una GUI o interfaz gráfica de usuario que puede ser utilizada por usuarios que no manejan R.

- `BaBooN`: *Bayesian Bootstrap Predictive Mean Matching - Multiple and single imputation for discrete data*, de Florian Meinfelder. Contiene dos versiones del algoritmo *Bayesian Bootstrap Predictive Mean Matching* para la imputación múltiple de datos valtantes. Se recomienda emplear la segunda variante para situaciones como la del enlace de encuestas (o fusión de datos), o, en general, situaciones en las que varias variables presentan un mismo patrón de datos faltantes.

Referencias

`StatMatch`
Marcello D'Orazio (2013). `StatMatch`: Statistical Matching (aka data fusion). http://CRAN.R-project.org/package=StatMatch

D'Orazio, M. (2013). Statistical Matching and Imputation of Survey Data with StatMatch: Viñeta de StatMatch.

`survey`
Thomas Lumley (2012) `survey`: analysis of complex survey samples. http://CRAN.R-project.org/package=survey

`mice`
Stef van Buuren, Karin Groothuis-Oudshoorn (2011). `mice`: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL http://www.jstatsoft.org/v45/i03/ www.multiple-imputation.com

`Hmisc`
Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2014). `Hmisc`: Harrell Miscellaneous. R package version 3.14-3. http://CRAN.R-project.org/package=Hmisc

`Amelia`
James Honaker, Gary King, Matthew Blackwell (2011). `Amelia II`: A Program for Missing Data. Journal of Statistical Software, 45(7), 1-47. URL http://www.jstatsoft.org/v45/i07/.

`BaBooN`
Florian Meinfelder (2011). `BaBooN`: Bayesian Bootstrap Predictive Mean Matching – Multiple and single imputation for discrete data. http://CRAN.R-project.org/package=BaBooN

Cuadro 3. *Paquetes de R para enlazar encuestas independientes.*

En el capítulo 5º: "*Desarrollo de* un paquete propio R", se describe un paquete propio, `micromatch`, basado en los anteriores, que se ha desarrollado a lo largo de este mismo proyecto y que se distribuye también a través de la página web de Eustat.

Capítulo

4

Aplicación práctica

Enlace de dos encuestas muestrales de EUSTAT: la Encuesta de Condiciones de Vida y la encuesta de Población en Relación con la Actividad

El objetivo de este estudio es probar las técnicas presentadas en el capítulo dedicado a la Metodología (página 6) en el caso real de dos encuestas muestrales independientes de EUSTAT, la Encuesta de Condiciones de Vida (2009) y la Encuesta de Población en Relación con la Actividad (4º Trimestre de 2009).

Utilizando los últimos datos disponibles de estas dos encuestas, paso a paso ilustraremos las principales fases del enlace, desde la definición de los objetivos, pasando por el estudio de la coherencia entre las fuentes y la selección de variables de enlace, y terminando con una breve validación y presentación de los resultados. Todos los cálculos mostrados se han realizado con el entorno de sofware libre R.

Descripción de las encuestas

La **Encuesta de Población en Relación con la Actividad** (en adelante, PRA) es un panel trimestral continuo que EUSTAT lleva realizando desde el año 1985 con el objetivo de conocer las características así como la dinámica del mercado de trabajo en la Comunidad Autónoma de Euskadi. La PRA realiza un muestreo probabilístico de un panel de viviendas que se renueva con frecuencia trimestral (véase en Bibliografía: ficha metodológica). La muestra actual tiene un tamaño aproximado de 5.000 viviendas (lo que afecta a un total aproximado de 13.500 individuos) y una rotación de un octavo de un trimestre.

La encuesta tiene dos objetivos principales:

- Facilitar información estadística continua sobre el volumen y las características de los principales colectivos en que se puede clasificar a la población de Euskadi según su participación en las distintas actividades económicas, así como de sus cambios de situación.

- Facilitar información estadística sobre las principales características demográficas y sociales de esta población, así como su grado de participación en actividades no productivas económicamente.

Uno de los principales resultados de la PRA es una clasificación de la población en relación con su actividad más detallada que la distinción básica entre Ocupados, Inactivos y Parados (véase la Tabla 1).

SISTEMA DE CLASIFICACIÓN PRA

→ **Población con actividad**: personas que realizan actividades para la producción de bienes y servicios, que se dedican a la realización de las tareas domésticas, que cursan estudios, o que se encuentran cumpliendo el servicio militar. Se distribuyen en dos grupos:

- **Población con actividad laboral:** personas que están trabajando. La noción de trabajo designa toda actividad ejercida con remuneración o beneficio, es decir, todo trabajo remunerado en el contexto de una relación empleador-empleado o todo trabajo independiente. Puede igualmente tratarse de un trabajo familiar no remunerado (ayudas familiares).

- **Población con actividad no laboral:** personas que no tienen empleo y no se encuentran al servicio de ningún empleador/a, y se dedican a la realización de las tareas del hogar, cursan estudios o se encuentran cumpliendo el servicio militar. Se clasifican según busquen o no una actividad laboral y por su disponibilidad subjetiva para incorporarse a ella. Se distinguen dos sub-colectivos:

  - Población ocupada en actividades no laborales: personas que no buscan un empleo, y que se dedican exclusivamente a la realización de las tareas del hogar, al estudio, o al cumplimiento del servicio militar.
  - Resto: personas con actividad no-laboral que buscan empleo

→ **Población sin actividad:** se divide en dos sub-colectivos:.

- **Población parada estricta:** personas que buscan empleo y que están disponibles para ocupar inmediatamente un puesto de trabajo.

- **Población jubilada estricta y otros:** personas que por su edad o su situación física no realizan ninguna actividad (pensionistas, incapacitadas para el trabajo, etc.).

Tabla 1. *Segmentación de la población en relación al tipo de actividad, tal como se recoge en la encuesta PRA. Los datos correspondientes a las personas que han participado en el cuestionario se extrapolan a toda la población.*

La **Encuesta de Condiciones de Vida** (en adelante, ECV) es una encuesta muestral que EUSTAT viene realizando con una periodicidad quinquenal desde el año 1989 con el objetivo de proporcionar información puntual sobre las condiciones de vida familiares, individuales y del entorno de la C.A. de Euskadi.

La ECV utiliza dos tipos de cuestionario (uno individual y otro familiar), y se basa en una muestra aleatoria estratificada en dos etapas (véase en Bibliografía: ficha

metodológica). En la primera etapa se seleccionan las viviendas dentro de un estrato (zona geográfica) sobre las que se tendrá que responder al cuestionario familiar y en una segunda etapa se selecciona aleatoriamente una persona de la vivienda, que responderá al cuestionario individual. El tamaño muestral inicial es de 7.500 viviendas.

La ECV persigue tres objetivos específicos:

1. Conocer las condiciones de salud, instrucción, trabajo, tiempo libre y relaciones sociales de los individuos

2. Describir el estado del medio ambiente físico y social del entorno o zona de residencia de las personas

3. Analizar las relaciones familiares y recursos económicos de la familia, así como los equipamientos de su vivienda.

Enlace ECV-PRA

En los siguientes apartados se describe cada una de las fases del enlace entre las encuestas ECV y PRA de EUSTAT, siguiendo el esquema de la Figura 3 (apartado Fases de un enlace de encuestas).

Datos disponibles y población de referencia

A la fecha de redacción de este cuaderno técnico los últimos datos disponibles y publicados de la ECV corresponden al último trimestre de 2009. Por tanto, para llevar a cabo el enlace, se emplean acordemente los datos de la PRA correspondientes a dicho periodo.

En ambas encuestas se han seleccionado las personas de 16 o más años. Así, nuestra población de referencia son las personas de 16 o más años, residentes de la C.A. de Euskadi en el último trimestre de 2009. La muestra disponible es de 12.658 observaciones en PRA y 5.242 en ECV, que al descartar a los menores de 16 años se quedan reducidas a 10865 observaciones y 4749 observaciones, respectivamente.

Fase 1: Fijar objetivo del enlace

El objetivo de este estudio es valorar la posibilidad de proporcionar estadísticas integradas que combinen aspectos relacionados con los estilos y condiciones de vida con el mercado laboral. Para ello se parte de la información proporcionada de forma independiente por las encuestas PRA y ECV.

