TRABAJO DE FIN DE MÁSTER

Máster en Estadística e Investigación Operativa

Título: FUSIÓN DE DATOS ENTRE EPA Y ECV PARA LA ESTIMACIÓN DE INDICADORES ESTADÍSTICOS REGIONALES

Autor: JENNY MONTOYA CUELLAR

Directores: Enric Ripoll Font

Mònica Gasulla Ramon

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Facultat de Matemàtiques i Estadística
Universitat Politècnica de Catalunya

Trabajo de fin de Máster

Fusión de datos entre EPA y ECV para la estimación de indicadores estadísticos regionales

Jenny Montoya Cuellar

Directores: Enric Ripoll Font
Mònica Gasulla Ramon

Departament d'Estadística i Investigació Operativa

Resumen

La fusión de datos es un enfoque basado en modelos para proveer información estadística conjunta sobre variables o indicadores observados en diferentes fuentes estadísticas, mediante la imputación de información en un fichero receptor, a partir de la información proveniente de otras observaciones similares en un fichero donante. En el marco del convenio de cooperación educativa entre la Facultad de Matemáticas y Estadística de la UPC y el Instituto de Estadística de Cataluña – Idescat, se llevó a cabo una fusión de datos con el fin de enriquecer el fichero de datos de la Encuesta de Condiciones de Vida (ECV) aprovechando la Encuesta de Población Activa (EPA) para estimar indicadores de pobreza regionales para Cataluña. Tras realizar el proceso de armonización de las fuentes de datos, la selección de variables puente y la aplicación y evaluación de las técnicas de fusión de datos, se estimó la tasa de riesgo de pobreza para Cataluña por dos métodos diferentes, en los ficheros de la ECV, la EPA y el fichero de los datos fusionados.

Palabras clave: Fusión de datos, encuestas sociales, indicadores regionales, tasa de riesgo de pobreza.

Abstract

Data fusion is a model-based approach to provide joint statistical information on variables or indicators observed in different statistical sources, by imputing information on a recipient file, from other similar observations in a donor file. Under the educational cooperation agreement between the Faculty of Mathematics and Statistics of the UPC and the Statistical Institute of Catalonia – Idescat, data fusion was carried on in order to enrich the database of Encuesta de Condiciones de Vida (ECV) by taking advantage of the Encuesta de Población Activa (EPA) to estimate regional poverty indicators for Catalonia. After performing the process of harmonization of data sources, selecting hinge variables and the implementation and evaluation of the data fusion techniques, the at risk of poverty rate for Catalonia was estimated by two different methods, in the files of ECV, EPA and the fused data file.

Keywords: data fusion, social surveys, regional indicators, at risk of poverty rate.

TABLA DE CONTENIDO

1. INTRODUCCIÓN

La Encuesta de Condiciones de Vida (ECV), cuya referencia europea se denomina European Union Statistics on Income and Living Conditions (EU-SILC), provee información sobre la distribución de ingresos y la exclusión social y tiene como objetivo la producción sistemática de estadísticas comunitarias sobre la renta y las condiciones de vida, que incluyan datos sobre la renta, el nivel y composición de la pobreza y la exclusión social.

La ECV desde su inicio en 2004 ha desempeñado un papel importante en la provisión de estadísticas socio-económicas claves, pero fue tras la adopción de los objetivos de Europa 2020, en junio de 2010 por el Consejo Europeo, que cobró más relevancia. La confirmación de los cinco objetivos principales de esta iniciativa, orientan la actuación de los Estados Miembros en materia de empleo, investigación y desarrollo, cambio climático y sostenibilidad energética, educación y la lucha contra la pobreza y la exclusión social. El seguimiento del progreso hacia la consecución del quinto de estos objetivos, se realiza sobe la base del indicador de pobreza y exclusión social (At-risk-of-poverty and social exclusión -AROPE), a partir de los datos recogidos por EU-SILC.

Las estadísticas que se obtienen de la ECV sirven de base para la formulación de políticas sociales y su seguimiento, relacionadas con este objetivo. En el caso español, se presentan algunos análisis y clasificaciones al nivel de comunidades autónomas (NUTS 2). La importancia de mejorar los datos recogidos y los indicadores calculados se ha evidenciado mediante la publicación de documentos estratégicos para la estadística europea oficial (Stiglitz et al, 2009; Comisión Europea, 2009), y ha incentivado al Sistema Estadístico Europeo (ESS) a tomar medidas para hacer frente a la nueva realidad.

En algunas ocasiones, la dificultad de obtener estimadores de pobreza regionales se debe a un tamaño de muestra pequeño que hace que los estimadores directos no tengan la precisión adecuada, y en otras se añaden las limitaciones ligadas a las encuestas como la falta de respuesta o la longitud de los cuestionarios que dificulta la obtención de toda la información específica deseada de una sola fuente. En cualquier caso, nos enfrentamos a un problema importante.