Más concretamente, se desea obtener un fichero sintético combinando variables de las dos encuestas. Para ello, se realiza una imputación hot-deck en la que ECV actúa como encuesta receptora y la PRA como encuesta donante. La principal variable de la PRA –la segmentación que hemos mostrado en la Tabla 1–, se imputa en el fichero de la encuesta ECV.

Así pues, a la encuesta ECV (con todos los ítems que recogen diferentes aspectos como el nivel de instrucción, el estado de salud, las relaciones sociales, el medio ambiente, la situación económica…), se le añade la principal variable de la PRA, a saber, la segmentación de la población de acuerdo con el tipo de actividad (Tabla 1). Este fichero sintético proporcionará la oportunidad de analizar las distintas condiciones de vida en función de la segmentación del mercado laboral. Esto es algo que actualmente no se puede llevar a cabo directamente, dado que las variables corresponden a encuestas (ficheros) independientes.

Para simplificar el objetivo de este estudio se han seleccionado 6 variables específicas de la ECV cubriendo las principales dimensiones recogidas por esta encuesta, véase la Tabla 2. (En el Anexo: Ficha A se incluye una tabla que proporciona el origen y tratamiento de estas variables con relación a los ficheros de microdatos.)

VARIABLES ESPECÍFICAS DE ECV (Muestra)

- Trastornos de salud: {1-Algún transtorno; 2-Ningún transtorno}
- Conocimiento de idiomas: {1-Sólo castellano; 2-Castellano y otros; 3-Castellano y euskara; 4-Castellano, euskara y otros}
- Tiempo libre: {1-Menos de 2h; 2-2h-4h; 3-Más de 4h}
- Situación económica objetiva: {1-Mala; 2-Normal; 3-Buena}
- Tenencia de vehículos a motor {1-Ninguno; 2-Uno; 3-Dos o más}
- Equipamiento del hogar[a]: {1-Equipamiento escaso; 2- Equipamiento suficiente}

[a] Niveles agregados

Tabla 2. *Muestra de variables específicas de la ECV, junto con sus categorías. Véase Anexo: Ficha A.*

Fase 2: Seleccionar variables de enlace

En esta fase el objetivo es seleccionar un subconjunto óptimo de variables (denominadas variables comunes o de enlace) entre todas aquellas variables (originales o derivadas) compartidas por las encuestas ECV y PRA.

Fase 2-1: Meta-análisis de los cuestionarios

A fin de identificar todas aquellas variables que (potencialmente) recogen la misma información, se lleva a cabo un meta-análisis de los cuestionarios. Dado que el objetivo es tratar de imputar la segmentación PRA dentro de la encuesta ECV, además de las habituales variables sociodemográficas como la edad, el sexo o el tamaño familiar, se buscan específicamente variables que aporten información relacionada con la actividad laboral. (Estas variables se conocen por variables proxy o variables "comunes-específicas"). Esto es factible en este caso, gracias a que en la ECV existe un apartado de "Condiciones de trabajo", del que se extraen algunos indicadores.

Como resultado, se han identificado las variables comunes siguientes entre ECV y PRA, 4º trimestre de 2009:

Variables sociodemográficas

- Edad
- Sexo
- Tamaño familiar

Variables relacionadas con el nivel de instrucción

- Personas realizando estudios de enseñanza reglada (en adelante "Estudiante S/N")
- Personas realizando estudios a distancia
- Analfabetos (que no saben leer o escribir)

Variables relacionadas con la relación con la actividad

- Identificación de parados/ocupados/inactivos
- Horas trabajadas de los ocupados
- Personas buscando empleo (en adelante "Buscando empleo S/N")
- Dedicación a las tareas domésticas

A fin de que el lector pueda trazar el origen de esta información y su posterior tratamiento los códigos de las variables en los ficheros de los microdatos, así como los niveles de agregación, se han incluido en el Anexo: Ficha B.

Fase 2-2: Estudio de coherencia

Tras el meta-análisis, para las variables identificadas en ambos cuestionarios se ha realizado un estudio de la coherencia en base a las distribuciones marginales observadas. Inicialmente se han descartado dos de las variables: "Analfabetos" y "Personas realizando estudios a distancia", por tener una probabilidad de ocurrencia demasiado baja. La coherencia de las distribuciones marginales se ha analizado tanto en sentido global (sin tener en cuenta otras variables) como por grupos definidos por doce estratos de definidos o cruces de las variables edad y sexo; véase la Tabla 3.

	Sexo	
Edad (años)	H: Hombres	M: Mujeres
16-24	H.16-24	M.16-24
25-34	H.25-34	M.25-34
35-44	H.35-44	M.35-44
45-54	H.45-54	M.45-54
55-64	H.55-64	M.55-64
65+	H.+65	M.+65

Tabla 3. *Codificación para los estratos por Edad y Sexo empleados en este estudio.*

Distinguir los estratos resulta esencial puesto que para la mayoría de las variables las implicaciones son distintas dentro de cada grupo de edad y sexo. Por ejemplo, para los estratos H.+65 y M.+65, la variable "Estudiante S/N" no aporta ninguna información y por tanto no se debería incluir en el enlace; al contrario, en los estratos H.16-24, M.16-24 esta variable es imprescindible.

Siguiendo las recomendaciones de A. Leulescu et al. (2013), para comparar las distribuciones observadas se han empleado una serie de medidas empíricas. En este estudio se ha empleado la distancia de Hellinger, que toma valores entre 0 (distribuciones iguales) y 1 (máxima disimilitud posible). Los resultados –en general y por grupos de edad y sexo– se muestran en el Anexo: Ficha C.

Fase 2-3: Estudio del valor predictivo

Para completar la selección de variables se ha estudiado el valor predictivo de las variables con respecto a las variables específicas. La idea es seleccionar aquellas variables que aporten información de valor para poder efectuar el enlace.

Al igual que el estudio de la coherencia, la capacidad predictiva de las variables comunes se ha analizado tanto en sentido global (incluyendo todas las observaciones) como dentro de doce estratos de definidos por las variables edad y sexo. Como en el caso de ECV-PRA todas las variables son categóricas, se han empleado medidas de asociación basadas en el estadístico Chi-cuadrado, como la V de Cramer. De nuevo, los resultados se muestran en el Anexo: Ficha C.

Fase 3: Aplicar un método de enlace

Finalmente se ha empleado un método hot-deck en cada estrato en base a las variables seleccionadas. A continuación se ilustra el procedimiento para uno de los estratos: los hombres de entre 25 y 34 años de edad (inclusive).

Ejemplo: Método hot-deck dentro del estrato M.25-34

Para este segmento, las **variables comunes seleccionadas** son "Búsqueda de empleo S/N" y "Dedicación a las tareas de hogar". La variable "Tamaño familiar", aunque concordante, no aporta información en este segmento ya que en casi todos los casos las familias son de un solo miembre (TF = 1). En las variables "Ocupados" e "Inactivos" no hay concordancia suficiente, y la variable "Parados" es redundante con la búsqueda de empleo.

El fichero receptor contiene 359 registros y el donante 746. Para cada registro receptor (i.e. hombre de entre 25 y 34 años, que ha contestado a la ECV) se busca un registro donante (i.e. hombre de entre 25 y 34 años, que ha contestado a la PRA) que más se le parece en las variables elegidas. El código en `R > StatMatch`, se muestra en el Cuadro 4:

EJEMPLO: IMPUTACIÓN HOT-DECK EN `R > StatMatch`

Paso 1.- Buscar parejas donante-receptor

```
out.nnd <- NND.hotdeck(data.rec = rec, data.don = don,
         dist.fun = "Gower", match.vars = c("BUSQ","DOM"),
         constrained = TRUE)
```

Paso 2.- Imputar el fichero receptor (i.e. `ecv` filtrado con las del estrato seleccionado)

```
fecv <- create.fused(data.rec  =  rec, data.don = don,
         mtc.ids = out.nnd$mtc.ids, z.vars = "PRA")
```

Donde:

- `rec`: fichero con registros filtrados de la ECV (Hombres de edad comprendida entre 25 y 34 años).
- `don`: fichero con registros filtrados de la PARA (ídem).
- `match.vars`: lista de variables comunes seleccionadas (En el ejemplo: BUSQ y DOM).
- `z.vars`: lista de variables específicas, en este caso es única, "PRA"
- `NND.hotdeck()`: Búsqueda de donantes similares
- `dist.fun = "Gower"`: se emplea la distancia de Gower. (Consúltese la documentación del paquete).
- `constrained = TRUE`: indica que el algoritmo es restringido i.e. cada registro donante se emplea una sola vez
- `create.fused()`: Función que genera el fichero ecv ampliado con los valores imputados
- `mtc.ids`: Contiene la correspondencia donante-receptor para poder generar el fichero fusionado
- z.vars: Variables imputadas (en este caso sólo una: la segmentación "PRA")

Resultado

`fecv`: Fichero inicial `rec` ampliado con la variable PRA.

Cuadro 4. *Ejemplo de imputación vía hot-deck para en enlace ECV-PRA, estrato H.25-34: (Hombres de edad comprendida entre 25 y 34 años).*

Fase 4: Evaluar la calidad de los resultados

Por último, se procede a evaluar los resultados comparando las distribuciones marginales observadas frente a las imputadas (cuarto nivel de validez). Los resultados

globales se muestran en las Figura 4-1 (resultados globales) y 4-1 (resultados por Edad y Sexo).

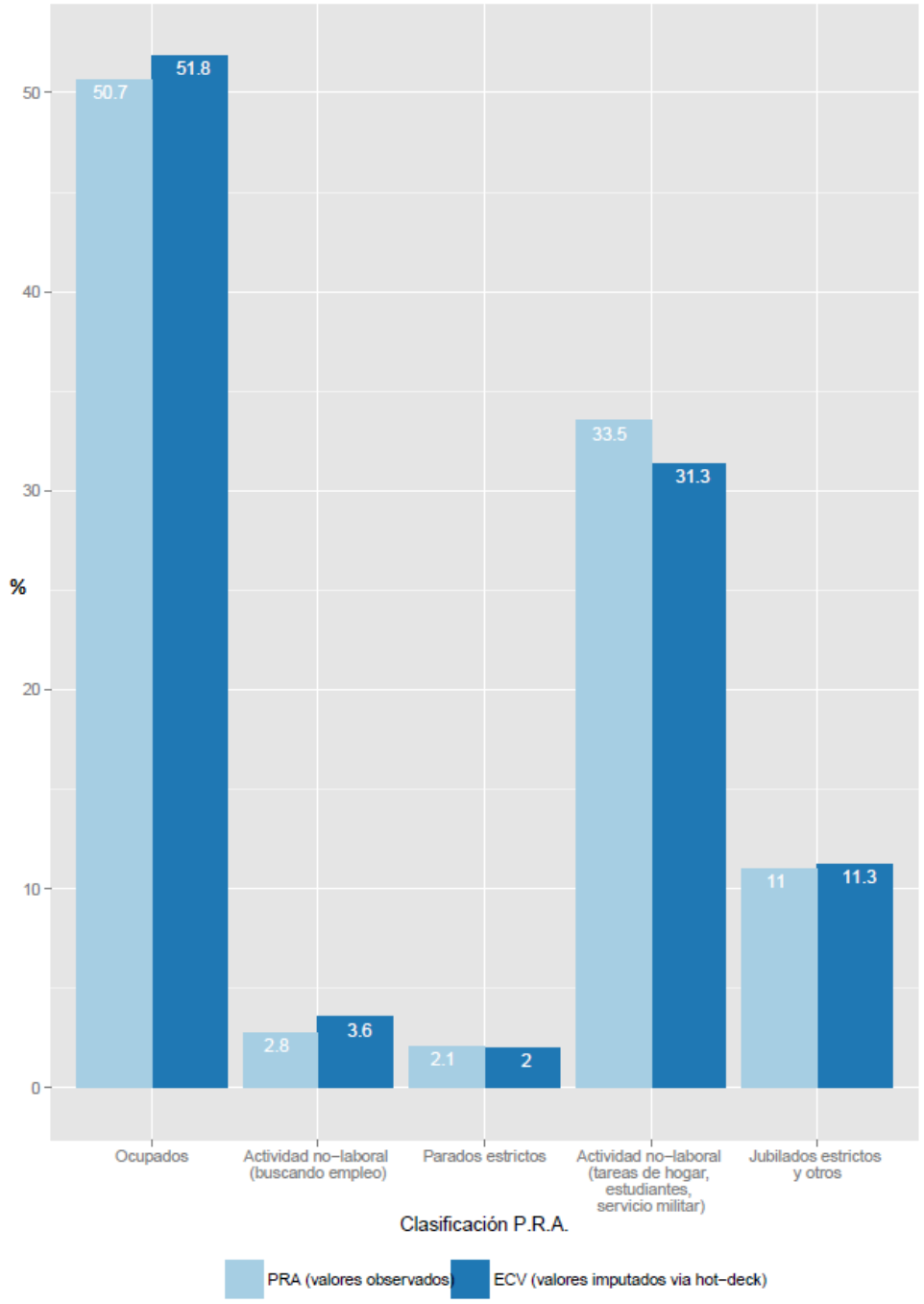

Figura 4-1. *Resultados tras el enlace ECV-PRA. Distribución marginal real observada (encuesta PRA, donante) frente a la imputada (encuesta ECV, receptora).*