En los últimos años el ESS se ha esforzado por buscar la forma de mejorar la explotación de las fuentes de datos existentes, por ejemplo por medio de técnicas basadas en modelos como la fusión de datos y la estimación de áreas pequeñas llevando a cabo proyectos como EURAREA, EU-SAE, ISAD, DATA INTEGRATION y Net-SILC[1] (Leulescu et al., 2011). Además, en los Institutos Nacionales de Estadística en Europa se realizan constantes esfuerzos para obtener resultados confiables y oportunos, al tiempo que se intenta minimizar los costos y las cargas

[1] Portal on Collaboration in Research and Methodology for Official Statistics. http://www.cros-portal.eu/

soportadas por los informantes, que tendrían que asumirse si se diseñara una nueva encuesta que cumpla con los objetivos deseados (Leulescu y Agafitei, 2013).

En este sentido, la fusión de datos se presenta como una alternativa que permite aprovechar mejor la información disponible y reducir costos para obtener datos que sirvan a múltiples propósitos, ampliando el alcance y satisfaciendo las nuevas demandas.

La integración de datos se puede realizar por medio de tres metodologías diferentes: data merging, record linkage y statistical matching (D'Orazio et al., 2001). Esta última metodología, que comprende la fusión de datos, es un enfoque basado en modelos para proveer información estadística conjunta sobre variables o indicadores observados en diferentes fuentes estadísticas, cuyas unidades de análisis provienen de una misma población y poseen variables en común pero no se superponen. Básicamente consiste en imputar a unos individuos (receptores) información para algunas variables a partir de la información proveniente de otros individuos (donantes), a los que se les han observado algunas características comunes y que se relacionan con la información que se quiere estimar.

Bien sea con el objetivo de ampliar el alcance de la información existente o de mejorar la precisión de las estimaciones realizadas, el statistical matching también se clasifica dependiendo la estructura de los conjunto de datos que se van a fusionar y la estructura que se desea que tengan los datos fusionados. Existen diferentes tipos de operaciones: 1) Inyección, 2) Fusión unilateral o fusión de datos, y 3) Fusión recíproca (Juárez, 2005).

Este trabajo, realizado bajo un convenio de cooperación educativa entre la Facultad de Matemáticas y Estadística de la Universidad Politécnica de Cataluña y el Instituto de Estadística de Cataluña – Idescat, se lleva a cabo una fusión de datos con el fin de enriquecer el fichero de datos de la Encuesta de Condiciones de Vida (ECV) aprovechando la Encuesta de Población Activa (EPA) para estimar indicadores de pobreza regionales, buscando aumentar la precisión de las estimaciones realizadas a partir de la ECV para Cataluña. Dada la correlación entre mercado laboral y la situación económica de los hogares, es relevante el uso complementario de ambas fuentes estadísticas (Aluja y Daunis, 2012).

Para realizar la fusión de datos se utilizó el software R. Antes de poder realizar cualquier tipo de integración de datos, existen unos ciertos pre-requisitos que se deben cumplir y una serie de pasos que se deben seguir, a partir de los cuales se estructura la presente memoria.

Primero, se presenta un estado del arte sobre el tema en el capítulo 2. En el capítulo 3 se describe la metodología detallada, y en el capítulo 4 se describen los ficheros y el proceso de armonización y coherencia entre las fuentes de datos, la selección de variables comunes y variables puente, y la aplicación de las técnicas de fusión de datos. La evaluación de la calidad de la fusión se presenta en el capítulo 5 y finalmente las conclusiones.

2. REVISIÓN DE LA LITERATURA

La integración de datos provenientes de diferentes fuentes ha tomado fuerza en la última década en la Unión Europea. Aunque la mayoría de la literatura inicial se centra en Estados Unidos, a nivel europeo se ha continuado con el desarrollo del tema y se han llevado a cabo aplicaciones de diferentes tipos. Si bien las más comunes se han dado en el marco de las estadísticas oficiales dentro del ESS, el uso de esta metodología también se hizo popular en el área de marketing (Aluja et al., 2007; Rius et al., 1999).

La investigación en el área de la economía personal en los Estados Unidos en la década de 1960 produjo una serie de estudios en los que la característica común era la recreación de ficheros de datos de muestras de los individuos y las familias, con el objetivo de incorporar información adicional que pudiera ser usada para realizar estimaciones sobre variables no incluidas en el fichero original, dando los primeros pasos en el área de la fusión de datos.

Algunos de estos estudios son mencionados por Budd, uno de los primeros artículos publicados sobre el tema en el que el autor describe los métodos utilizados por el Departamento de Comercio de los Estados Unidos para crear un fichero de microdatos que permitiera complementar y estimar información sobre la distribución de los ingresos de los hogares a partir de la Encuesta Continua de Población (CPS) (Budd, 1971). Al mismo tiempo, otro trabajo de este tipo era desarrollado por Okner , quien combinó información de registros administrativos del Tax File 1966 con la Encuesta de Oportunidades Económicas (SEO) y estableciendo unas "clases de equivalencia" de las variables X para hacer una asignación aleatoria de (X,Z) entre (X,Z)' "equivalentes" que alcanzaran una puntuación mínima de cercanía (Okner, 1972a). Este procedimiento fue criticado por Sims, argumentando que Okner hizo el supuesto implícito de que Y y Z eran independientes dado X (conocido como supuesto de independencia condicional), hace hincapié en la necesidad de una teoría para la integración de datos (Sims, 1972), A partir de aquí, se genera una primera discusión en la que Peck defiende el supuesto (Peck, 1972) y Okner responde sobre su validez (Okner, 1972b). En una segunda ronda de la discusión un par de años después, Okner, Ruggles y Ruggles y Alter muestran algunas mejoras en el método aunque continúan concentrados en las clases de equivalencia (Okner, 1974), (Ruggles y Ruggles, 1974), (Alter, 1974).

Hasta ese momento la mayoría de las publicaciones se centraban más en mostrar la aplicación de las diferentes metodologías desarrolladas y describirlas sólo con palabras, que en perfilar un marco teórico más sólido sobre el tema. En 1978 Kadane hizo un gran avance formalizando la notación para describir el problema fundamental al que se refiere la integración de datos (Kadane, 1978). En una primera sección considera el caso en el que los ficheros a fusionar están constituidos por los mismos individuos, lo que ahora se conoce como Exact Matching o Record Linkage, y en

la segunda sección se ocupa del caso en que los ficheros son muestras independientes tomadas de una misma población (Statistical Matching).

En 1980 el Comité Federal de Metodología Estadística de Estados Unidos preparó un reporte sobre Exact Matching y Statistical Matching, resumiendo básicamente los trabajos realizados en la década de 1970 (Radner, 1980). En él se presentan algunos casos de aplicaciones, sobre todo en agencias gubernamentales, y se describe diversos procedimientos desarrollados. Además, de discutir los procedimientos, también trata otros aspectos como su desarrollo, ventajas y desventajas respecto a otras alternativas, la confidencialidad y la precisión de los resultados de la fusión, empezando por intentar convenir las definiciones básicas y terminando con algunas recomendaciones para el uso de estas técnicas.

En los años posteriores aparecen publicaciones que desarrollan nuevas técnicas, no solo para la ejecución de la integración de datos, sino también para su evaluación, como es el caso de Rodgers, que destaca que los procedimientos existentes se habían desarrollado sin mucha fundamentación teórica o justificación empírica y sólo en los años más recientes se estaban realizando algunos esfuerzos por corregir esa deficiencia (Rodgers, 1984). El autor hace una revisión de esos esfuerzos y evalúa la utilidad potencial de la fusión de datos. Argumenta que la validez de los conjuntos de datos obtenidos a partir de la fusión de datos depende de la precisión de los supuestos subyacentes sobre las relaciones entre las variables y basándose en simulaciones provee una base para la elección de las técnicas de fusión y la evaluación de su utilidad.

Rubin se centra en la idea de que al no poder recrear los verdaderos valores de las variables específicas para el archivo receptor, los ficheros creados tienen incertidumbre y que eso genera incertidumbre en las inferencias realizadas a partir de esos datos (Rubin, 1986). Este autor propone un nuevo enfoque llamado concatenación de ficheros con pesos ajustados e imputaciones múltiples, con el que se podrían crear ficheros a partir de la fusión de datos que permitan la valoración directa de la incertidumbre debida tanto a la varianza de la muestra como a los supuestos, que denomina "implícitos e inestables", respecto a la relación entre las variables en los diferentes ficheros.

El tema de la fusión de datos continua su evolución en la década de 1980 y 1990, con aplicaciones en la estadística oficial y el área de marketing principalmente, para el aprovechamiento de datos de encuestas. La consideración del uso de información auxiliar para mejorar los resultados y como alternativa al supuesto de la independencia condicional toma fuerza (Paass, 1986; Singh et al., 1993).

El cumplimiento de ciertos requisitos para llevar a cabo una fusión de datos quedó "establecido" con Van der Laan, referencia utilizada en la mayoría de los casos desde su aparición (Van der

Laan, 2000). Lo que denominó como proceso de micro-integración consiste en la reconciliación de los ficheros de datos en 9 pasos, y es ahora el punto de partida para realizar una fusión de datos.

D'Orazio et al., resume la metodología de la integración de datos haciendo notar que no solo es un problema de tecnología de la información sino también estadístico e introduce la definición de objetivos micro y macro (D'Orazio et al., 2001). Establece que un primer paso debe ser el de verificar si los ficheros a fusionar cumplen ciertas condiciones para hacer la integración o si es necesario hacer el paso preliminar reconciliación al que llama "armonización". Dependiendo del tipo de objetivo definido hace un resumen mencionando las técnicas disponibles (imputación por regresión, hot deck, imputación múltiple, etc.). Finalmente, trata la cuestión de la evaluación de la calidad de la fusión, diciendo que es un tema aún abierto al que Rässler presenta una propuesta (Rässler, 2004). En su artículo parte de que la integración de datos está relacionada con un problema de identificación en lo que se refiere a la asociación de variables no observadas conjuntamente. Esa asociación condicional no puede ser estimada de los datos observados, sin embargo, dependiendo del poder explicativo de las variables comunes y usando métodos de imputación múltiple en modelos explícitos, hay alguna posibilidad de estimar unos límites para la $\mathrm{cov}(Y, Z)$. Rässler propone la validación del proceso de integración de datos basándose en esos límites, distinguiendo cuatro niveles de validación: 1) Preservación de los valores individuales, 2) Preservación de las distribuciones conjuntas, 3) Preservación de las estructuras de correlación, y 4) Preservación de las distribuciones marginales.