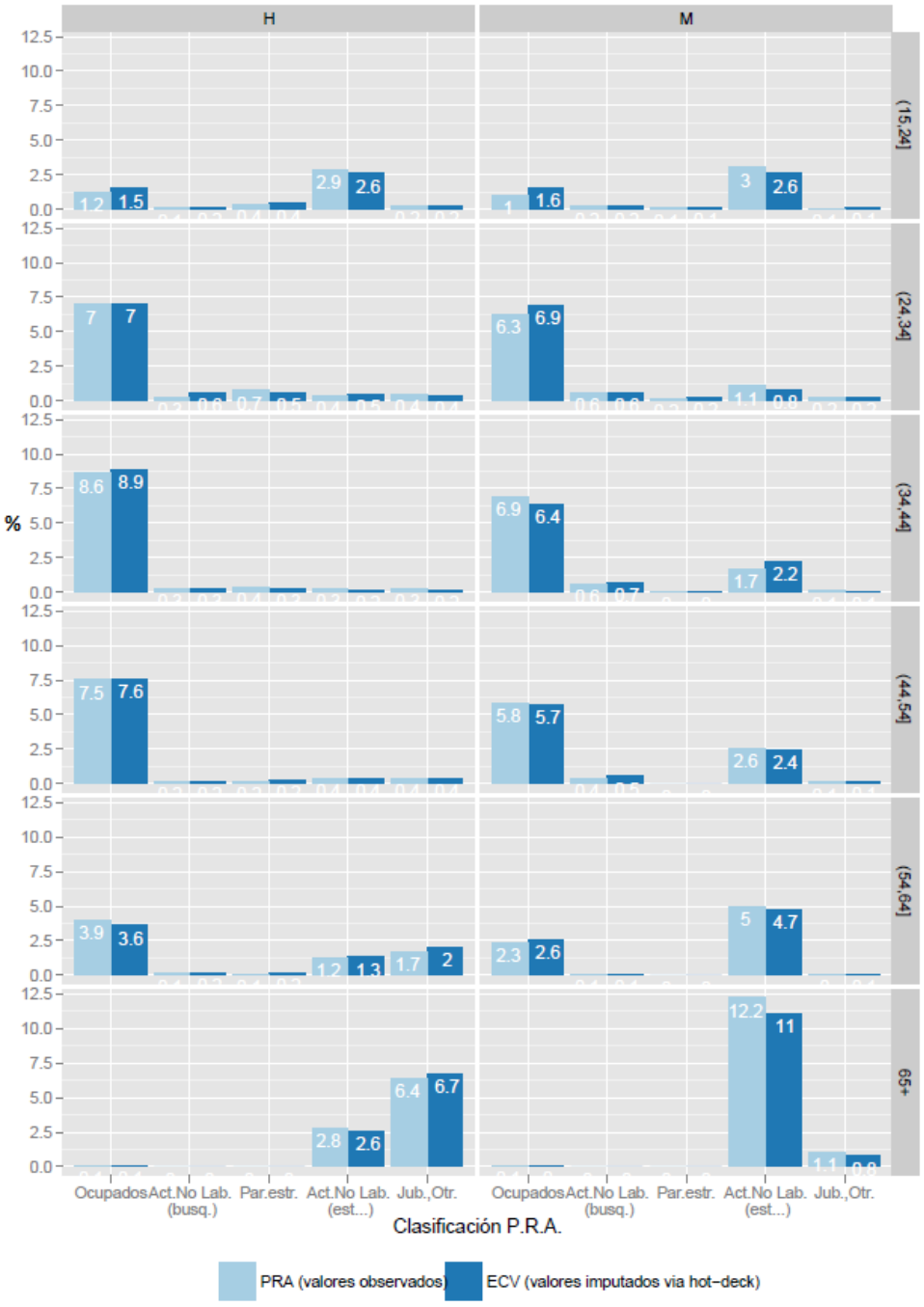

Figura 4-2. *Resultados tras el enlace ECV-PRA, por grupos de Edad y Sexo. Distribución marginal real observada (encuesta PRA, donante) frente a la imputada (encuesta ECV, receptora).*

Resultados

Para los objetivos planteados y la muestra de variables específicas seleccionadas, el resultado se traduce en una serie de tablas de contingencia que mostramos a continuación. Los resultados se presentan tanto para el total de la población, como para determinados estratos de edad o sexo que hemos seleccionado a modo de ilustración.

El interés de estas tablas reside en que permiten realizar una exploración de las condiciones de vida (proporcionadas por la ECV), en función de la segmentación del mercado laboral (proporcionada por la PRA). Estas variables inicialmente residen en ficheros separados, y, en ausencia de información adicional, sólo es posible "cruzarlas" empleando alguna técnica de enlace de encuestas. La estrategia seguida en este caso ha sido la de imputar la variable principal de la PRA (donante) dentro de la encuesta ECV (receptora), mediante una imputación hot-deck por estratos de edad y sexo.

Tablas de contingencia para explorar ítems de la Encuesta de Condiciones de Vida (origen: ECV) en función de la Relación con la Actividad (origen: PRA).

Variable: Condiciones de Salud

Fuente: PRA	**Fuente: ECV**		
	1-Algún trastorno	2-Ningún trastorno	Total
Ocupados	154.204	806.109	**960.312**
Actividad no-laboral (buscando empleo)	14.633	51.196	**65.830**
Parados estrictos	7.553	30.079	**37.633**
Actividad no-laboral (hogar,estud.,serv.mil.)	181.454	399.219	**580.673**
Jubilados estrictos y otros	86.556	121.987	**208.543**
Total	**444.400**	**1.408.590**	**1.852.991**
% fila	1-Algún trastorno	2-Ningún trastorno	Total
Ocupados	16,1%	83,9%	100%
Actividad no-laboral (buscando empleo)	22,2%	77,8%	100%
Parados estrictos	20,1%	79,9%	100%
Actividad no-laboral (hogar,estud.,serv.mil.)	31,2%	68,8%	100%
Jubilados estrictos y otros	41,5%	58,5%	100%
Total	*24,0%*	*76,0%*	*100%*
% columna	1-Algún trastorno	2-Ningún trastorno	Total
Ocupados	34,7%	57,2%	*51,8%*
Actividad no-laboral (buscando empleo)	3,3%	3,6%	*3,6%*
Parados estrictos	1,7%	2,1%	*2,0%*
Actividad no-laboral (hogar,estud.,serv.mil.)	40,8%	28,3%	*31,3%*
Jubilados estrictos y otros	19,5%	8,7%	*11,3%*
Total	*100%*	*100%*	*100%*

Resultado 1-1. *Condiciones de Salud (Fuente: ECV) vs Segmentación PRA (Fuente: PRA). Totales para toda la población, y porcentajes de filas y columnas.*

Variable: Condiciones de Salud			
Edad 45-54 años	**Fuente: ECV**		
Fuente: PRA	1-Algún trastorno	2-Ningún trastorno	Total
Ocupados	45.002	201.485	**246.487**
Actividad no-laboral (buscando empleo)	2.651	1.046	**3.698**
Parados estrictos	1.178	3.370	**4.548**
Actividad no-laboral (hogar,estud.,serv.mil.)	181.454	42.995	**224.449**
Jubilados estrictos y otros	1.467	8.534	**10.002**
Total	**231.753**	**257.430**	**489.184**
% fila	1-Algún trastorno	2-Ningún trastorno	Total
Ocupados	18,3%	81,7%	100%
Actividad no-laboral (buscando empleo)	71,7%	28,3%	100%
Parados estrictos	25,9%	74,1%	100%
Actividad no-laboral (hogar,estud.,serv.mil.)	80,8%	19,2%	100%
Jubilados estrictos y otros	14,7%	85,3%	100%
Total	*47,4%*	*52,6%*	*100%*
% columna	1-Algún trastorno	2-Ningún trastorno	Total
Ocupados	19,4%	78,3%	*19,4%*
Actividad no-laboral (buscando empleo)	1,1%	0,4%	*1,1%*
Parados estrictos	0,5%	1,3%	*0,5%*
Actividad no-laboral (hogar,estud.,serv.mil.)	78,3%	16,7%	*78,3%*
Jubilados estrictos y otros	0,6%	3,3%	*0,6%*
Total	*100%*	*100%*	*100%*

Resultado 1-2. *Condiciones de Salud (Fuente: ECV) vs Segmentación PRA (Fuente: PRA). Estrato seleccionado: Edad entre 45 y 54 años (inclusive). Totales y porcentajes de filas y columnas.*

Variable: Conocimiento de Idiomas

Fuente: PRA	Fuente: ECV				
	1-Sólo cast.	2-C.+Otr	3-C+Eusk	4-C+E+Ot.	Total
Ocupados	310.088	181.823	165.623	302.778	**960.312**
Actividad no-laboral (buscando empleo)	19.586	17.525	10.651	18.067	**65.830**
Parados estrictos	10.867	4.874	4.762	17.129	**37.633**
Actividad no-laboral (hogar,estud.,serv.mil.)	265.572	55.413	112.938	146.750	**580.673**
Jubilados estrictos y otros	113.300	27.900	46.692	20.651	**208.543**
Total	**719.413**	**287.535**	**340.666**	**505.375**	**1.852.992**
% fila	1-Sólo cast.	2-C.+Otr	3-C+Eusk	4-C+E+Ot.	Total
Ocupados	32,3%	18,9%	17,2%	31,5%	100%
Actividad no-laboral (buscando empleo)	29,8%	26,6%	16,2%	27,4%	100%
Parados estrictos	28,9%	13,0%	12,7%	45,5%	100%
Actividad no-laboral (hogar,estud.,serv.mil.)	45,7%	9,5%	19,4%	25,3%	100%
Jubilados estrictos y otros	54,3%	13,4%	22,4%	9,9%	100%
Total	*38,8%*	*15,5%*	*18,4%*	*27,3%*	*100%*
% columna	1-Sólo cast.	2-C.+Otr	3-C+Eusk	4-C+E+Ot.	Total
Ocupados	43,1%	63,2%	48,6%	59,9%	*51,8%*
Actividad no-laboral (buscando empleo)	2,7%	6,1%	3,1%	3,6%	*3,6%*
Parados estrictos	1,5%	1,7%	1,4%	3,4%	*2,0%*
Actividad no-laboral (hogar,estud.,serv.mil.)	36,9%	19,3%	33,2%	29,0%	*31,3%*
Jubilados estrictos y otros	15,7%	9,7%	13,7%	4,1%	*11,3%*
Total	*100%*	*100%*	*100%*	*100%*	*100%*

Resultado 2-1. *Conocimiento de Idiomas (Fuente: ECV) vs Segmentación PRA (Fuente: PRA). Totales para toda la población, y porcentajes de filas y columnas.*