Respecto a la fusión unilateral, Aluja et al. describe tres enfoques básicos de la metodología de fusión de datos y presenta GRAFT, un sistema multipropósito basado en el método de imputación KNN hot deck (Aluja et al., 2007). En 2013, Aluja y Daunis se resumen las técnicas más comúnmente usadas y se exponen los diferentes pasos a seguir en un manual de procedimiento para fusión de datos (Aluja y Daunis, 2013).

Dentro del Sistema Estadístico Europeo, se destacan dos proyectos relacionados con el tema. Finalizado en 2008, el proyecto de integración de encuestas y datos administrativos, ISAD, promueve la aplicación de 3 metodologías para el uso de las fuentes de datos existentes en la producción de estadísticas oficiales: Record Linkage, Statistical Matching y Microintegration Processing. En el primer informe, se presenta el estado del arte de estas metodologías y algunas experiencias prácticas. En el segundo, se hace una serie de recomendaciones sobre su uso y el último abarca las herramientas informáticas para su implementación.

El segundo proyecto llamado *Data Integration*, que fue realizado entre 2009 y 2011, discute la precisión de los métodos de integración y la utilidad de los datos fusionados, enfocándose en que sean aplicables por los Institutos Nacionales de Estadística. En este proyecto se actualizó estado de la arte de las mismas metodologías presentadas en el ISAD, se discutió su desarrollo

metodológico y se expusieron algunos casos de estudio y herramientas de software disponibles para la integración de datos. En uno de los casos allí presentados, se utiliza la metodología de la fusión de datos en un estudio de simulación para enriquecer la base de datos del Microcenso con las variables de la encuesta de fuerza laboral (LFS) y aumentar el tamaño de muestra de esta última en Polonia.

Existen otros casos prácticos relacionados con las dos encuestas objeto de estudio en este trabajo. Uno de los Institutos Nacionales de Estadística más activo en el tema es el italiano (ISTAT). En 2008 ya presentaban unos resultados preliminares de la fusión entre la LFS y la encuesta de uso del tiempo (TUS) para Italia (Gazzelloni et al., 2008).

La fusión de datos de la encuesta de condiciones de vida (EU-SILC) y la encuesta de presupuestos familiares (HBS) se ha llevado a cabo con datos del Reino Unido para comparar estimadores de pobreza usando los ingresos, gastos y la privación material (Eurostat, 2013), y para el caso de Italia con el fin de elaborar estadísticas conjuntas sobre ingresos del hogar, el consumo y la riqueza (Donatiello et al., 2014).

En 2011 se llevó a cabo una fusión de datos entre LFS y EU-SILC utilizando datos de Austria (Leulescu et al., 2011). Estas mismas encuestas son fusionadas para siete países y sus resultados se presentan en una publicación de Eurostat que da un panorama general de la metodología de fusión de datos y su implementación, mostrando un par de estudios piloto llevados a cabo. El otro estudio presentado en esta publicación utiliza las encuestas EU-SILC y European Quality of Life Survey (EQLS) de Finlandia y España (Leulescu y Agafitei, 2013).

3. FUNDAMENTACIÓN TEÓRICA

Statistical Matching es un enfoque basado en modelos para proveer información estadística conjunta sobre variables o indicadores observados en diferentes fuentes estadísticas, cuyas unidades de análisis generalmente provienen de una misma población y poseen variables en común pero no se superponen (Leulescu y Agafitei, 2013).

El objetivo principal es la integración de diferentes fuentes con el fin de estudiar la relación existente entre algunas variables no observadas conjuntamente. Esto se puede lograr mediante dos enfoques diferentes según el resultado que se desea. D'orazio et al. los define como enfoques micro y macro (D'orazio et al, 2006).

El enfoque micro tiene como objetivo construir un conjunto de datos sintético completo, transformando los distintos conjuntos de datos en uno solo integrado cuyos registros se refieren a la misma unidad de análisis y donde todas las variables de interés están presentes. El término "sintético" se refiere al hecho que este archivo no es el resultado de observaciones directas de todas las variables sobre un conjunto de individuos que pertenecen a la población de interés.

El objetivo del enfoque macro es transformar los conjuntos de datos en resultados agregados, identificando estructuras que describan la relación entre las variables de las dos fuentes que no han sido observadas de manera conjunta (tablas de contingencia, matrices de correlación, distribuciones conjuntas y marginales, etc.).

En el caso típico de statistical matching, se tiene dos fuentes de datos A y B provenientes de dos encuestas muestrales independientes que comparten un conjunto de variables comunes X observadas en ambas, pero existen unas variables específicas Y observadas solamente en A y unas variables específicas Z observadas sólo en B, de manera que Y y Z no se observan de manera conjunta (Kadane, 1978). La relación entre las variables comunes y específicas se observa sólo en uno de conjunto de datos. Es decir que X y Y se observan en la muestra A, mientras que X y Z se observan en la muestra B. Cada conjunto de datos se usa para imputar las variables que no se han observado directamente en las unidades del otro conjunto de datos. Así, se genera un conjunto de datos sintético con información completa de X, Y y Z.

Bien sea con el objetivo de ampliar el alcance de la información existente o de mejorar la precisión de las estimaciones realizadas, el statistical matching se clasifica dependiendo la estructura de los conjunto de datos que se van a fusionar y la estructura que deseamos que tengan los datos fusionados. Los casos más comunes son:

1. A y B son submuestras de una misma encuesta.

2. Las muestras A y B son cada una donante y receptora de la otra. Este caso es conocido como fusión recíproca (Figura 1).
3. A y B son extraídas de dos encuestas diferentes, son muestras independientes de una misma población que comparten un cierto número de variables a partir de las cuales se transfieren variables específicas a una muestra receptora B, desde otra donante A. Este caso se conoce como fusión unilateral, fusión de datos o enriquecimiento de la base de datos (Figura 2).

El fichero A que contiene n_0 observaciones con información sobre $p+q = u$ variables. El fichero B contiene n_1 observaciones con información sobre $p+r = v$ variables. Ambos ficheros tienen p variables comunes X correlacionadas con Y y Z, siendo X_0 las variables comunes del fichero A y X_1 las del fichero B. Existen q variables específicas en el fichero A denotadas como Y_0 y r variables específicas en el fichero B denotadas como Z_1.

	Y	**X**	**Z**
Muestra A	Y_0	X_0	*?*
Muestra B	*?*	X_1	Z_1

Figura 1. Fusión Recíproca.

En el caso de la fusión recíproca, se trata de imputar las variables especificas Y_1 y Z_0 en B y A, respectivamente. Para esto es necesario adoptar el supuesto de independencia condicional, CIA por sus siglas en inglés (Conditional Independence Assumption), o usar información auxiliar, ya que los parámetros críticos son aquellos que involucran las relaciones entre las variables Y y Z que no se observan conjuntamente. El CIA asume que Y y Z son estadísticamente independientes condicionadas a X. Así, la distribución condicional conjunta de Y y Z dado X es

$$f(Y,Z/X) = f(Y \mid X) f(Z \mid X)$$

Este supuesto no puede comprobarse con los datos disponibles de los ficheros A y B, por lo que debe asumirse en base a otras informaciones sobre el fenómeno investigado, o comprobando la independencia con otros datos históricos o similares en los que se hayan observado las variables de manera conjunta (D'Orazio et al., 2001; ISAD, 2008).

La situación que se presenta en este trabajo corresponde a una fusión unilateral con enfoque micro. Esta situación, también conocida como fusión de datos, se representa en la Figura 2, en la que el

fichero de la muestra A es el donante y su información se usa para completar el fichero receptor B a nivel individual. La muestra donante es aquella que tiene todas las variables de interés mientras la receptora tiene un bloque de variables faltantes y por lo general posee más observaciones que la donante.

	X	**Y**
Muestra A	X_0	Y_0
Muestra B	X_1	*?*

Figura 2. Fusión Unilateral.

(X_0, Y_0) constituye el fichero donante A, con información completa sobre las variables comunes X_0 y las específicas Y_0, y (X_1) el fichero receptor B se compone de las variables comunes X_1 y también puede contener variables específicas Z_1, aunque a efectos prácticos estas últimas no se tienen en cuenta para la fusión.

En el caso de la fusión unilateral la independencia condicional estaría dada por $f(Y/X, Z) = f(Y/X)$ para cualquier conjunto de variables Z.

En resumen, la fusión de datos es un caso particular del statistica matching que consiste en imputar a unos individuos, en el fichero receptor, información sobre algunas variables a partir de la información proveniente de otros individuos, del fichero donante, a los que además se les ha observado algunas características comunes y que se relacionan con la información que se quiere estimar. Esta imputación equivale a estimar los valores que se podrían haber observado de las variables específicas si se hubieran medido (Aluja y Daunis, 2013). Puede verse como un problema de valores faltantes, con la diferencia de que el bloque de variables Y_1 es faltante por diseño, asumiendo que los valores faltantes son completamente aleatorios (MCAR) en el caso de que tanto el fichero donante como el receptor provengan de muestras tomadas de una misma población, o que los valores faltantes son aleatorios (MAR) en caso contrario (Aluja, et al., 2007).

Para hacer la fusión de datos es necesario que las bases de datos que se van a fusionar cumplan ciertas condiciones y se sigan una serie de pasos cuidadosamente. Esta verificación de las condiciones y la adaptación a ellas es el primer paso del proceso, al que se le llama *armonización*. El segundo paso es la selección de las variables que servirán para fusionar los ficheros, llamadas variables puente. El tercer paso consiste en la aplicación de las técnicas de fusión de datos y finalmente, se evalúa la calidad de la fusión.