Variable: Conocimiento de Idiomas

Edad 35-44 años	Fuente: ECV				
Fuente: PRA	1-Sólo cast.	2-C.+Otr	3-C+Eusk	4-C+E+Ot.	Total
Ocupados	76.100	59.203	50.488	96.088	**281.879**
Actividad no-laboral (buscando empleo)	7.081	4.679	4.096	3.464	**19.319**
Parados estrictos	2.536	415	321	2.311	**5.584**
Actividad no-laboral (hogar,estud.,serv.mil.)	16.426	7.238	7.056	13.743	**44.462**
Jubilados estrictos y otros	1.899	1.051	155	993	**4.099**
Total	**104.042**	**72.587**	**62.116**	**116.599**	**355.344**
% fila	1-Sólo cast.	2-C.+Otr	3-C+Eusk	4-C+E+Ot.	Total
Ocupados	27,0%	21,0%	17,9%	34,1%	100%
Actividad no-laboral (buscando empleo)	36,7%	24,2%	21,2%	17,9%	100%
Parados estrictos	45,4%	7,4%	5,8%	41,4%	100%
Actividad no-laboral (hogar,estud.,serv.mil.)	36,9%	16,3%	15,9%	30,9%	100%
Jubilados estrictos y otros	46,3%	25,6%	3,8%	24,2%	100%
Total	*29,3%*	*20,4%*	*17,5%*	*32,8%*	*100%*
% columna	1-Sólo cast.	2-C.+Otr	3-C+Eusk	4-C+E+Ot.	Total
Ocupados	73,1%	81,6%	81,3%	82,4%	*79,3%*
Actividad no-laboral (buscando empleo)	6,8%	6,4%	6,6%	3,0%	*5,4%*
Parados estrictos	2,4%	0,6%	0,5%	2,0%	*1,6%*
Actividad no-laboral (hogar,estud.,serv.mil.)	15,8%	10,0%	11,4%	11,8%	*12,5%*
Jubilados estrictos y otros	1,8%	1,4%	0,3%	0,9%	*1,2%*
Total	*100%*	*100%*	*100%*	*100%*	*100%*

Resultado 2-2. *Conocimiento de Idiomas (Fuente: ECV) vs Segmentación PRA (Fuente: PRA). Estrato seleccionado: Edad entre 35 y 44 años (inclusive). Totales y porcentajes de filas y columnas.*

Variable: Tiempo libre (horas/día)

Fuente: PRA	Fuente: ECV			
	<2h	2-4h	+4h	Total
Ocupados	171.098	551.052	238.162	**960.312**
Actividad no-laboral (buscando empleo)	8.430	28.436	28.963	**65.830**
Parados estrictos	1.161	17.341	19.131	**37.633**
Actividad no-laboral (hogar,estud.,serv.mil.)	48.494	247.203	284.977	**580.673**
Jubilados estrictos y otros	11.415	48.415	148.713	**208.543**
Total	**240.598**	**892.447**	**719.946**	**1.852.991**
% fila	<2h	2-4h	+4h	Total
Ocupados	17,8%	57,4%	24,8%	100%
Actividad no-laboral (buscando empleo)	12,8%	43,2%	44,0%	100%
Parados estrictos	3,1%	46,1%	50,8%	100%
Actividad no-laboral (hogar,estud.,serv.mil.)	8,4%	42,6%	49,1%	100%
Jubilados estrictos y otros	5,5%	23,2%	71,3%	100%
Total	*13,0%*	*48,2%*	*38,9%*	*100%*
% columna	<2h	2-4h	+4h	Total
Ocupados	71,1%	61,7%	33,1%	*51,8%*
Actividad no-laboral (buscando empleo)	3,5%	3,2%	4,0%	*3,6%*
Parados estrictos	0,5%	1,9%	2,7%	*2,0%*
Actividad no-laboral (hogar,estud.,serv.mil.)	20,2%	27,7%	39,6%	*31,3%*
Jubilados estrictos y otros	4,7%	5,4%	20,7%	*11,3%*
Total	*100%*	*100%*	*100%*	*100%*

Resultado 3-1. *Horas libres al día (Fuente: ECV) vs Segmentación PRA (Fuente: PRA). Totales para toda la población, y porcentajes de filas y columnas.*

Variable: Tiempo libre (horas/día)

Mujeres de Edad 35-44 años	Fuente: ECV			
Fuente: PRA	<2h	2-4h	+4h	Total
Ocupados	35.916	67.250	14.515	**117.680**
Actividad no-laboral (buscando empleo)	2.246	4.291	7.230	**13.767**
Parados estrictos	0	0	159	**159**
Actividad no-laboral (hogar,estud.,serv.mil.)	4.169	23.519	7.230	**34.917**
Jubilados estrictos y otros	0	0	1.027	**1.027**
Total	**42.330**	**95.060**	**30.160**	**167.550**
% fila	<2h	2-4h	+4h	Total
Ocupados	30,5%	57,1%	12,3%	100%
Actividad no-laboral (buscando empleo)	16,3%	31,2%	52,5%	100%
Parados estrictos	0,0%	0,0%	100,0%	100%
Actividad no-laboral (hogar,estud.,serv.mil.)	11,9%	67,4%	20,7%	100%
Jubilados estrictos y otros	0,0%	0,0%	100,0%	100%
Total	*25,3%*	*56,7%*	*18,0%*	*100%*
% columna	<2h	2-4h	+4h	Total
Ocupados	84,8%	70,7%	48,1%	*70,2%*
Actividad no-laboral (buscando empleo)	5,3%	4,5%	24,0%	*8,2%*
Parados estrictos	0,0%	0,0%	0,5%	*0,1%*
Actividad no-laboral (hogar,estud.,serv.mil.)	9,8%	24,7%	24,0%	*20,8%*
Jubilados estrictos y otros	0,0%	0,0%	3,4%	*0,6%*
Total	*100%*	*100%*	*100%*	*100%*

Resultado 3-2. *Tiempo libre (Fuente: ECV) vs Segmentación PRA (Fuente: PRA). Estrato seleccionado:Mujeres de Edad entre 35 y 44 años (inclusive). Totales y porcentajes de filas y columnas.*

Variable: Situación Económica Objetiva

Fuente: PRA	Fuente: ECV			
	Mala	Normal	Buena	Total
Ocupados	64.393	420.064	475.855	**960.312**
Actividad no-laboral (buscando empleo)	19.567	31.043	15.219	**65.830**
Parados estrictos	6.935	18.128	12.570	**37.633**
Actividad no-laboral (hogar,estud.,serv.mil.)	81.399	319.399	179.875	**580.673**
Jubilados estrictos y otros	40.017	112.727	55.799	**208.543**
Total	**212.311**	**901.361**	**739.318**	**1.852.991**
% fila	Mala	Normal	Buena	Total
Ocupados	6,7%	43,7%	49,6%	100%
Actividad no-laboral (buscando empleo)	29,7%	47,2%	23,1%	100%
Parados estrictos	18,4%	48,2%	33,4%	100%
Actividad no-laboral (hogar,estud.,serv.mil.)	14,0%	55,0%	31,0%	100%
Jubilados estrictos y otros	19,2%	54,1%	26,8%	100%
Total	*11,5%*	*48,6%*	*39,9%*	*100%*
% columna	Mala	Normal	Buena	Total
Ocupados	30,3%	46,6%	64,4%	*51,8%*
Actividad no-laboral (buscando empleo)	9,2%	3,4%	2,1%	*3,6%*
Parados estrictos	3,3%	2,0%	1,7%	*2,0%*
Actividad no-laboral (hogar,estud.,serv.mil.)	38,3%	35,4%	24,3%	*31,3%*
Jubilados estrictos y otros	18,8%	12,5%	7,5%	*11,3%*
Total	*100%*	*100%*	*100%*	*100%*

Resultado 4-1. *Situación económica objetiva (Fuente: ECV) vs Segmentación PRA (Fuente: PRA). Totales para toda la población, y porcentajes de filas y columnas.*

Variable: Situación Económica Objetiva

Edad 35-34 años	Fuente: ECV			
Fuente: PRA	Mala	Normal	Buena	Total
Ocupados	21.066	108.143	128.270	**257.479**
Actividad no-laboral (buscando empleo)	6.179	10.297	4.341	**20.818**
Parados estrictos	1.640	8.015	4.400	**14.054**
Actividad no-laboral (hogar,estud.,serv.mil.)	2.194	9.478	12.087	**23.760**
Jubilados estrictos y otros	1.251	6.852	3.367	**11.470**
Total	**32.330**	**142.785**	**152.465**	**327.581**
% fila	Mala	Normal	Buena	Total
Ocupados	8,2%	42,0%	49,8%	100%
Actividad no-laboral (buscando empleo)	29,7%	49,5%	20,9%	100%
Parados estrictos	11,7%	57,0%	31,3%	100%
Actividad no-laboral (hogar,estud.,serv.mil.)	9,2%	39,9%	50,9%	100%
Jubilados estrictos y otros	10,9%	59,7%	29,4%	100%
Total	*9,9%*	*43,6%*	*46,5%*	*100%*
% columna	Mala	Normal	Buena	Total
Ocupados	65,2%	75,7%	84,1%	*78,6%*
Actividad no-laboral (buscando empleo)	19,1%	7,2%	2,8%	*6,4%*
Parados estrictos	5,1%	5,6%	2,9%	*4,3%*
Actividad no-laboral (hogar,estud.,serv.mil.)	6,8%	6,6%	7,9%	*7,3%*
Jubilados estrictos y otros	3,9%	4,8%	2,2%	*3,5%*
Total	*100%*	*100%*	*100%*	*100%*

Resultado 4-2. *Situación económica objetiva (Fuente: ECV) vs Segmentación PRA (Fuente: PRA). Estrato seleccionado:Edad entre 35 y 44 años (inclusive). Totales y porcentajes de filas y columnas.*

Variable: Vehículos en propiedad

Fuente: PRA	Fuente: ECV			
	Ninguno	Uno	Dos o más	Total
Ocupados	114.756	558.465	287.091	**960.312**
Actividad no-laboral (buscando empleo)	19.094	34.008	12.728	**65.830**
Parados estrictos	5.770	19.520	12.343	**37.633**
Actividad no-laboral (hogar,estud.,serv.mil.)	211.396	274.670	94.608	**580.673**
Jubilados estrictos y otros	71.811	110.357	26.375	**208.543**
Total	**422.827**	**997.020**	**433.145**	**1.852.991**
% fila	Ninguno	Uno	Dos o más	Total
Ocupados	11,9%	58,2%	29,9%	100%
Actividad no-laboral (buscando empleo)	29,0%	51,7%	19,3%	100%
Parados estrictos	15,3%	51,9%	32,8%	100%
Actividad no-laboral (hogar,estud.,serv.mil.)	36,4%	47,3%	16,3%	100%
Jubilados estrictos y otros	34,4%	52,9%	12,6%	100%
Total	*22,8%*	*53,8%*	*23,4%*	*100%*
% columna	Ninguno	Uno	Dos o más	Total
Ocupados	27,1%	56,0%	66,3%	*51,8%*
Actividad no-laboral (buscando empleo)	4,5%	3,4%	2,9%	*3,6%*
Parados estrictos	1,4%	2,0%	2,8%	*2,0%*
Actividad no-laboral (hogar,estud.,serv.mil.)	50,0%	27,5%	21,8%	*31,3%*
Jubilados estrictos y otros	17,0%	11,1%	6,1%	*11,3%*
Total	*100%*	*100%*	*100%*	*100%*

Resultado 5-1. *Número de vehículos en propiedad (Fuente: ECV) vs Segmentación PRA (Fuente: PRA). Totales para toda la población, y porcentajes de filas y columnas.*

Variable: Vehículos en propiedad

Estrato: 55-64 años	Fuente: ECV			
Fuente: PRA	Ninguno	Uno	Dos o más	Total
Ocupados	13.009	69.510	31.889	**114.407**
Actividad no-laboral (buscando empleo)	970	2.345	2.036	**5.351**
Parados estrictos	276	2.582	449	**3.307**
Actividad no-laboral (hogar,estud.,serv.mil.)	23.395	68.120	20.161	**111.677**
Jubilados estrictos y otros	9.495	20.390	8.382	**38.267**
Total	**47.144**	**162.947**	**62.917**	**273.009**
% fila	Ninguno	Uno	Dos o más	Total
Ocupados	11,4%	60,8%	27,9%	100%
Actividad no-laboral (buscando empleo)	18,1%	43,8%	38,0%	100%
Parados estrictos	8,3%	78,1%	13,6%	100%
Actividad no-laboral (hogar,estud.,serv.mil.)	20,9%	61,0%	18,1%	100%
Jubilados estrictos y otros	24,8%	53,3%	21,9%	100%
Total	*17,3%*	*59,7%*	*23,0%*	*100%*
% columna	Ninguno	Uno	Dos o más	Total
Ocupados	27,6%	42,7%	50,7%	*41,9%*
Actividad no-laboral (buscando empleo)	2,1%	1,4%	3,2%	*2,0%*
Parados estrictos	0,6%	1,6%	0,7%	*1,2%*
Actividad no-laboral (hogar,estud.,serv.mil.)	49,6%	41,8%	32,0%	*40,9%*
Jubilados estrictos y otros	20,1%	12,5%	13,3%	*14,0%*
Total	*100%*	*100%*	*100%*	*100%*

Resultado 5-2. *Vehículos en propiedad (Fuente: ECV) vs Segmentación PRA (Fuente: PRA). Estrato seleccionado: Edad entre 55 y 64 años (inclusive). Totales y porcentajes de filas y columnas.*

Variable: Nivel de Equipamiento del Hogar

Fuente: PRA	Fuente: ECV		
	Escaso	Suficiente	Total
Ocupados	33.187	927.125	**960.312**
Actividad no-laboral (buscando empleo)	0	65.830	**65.830**
Parados estrictos	446	37.186	**37.633**
Actividad no-laboral (hogar,estud.,serv.mil.)	92.649	488.024	**580.673**
Jubilados estrictos y otros	43.476	165.067	**208.543**
Total	**169.758**	**1.683.232**	**1.852.991**
% fila	Escaso	Suficiente	Total
Ocupados	3,5%	96,5%	100%
Actividad no-laboral (buscando empleo)	0,0%	100,0%	100%
Parados estrictos	1,2%	98,8%	100%
Actividad no-laboral (hogar,estud.,serv.mil.)	16,0%	84,0%	100%
Jubilados estrictos y otros	20,8%	79,2%	100%
Total	*9,2%*	*90,8%*	*100%*
% columna	Escaso	Suficiente	Total
Ocupados	19,5%	55,1%	*51,8%*
Actividad no-laboral (buscando empleo)	0,0%	3,9%	*3,6%*
Parados estrictos	0,3%	2,2%	*2,0%*
Actividad no-laboral (hogar,estud.,serv.mil.)	54,6%	29,0%	*31,3%*
Jubilados estrictos y otros	25,6%	9,8%	*11,3%*
Total	*100%*	*100%*	*100%*

Resultado 6-1. *Nivel de equipamiento del hogar (Fuente: ECV) vs Segmentación PRA (Fuente: PRA). Totales para toda la población, y porcentajes de filas y columnas.*

Variable: Nivel de Equipamiento del Hogar

Mayores de 65 años	Fuente: ECV		
Fuente: PRA	Escaso	Suficiente	Total
Ocupados	729	1.877	**2.606**
Actividad no-laboral (buscando empleo)	0	0	**0**
Parados estrictos	0	0	**0**
Actividad no-laboral (hogar,estud.,serv.mil.)	78.202	174.850	**253.051**
Jubilados estrictos y otros	39.083	98.975	**138.058**
Total	**118.014**	**275.701**	**393.715**
% fila	Escaso	Suficiente	Total
Ocupados	28,0%	72,0%	100%
Actividad no-laboral (buscando empleo)	.	.	.
Parados estrictos	.	.	.
Actividad no-laboral (hogar,estud.,serv.mil.)	30,9%	69,1%	100%
Jubilados estrictos y otros	28,3%	71,7%	100%
Total	*30,0%*	*70,0%*	*100%*
% columna	Escaso	Suficiente	Total
Ocupados	0,6%	0,7%	*0,7%*
Actividad no-laboral (buscando empleo)	0,0%	0,0%	*0,0%*
Parados estrictos	0,0%	0,0%	*0,0%*
Actividad no-laboral (hogar,estud.,serv.mil.)	66,3%	63,4%	*64,3%*
Jubilados estrictos y otros	33,1%	35,9%	*35,1%*
Total	*100%*	*100%*	*100%*

Resultado 6-2. *Nivel de equipamiento del hogar (Fuente: ECV) vs Segmentación PRA (Fuente: PRA). Estrato seleccionado: Mayores de 65 años (. Totales y porcentajes de filas y columnas.*

Capítulo

5

Desarrollo de un paquete propio de R

A lo largo del enlace de las encuestas ECV y PRA, se crearon una serie de funciones para agilizar los cálculos en las distintas fases del proceso: selección de variables, imputación por estratos y validación de los resultados. Estas funciones, basadas en paquetes ya probados y contrastados (véase el capítulo tercero: Software), fueron englobadas en un paquete propio de R, denominado `micromatch`, con la idea de que fuesen distribuidas junto con este cuaderno técnico.