3.1. Selección de variables comunes y armonización

Van Der Laan propone un listado de puntos a verificar en el proceso de reconciliación de las fuentes con especial referencia a la comparabilidad en el tiempo (Van der Laan, 2000). Estas tareas de armonización pueden agruparse en las relacionadas con las unidades estadísticas (unidad de análisis, períodos de referencia, población de referencia), las relacionadas con las variables (definición, clasificación y distribución) y las relacionadas con otros aspectos operacionales como el ajuste de datos missing y la derivación de variables (ISAD, 2008).

Básicamente, para que sea posible una fusión de datos entre dos fuentes estadísticas provenientes de encuestas muestrales, estas deben ser muestras independientes sobre la misma población con una misma unidad de análisis, que tengan algunas variables específicas y otras variables comunes. Estas últimas con las mismas definiciones, escalas, clasificaciones (Leulescu et al., 2011).

Normalmente las variables comunes necesitan ser recodificadas, agregadas, cambiadas de formatos cuantitativo a cualitativo o viceversa y en algunos casos se deben crear nuevas variables a partir de las existentes, para que estén definidas y clasificadas de la misma manera en ambas fuentes y poder usarlas en el análisis. Las variables que no se pueden armonizar se deben descartar (D'Orazio et al., 2006).

3.2. Selección de variables puente

Del conjunto de variables comunes, deberán seleccionarse aquellas que serán las variables predictivas en el modelo de imputación. No todas las variables comunes podrán ser utilizadas como variables predictivas ya que es posible que algunas variables no se puedan armonizar, no tengan poder explicativo sobre la variable Y o podrían existir problemas de colinealidad. Además, entre mayor sea el número de variables, mayor será el coste computacional a la hora de estimar los parámetros del modelo.

La correcta elección de estas variables tiene un gran impacto sobre la validez de la fusión, que depende también de que las variables elegidas sean capaces de predecir la información específica que se transferirá del fichero donante al receptor, y el supuesto de independencia condicional es el punto de partida. Que la distribución de Y condicionada a X sea independiente de cualquier conjunto de variables Z, valida los procedimientos de inferencia sobre la relación no observada e induce una fuerte relación predictiva entre las variables comunes y las especificas del fichero donante (Leulescu et al., 2011).

Una primera opción, aunque no es la más óptima, sería identificar las variables comunes X que están estadísticamente relacionadas con las variables Y, mediante el cálculo de medidas de

asociación entre todos los pares de variables X y Y en el fichero donante, y elegir aquellas que presentan las asociaciones más altas (ISAD, 2008).

Un aspecto a tener en cuenta para seleccionar las variables puente, es la calidad de las variables comunes. Serán preferibles variables sin errores y sin datos faltantes, y deben evitarse aquellas que tengan un alto grado de imputación (D'Orazio et al., 2006).

Como este procedimiento no tiene en cuenta que podrían seleccionar dos variables comunes correlacionadas entre ellas, es mejor analizar la asociación de todas las variables comunes en conjunto. Se pueden utilizar diversos métodos multivariantes o hacer un modelo de regresión de Y sobre las variables X y utilizar un procedimiento de selección de variables como el paso a paso, o modelos lineales generalizados en el caso de tener una variable Y continua y predictores mixtos. Si se supone que la relación entre X y Y no es lineal, una posibilidad serían los árboles de clasificación y regresión (CART) (ISAD, 2008).

3.3. Técnicas de fusión de datos

Para elegir la técnica de fusión de datos que se va a utilizar, es necesario tener en cuenta algunos aspectos como el enfoque de la fusión (micro o macro), la validación a priori (si se asume algún supuesto o se utiliza información auxiliar) y las características del modelo de imputación: paramétrico o no paramétrico, condicional o no condicional, determinista o estocástico, imputación simple o múltiple

Respecto a la validación a priori, el supuesto CIA es el más utilizado en el caso de fusión recíproca. Para la fusión unilateral, se describen los siguientes tres supuestos que podrían considerarse (Aluja y Daunis, 2013):

1. **Relevancia predictiva**: asume que la variable específica Y es explicada en su totalidad por las variables puente. Este supuesto implica la independencia condicional, ya que si todo el poder predictivo de Y está contenido en las variables puente X, cualquier otro conjunto de variables Z tendría que ser independiente de Y dado X.

$$Y = i(X) + \varepsilon$$

donde i(X) representa el modelo de imputación elegido y ε la fluctuación aleatoria.

2. **Preservación de la distribución condicional**: asume que la distribución condicional de Y dado X es igual en ambos ficheros $f(Y_1 \,/\, X_1) = f(Y_0 \,/\, X_0)$, aunque la distribución conjunta puede ser diferente, ya que lo que interesa es transferir las variables específicas del fichero donante al receptor a nivel individual, basados en la distribución $f(Y_0 \,/\, X_0)$.

$$f\left(X_0, Y_0/\theta_{f(X,Y)}\right) = f\left(Y/X, \theta_{Y/X}\right) f\left(X_0/\theta_{f(X)}\right)$$

$$f\left(X_1, Y_1/\theta_{g(X,Y)}\right) = f\left(Y/X, \theta_{Y/X}\right) g\left(X_1/\theta_{g(X)}\right)$$

3. **Representatividad del fichero donante**: No se asume que las muestras de los dos ficheros provienen de la misma población, sólo se asume que el fichero donante es una muestra representativa de la población

$$f(X_0, Y_0) \neq g(X_1, Y_1)$$

Existen diversos tipos de modelos de imputación según sus características:

Condicional o No condicional: Uno de los métodos más conocido es el de imputación por la media, en el que los valores faltantes se sustituyen por la media de las unidades observadas de esa variable. Puede hacerse imputación por media no condicional, que no tiene en cuenta los valores de las variables puente para hacer la imputación y preserva el valor medio de la variable pero los estadísticos que definen la forma de la distribución (varianza, sesgo, etc.) y las relaciones con otras variables pueden verse afectados. La imputación por media condicional imputa medias condicionadas a los valores observados. Un método común consiste en agrupar los valores observados y no observados en clases e imputar los valores faltantes por la media de los valores observados en la misma clase.

Deterministas o Estocástico: Los métodos deterministas $i(X) \leftarrow E(Y/X)$ producen las mismas respuestas al repetir la imputación a los mismos individuos bajo las mismas condiciones, es decir que a un individuo concreto se le imputa siempre el mismo valor. Los métodos estocásticos o aleatorios $i(X) \leftarrow f(Y/X)$ producen resultados diferentes cuando se repite el método de imputación bajo las mismas condiciones, el valor que se imputa se extrae de una distribución de probabilidad.

Simple o múltiple: La imputación simple consiste en asignar un valor por cada valor faltante basándose en el valor de la propia variable o de otras variables generando un conjunto de datos completo, mientras que la imputación múltiple propuesta por Rubin, asigna a cada valor faltante

m valores, m $>$ 1, generando m conjuntos de datos completos. En cada conjunto de datos completo se estiman los parámetros de interés y posteriormente se combinan los resultados obtenidos. A través de un proceso aleatorio se muestran posibles valores para los datos faltantes y la utilización de dichos valores recoge el componente aleatorio del dato imputado (Rubin, 1986).

Paramétrico o no paramétrico: Se usan las técnicas paramétricas si se puede suponer que la distribución conjunta $f(X,Y/\theta)$ pertenece a una familia de distribuciones conocida, usualmente la Normal para variables continuas y Multinomial para las categóricas. Como $f(X,Y/\theta) = f\left(Y/X,\theta_{Y/X}\right)f(X,\theta_X)$, es posible estimar los parámetros $\theta_{Y/X}$ y θ_X a partir de la información disponible y usarlos para la imputación (Aluja, et al., 2007).

Los métodos no paramétricos difieren de los paramétricos en que no se especifica la estructura del modelo a priori. Entre las técnicas no paramétricas más utilizadas están las pertenecientes a los métodos denominados hot deck.

En la literatura existen también los métodos mixtos y la modelización explicita. Los primeros consisten en una combinación en un proceso de dos etapas entre métodos paramétricos y no paramétricos, con el que se intenta añadir a la parsimonia del enfoque paramétrico, la robustez de técnicas no paramétricas, estimando primero un modelo paramétrico y luego aplicando técnicas de imputación hot deck (Leulescu y Agafitei, 2013). La modelización explicita intenta encontrar un modelo predictivo de las variables específicas en función del as variables puente, a partir del fichero donante y aplicar ese modelo estimado a los datos del fichero receptor (Aluja y Daunis, 2013).

3.3.1. Técnicas paramétricas

Se supone que las muestras A y B se han extraído aleatoriamente de una distribución de probabilidad normal

$$f(x,y/\theta) \sim N_{p+q}(\mu,\Sigma)$$

donde $\mu = \begin{bmatrix}\mu_X \\ \mu_Y\end{bmatrix}$ y $\Sigma = \begin{bmatrix}\Sigma_{\mathrm{X}} & \Sigma_{\mathrm{XY}} \\ \Sigma_{\mathrm{YX}} & \Sigma_{\mathrm{Y}}\end{bmatrix}$

Asumir la relevancia predictiva garantiza que los datos disponibles son suficientes para estimar los parámetros del modelo. Bajo este supuesto, la función de verosimilitud de la distribución conjunta se puede calcular como un producto de las funciones de verosimilitud condicionales para

cada uno de los conjuntos de datos y la función de verosimilitud de la distribución marginal de las variables. Así, se pueden aplicar métodos de máxima verosimilitud (MV) para la estimación de los parámetros y la identificación de la distribución. En algunas ocasiones se ha empleado los estimadores de mínimos cuadrados, que arrojan resultados similares para muestras grandes (Leulescu y Agafitei, 2013).

Factorizando la distribución conjunta

$$f(x, y/\theta) = f_{Y/X}\left(y/x, \theta_{Y/X}\right) f_X(x, \theta_X),$$

tenemos la función de verosimilitud

$$L\left(\theta_{Y/X}, \theta_X | \, x, y\right) = \prod_{i=1}^{n_0} f_{Y/X}\left(y/x, \theta_{Y/X}\right) \prod_{i=1}^{n} f_X(x, \theta_X)$$

donde $n = n_0 + n_1$.