Los paquetes propios de R (i.e. asociados a un proyecto) presentan una serie de ventajas (Chambers, J., 2008):

- Articulan las funciones de R creadas para el proyecto como un todo integrado, de forma que los cálculos resultan más eficientes y fiables
- Permiten crear de forma sencilla una documentación que resulta de gran ayuda incluso para el programador
- Posibilitan vías para generalizar los cálculos a otros problemas similares o asociados
- Permiten distribuir fácilmente el código a otros potenciales usuarios

El proyecto de enlace ECV-PRA, disfrutó ya de estas ventajas.

En una fase posterior de este proyecto, se vio que cabía la posibilidad de generalizar el paquete `micromatch` para abordar de forma genérica cualquier enlace de encuestas independientes. Para poder dar este paso se diseñó una estructura de clases y métodos en el sistema S4 de *programación orientada a objetos* (Chambers, J., 2008). La idea de las **clases** es que permiten encapsular conceptos complejos en *objetos*, tales como "una encuesta a enlazar" (en `micromatch`, la clase `filetomatch`). Después, se diseñan **métodos** que actúan sobre dichos objetos, por ejemplo: "compárense todas las variables con respecto a otra encuesta a enlazar".

Al contrario que otros paquetes, el paquete `micromatch` no proporciona métodos propios de enlace de encuestas, sino que, basándose sobre paquetes ya probados y contrastados, implementa una solución genérica que cubre (y agiliza) todas las fases de un enlace.

Así, la idea de `micromatch` es crear un *entorno* en el usuario puede emplear y probar todo tipo de métodos de enlace de una forma sencilla. En definitiva, con este paquete se persigue:

- Ofrecer un entorno donde poder agilizar los cálculos que implica un enlace encuestas

- Ofrecer un entorno de computación eficaz y robusto, en la que gracias al sistema de clases todos los cálculos están eficientemente inter-relacionados

- Difundir la metodología de encuestas, mediante la implementación en un único paquete de un conjunto amplio de técnicas que actualmente residen en paquetes esparados

- Difundir el trabajo realizado durante esta beca, ofreciendo un entorno donde el usuario puede reproducir los cálculos ECV-PRA.

Este paquete fue presentado durante las VI Jornadas de Usuarios de R celebradas en Santiago de Compostela, en Octubre de 2014, también disponible en la página web de Eustat.

Capítulo

6

Conclusiones

El enlace de encuestas nos brinda la oportunidad de explotar de forma más eficiente la información recogida a través de encuestas independientes referidas a la misma población, mediante la obtención de estadísticas e indicadores integrados.

El caso práctico desarrollado en este cuaderno técnico, i.e. el enlace de dos encuestas independientes de EUSTAT: la Encuesta de Condiciones de Vida y la Encuesta de Población en Relación con la Actividad, nos ha llevado a apreciar la importancia de varios elementos en todo proceso de un enlace:

- Una selección adecuada de **variables comunes**, lo que exige realizar un meta-análisis profundo de los cuestionarios a fin de poder identificar la información comparable en cuanto a la definición y a las distribuciones empíricas observadas en los ficheros de datos.

- El empleo de **variables de estrato** (en ECV-PRA, grupos de edad y sexo), pues generalmente tendrán una estrecha relación con las variables específicas (condiciones de vida: tiempo libre, relaciones sociales… en ECV; y mercado de trabajo en PRA)

- La necesidad de **validar los resultados** exigiendo que el algoritmo produzca distribuciones marginales comparables a las observadas, pero siempre teniendo presente que puede haber más fuentes de incertidumbre implicadas

Más en general, la adopción de una metodología de enlace de encuestas resulta beneficiosa pues nos obliga a ver las encuestas no como instrumentos independientes, sino como un todo integrado. A este respecto, A. Leulescu y M. Agafitei (2013) lanzan una serie de recomendaciones:

- **Estandarizar** los cuestionarios en la medida de lo posible, formulando las preguntas de una manera comparable, para poder garantizar la consistencia entre ellas

- A ser posible, **incluir un pequeño módulo común** a todas las encuestas recogiendo determinados aspectos “específicos” pero básicos, tales como los ingresos o la salud autopercibida. A la hora de enlazar las encuestas, estas variables jugarían un papel clave en la mejora de la calidad (i.e. reducción de incertidumbre) en los resultados del enlace.

En el caso concreto del enlace ECV-PRA, en un futuro sería deseable poder contar con variables armonizadas de ingresos o el nivel de instrucción, que ayudarían a reducir la incertidumbre y, consecuentemente, a mejorar los resultados del enlace.

En cuanto a la **implementación** de estas técnicas, el desarrollo del entorno de software libre R está proporcionando cada vez más herramientas, y es posible que en un futuro veamos una mayor coordinación entre todos los paquetes. Durante este proyecto, se han dado ya algunos pasos en este sentido, gracias al desarrollo del paquete `micromatch`.

Bibliografía

Referencias generales

Agresti (2014). Categorical data analysis. John Wiley & Sons.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. Journal of statistical software, 45(3).

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). Statistical matching: Theory and practice. John Wiley & Sons.

D'Orazio, M., Di Zio, M., & Scanu, M. (2010). Old and new approaches in statistical matching when samples are drawn with complex survey designs. Proceedings of the 45th "Riunione Scientifica della Societa'Italiana di Statistica", Padova, 16-18.

Leulescu, A. & Agafitei, M. (2013). Statistical matching: a model based approach for data integration. Eurostat Methodologies and working papers.

Rässler, S. (2002). Statistical matching. Springer.

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

Proyectos europeos

Data Integration: http://www.cros-portal.eu/content/data-integration-finished

ISAD Integration of Survey and Administrative Data. http://www.cros-portal.eu/content/isad-finished

Report WP2 ESS-net, Statistical Methodology Project on Integration of Surveys and Administrative Data. Recommendations on the use of methodologies for the integration of surveys and administrative data.

Documentos internos EUSTAT

EUSTAT, Ficha metodológica de la encuesta Población en Relación con la Actividad (PRA) http://es.eustat.es/document/poblact_c.html

EUSTAT, Ficha metodológica de la Encuesta de Condiciones de Vida (ECV), http://es.eustat.es/document/ecvida_c.html#axzz37QhmZ17l

Software

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Chambers, J. (2008). Software for data analysis: programming with R. Springer.

D'Orazio, M. (2013). StatMatch: Statistical Matching (aka data fusion). http://CRAN.R-project.org/package=StatMatch

Lumley, T. (2012) survey: analysis of complex survey samples. http://CRAN.R-project.org/package=survey

Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL http://www.jstatsoft.org/v45/i03/

Harrell Jr, F. E. (2014). Hmisc: Harrell Miscellaneous. R package version 3.14-3. http://CRAN.R-project.org/package=Hmisc

Honaker, J.,King, G. & Blackwell, M. (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), 1-47. URL http://www.jstatsoft.org/v45/i07/.

Meinfelder, F. (2011). BaBooN: Bayesian Bootstrap Predictive Mean Matching – Multiple and single imputation for discrete data. http://CRAN.R-project.org/package=BaBooN

Anexos

VARIABLES ESPECÍFICAS PARA EL ENLACE				
Variable PRA (única)				
Variable	**Variable microdatos**	**--**	**Nombre corto**	**Categorías agregadas[a]**
Relación con la actividad ampliada	PV1_PRA2		PRA22	1-Actividad laboral (Ocupados)[a]; 2-Actividad no laboral (Tareas hogar, estudiates, servicio militar); 3-Actividad no laboral (Buscando empleo); 4-Parados

Variables de ECV (muestra)				
Variable	**Variable microdatos**	**Cuestionario[c]**	**Nombre corto**	**Categorías**
Trastornos de salud	CVI_TRASPI	ind	SAL	1-Algún transtorno; 2-Ningún transtorno
Conocimiento de idiomas	CVI_COIDI	ind	IDM	1-Sólo castellano; 2-Castellano y otros; 3-Castellano y euskara; 4-Castellano, euskara y otros
Tiempo libre	CVI_TLIBRR	ind	LIB	1-Menos de 2h; 2-2h-4h; 3-Más de 4h
Situación económica objetiva	CVI_SITEC2	fam	ECO	1-Mala; 2-Normal; 3-Buena
Tenencia de vehículos a motor[b]	CV1_NMOTOR; CV1_NCOCHR;	fam	VEHICL	1-Ninguno; 2-Uno; 3-Dos o más
Equipamiento del hogar	CV1_EQUIP6	fam	EQP	1-Sin equipamiento; 2-Con algunos; 3-Con bastantes

[a] Dentro de los ocupados la PRA distingue sub-categorías en función del grado de ocupación. En este estudio se han agregado en una única categoría: 'Ocupados'

[b] Variable generada a partir de: CV1_NMOTOR 'Número de motocicletas de más de 50 cc', CV1_NCOCHR 'Número de coches', CV1_NFURGR 'Número de furgonetas', CV1_OTRVEHICR 'Otros vehículos'

[c] Cuestionario de origen: individual (ind) o familiar (fam).

Ficha A: Variables específicas para el enlace de las encuestas independientes de EUSTAT , ECV y PRA.

	PRA-2009 4T		ECV-2009			
Variable	**Nombre[a]**	**nº pregunta cuestionario[b]**	**Nombre[a]**	**Fichero origen[c]**	**nº pregunta cuestionario[b]**	**Niveles agregados[c]**
Variables sociodemográficas						
Sexo	PV1_SEXO	p12	ind: CV1_SEXOI	ind	--	H-Hombre; M-Mujer
Edad	PV1_EDAD	p10	ind: CV1_EDADIR	ind		01-"<=15 años"; 02-"16-24 años"; 03-"25-34 años"; 04-"35-44 años"; 05-"45-54
Tamaño familiar	TAMAÑO_FAM	--	fam: CV1_TFAMR	fam	--	1 ,2, 3+ miembros
Nivel de instrucción						
Cursa estudios reglados S/N[e]	PV1_ENRE (!D)[d]	p43	CV1_SITES (B)	ind	pI2	1- Estudia; 0-No estudia
Realiza estudios a distancia S/N	PV1_ENRE (C)	p43	CVI_SISTE (F)	ind	pI4	1- Sí; 0-No
Analfabeto S/N	PV1_LEES	p34	CVI_ANALF	ind	pI15	1-Sí; 0-No
D. Variables relacionadas con situación laboral						
Relación con Actividad - OIT	PV1_PRA1	--	CV1_RELA1	ind	--	Ocupados; Parados; Inactivos
Buscando empleo S/N	PV1_BUSQ	p140	CVI_BUSQ	ind	pT23	1- Sí; 0-No
Horas trabajadas semana (Ocup.)	PV1_HTRA[g]	p107	CVI_HOTAT	ind	pT22	Numérica
Dedicación a tareas del hogar	PV1_SILH	p55	CV1_TDOME1	ind	Indicador sintético[f]	Se dedica; No se dedica

[a] Nombre de la variable en el fichero de microdatos.

[b] Número de pregunta en el cuestionario.

[c] Fichero origen para la ECV: indica en el fichero de microdatos de origen (individual: ind; familiar: fam)

[d] Símbolo '!': se han tomado todas las categorías excepto la indicada.

[e] En adelante 'Estudiante S/N'.

[f] Indicador derivado de 4 ítems de la pregunta pT27: COMPR2-Comprar alimentos ; COMID2-Preparar comida; FREG2-Fregar vajilla; ROPA2-Preparar la ropa; LIMPC2-Limpiar la casa.

[g] Segmento OIT=Ocupados

Ficha B: Meta-análisis de las variables compartidas entre las encuestas ECV y PRA.

Variables comunes		Coherencia	Valor predictivo						
Variable	Nombre corto	Dist. Hell.	V de Cramer (global)						
			PRA: "PRA"	**ECV:** "SAL"	**ECV:** "IDM"	**ECV:** "LIB"	**ECV:** "ECO"	**ECV:** "VEH"	**ECV:** "EQP"
Edad	ED	0,002	0,372	0,304	0,267	0,230	0,152	0,289	0,396
Sexo	S	7,68E-06	0,409	0,014	0,027	0,075	0,051	0,123	0,054
Tamaño familiar	TF	0,001	0,192	0,169	0,099	0,167	0,194	0,285	0,275
Estudiante S/N	EST	0,009	0,329	0,097	0,333	0,138	0,028	0,084	0,089
Ocupado S/N	OCP	0,002	1,000	0,275	0,291	0,402	0,345	0,373	0,285
Parado S/N	PAR	0,012	0,998	0,021	0,046	0,105	0,142	0,029	0,073
Inactivo S/N	INA	0,007	0,999	0,288	0,309	0,380	0,285	0,366	0,322
Buscando empleo S/N	BUSQ	0,025	0,850	0,067	0,118	0,037	0,073	0,035	0,087
Tareas hogar	DOM	0,054	0,548	0,054	0,061	0,105	0,032	0,096	0,002

Tabla C1. *Medidas globales.*

Estrato	Variables comunes seleccionadas	Tamaño muestra		Coherencia	Valor predictivo
(Nombre corto)	Variables seleccionadas por estrato	ECV (receptor)	PRA (donante)	Distancias Hell. (por variable)	V de Cramer Var. dependiente: "PRA"
H.15-24	EST, BUSQ	157	512	EST: 0,048 - BUSQ: 0,052	EST: 0,887 - BUSQ: 0,907
M.15-24	EST, BUSQ	166	491	EST: 0,09 - BUSQ: 0,003	EST: 0,835 - BUSQ: 0,848
H.25-34	BUSQ, DOM	359	746	BUSQ: 0,002 - DOM: 0,064	BUSQ: 0,895 - DOM: 0,389
M.25-34	PAR, DOM	372	740	PAR: 0,004 - DOM: 0,049	PAR: 1,000 - DOM: 0,378
H.35-44	PAR, TF2	426	920	PAR: 0,011 - TF2: 0,047	PAR: 1,000 - TF2: 0,117
M.35-44	PAR, OCP	386	957	PAR: 0,019 - OCP: 0,051	PAR: 1,000 - OCP: 1,000
H.45-54	PAR, BUSQ	377	966	PAR: 0,004 - BUSQ: 0,034	PAR: 1,000 - BUSQ: 0,870
M.45-54	PAR, BUSQ	403	1034	PAR: 0,032 - BUSQ: 0,004	PAR: 1,000 - BUSQ: 0,730
H.55-64	INA, BUSQ	349	843	INA: 0,033 - BUSQ: 0,023	INA: 1,000 - BUSQ: 0,848
M.55-64	INA	391	887	INA: 0,027	INA: 1,000
H.+65	TF2	540	1158	TF2: 0,018	TF2: 0,295
M.+65	TF2	823	1611	TF2: 0,030	TF2: 0,111

Tabla C2. *Medidas por estratos de Edad y Sexo para las variables seleccionadas.*

Ficha C: Selección de variables por estrato. Medidas empíricas de coherencia (distancias de Hellinger) y del valor predictivo (V de Cramer). Medidas globales (Tabla C1) y por estrato (Tabla C2).

Organismo Autónomo del

www.eustat.eus