Utilizando los n_0 individuos del fichero donante A se estimarían los paramétros de la distribución condicional $f_{Y/X}\left(y/x, \theta_{Y/X}\right)$ y utilizando todos los *n* individuos se estimarían los parámetros de la distribución marginal $f_X(x, \theta_X)$.

Así, se puede hacer la imputación de manera determinista $E[y/x] = \mu_Y + \Sigma_{YX_{n_0}} \Sigma_{X_{n_0}}^{-1} \left(x - \mu_{X_{n_0}}\right)$ o estocástica $y = E[y/x] + \varepsilon, \quad \varepsilon \sim N\left(0, \Sigma_{Y/X}\right)$.

3.3.2. Técnicas no paramétricas: Hot deck

Son las técnicas más populares en la fusión de datos ya que imputa los valores no observados en el fichero receptor con valores del fichero donante.

Random hot deck consiste en la elección al azar de un registro en el archivo donante para cada registro del fichero receptor, generalmente haciendo una estratificación según variables que permitan tener un conjunto de unidades homogéneo (ISAD, 2008).

En el *distance hot deck* se sustituyen los valores faltantes con valores observados en unidades estadísticas similares observadas en el fichero donante. La forma de medir esa similitud es calculando una medida de distancia para las variables comunes, de manera que cada registro del

fichero receptor es relacionado con el registro más cercano (el que presente menor distancia) en el fichero donante. Se puede utilizar el concepto de distancia euclidiana, de Mahalanobis o una distancia ponderada, dándole mayor peso a las variables en función de su poder explicativo, por ejemplo.

La asignación se puede realizar de diferentes maneras. En caso de que dos o más registros de los donantes presenten la misma distancia con los registros receptores, puede optarse por elegir uno al azar. En el caso no restringido, se imputa con el valor del registro más cercano, siendo posible la utilización de un mismo valor varias veces, existiendo la posibilidad de que algunos registros del fichero donante se queden sin utilizar. Algunas soluciones a este problema pasan por imputar a un registro receptor el valor medio de todos los registros del donante que encuentra dentro de una distancia establecida o imponer penalizaciones. El caso restringido permite que cada registro sea usado sólo una vez, teniendo en cuenta que el fichero donante debe ser de igual o mayor longitud que el receptor.

En el *método knn*, los donantes y receptores se ubican en un mismo subespacio definido por las variables comunes y para cada receptor se encuentran sus k vecinos donantes más cercanos o similares. En un punto intermedio entre los métodos anteriores, se encuentra el *método knn* con restricciones denominado como *método híbrido* (Aluja y Daunis, 2013). Las restricciones más comunes consisten en imponer que los donantes más próximos verifiquen una serie de condiciones sobre la igualdad en ciertas variables e imponer una repetición mínima de los donantes.

En el *método knn*, se trata de estimar los valores de las variables específicas Y en un punto, como media local de sus donantes vecinos. El valor a imputar consiste en calcular la mediana de la variable Y para los donantes seleccionados en la vecindad de x_i. El número de vecinos seleccionados es *k*.

$$i(x_i) \sim average\big(i(x), x \in L(x_i)\big)$$

La vecindad $L(x_i)$ define una probabilidad a priori k/n. Para muestras grandes, el número de vecinos es grande pero su proporción respecto al total de individuos en la muestra es muy pequeña $k/n \to 0$, lo que significa que la imputación se haría sin error. Para variables continuas, la media y la varianza local tienden a su valor verdadero

$$\bar{y}_0 \to E(y/x_0)$$

$$s_0^2 \rightarrow V(y/x_0)$$

3.3.3. Modelización explícita

Consiste en definir una función $r(X)$ para la predicción de las variables específicas de manera que

$$E[Y/X] = r(X) + \varepsilon$$

Establecer una relación matemática entre el conjunto de variables y las variables específicas Y se puede hacer mediante una regresión sobre el conjunto de datos donante.

La imputación por modelización explicita se trata de una imputación condicional en la que se debe especificar la forma funcional de la relación entre los bloques de datos X y Y, y no es necesario hacer supuestos sobre la naturaleza probabilística de los datos (Aluja y Daunis, 2013).

3.4. Evaluación de la calidad de la fusión

Aunque para realizar la fusión de datos se deben cumplir ciertos prerrequisitos, su cumplimiento no hace prescindible la validación del fichero obtenido, con el propósito de evaluar el potencial que tiene para realizar estimaciones confiables. Se habla en términos de validación para hacer notar que la evaluación de la calidad de la fusión debe ir más allá de la estimación del error cuadrático medio (Leulescu y Agafitei, 2013).

El procedimiento de evaluación consiste en hacer una validación empírica utilizando el mismo procedimiento de imputación de la fusión, pero sobre unos datos específicos conocidos, la matriz Y_0. Por lo tanto, la comparación que se realiza es la de Y_0 con $\hat{Y}_0$ estimado (Aluja y Daunis, 2013).

Los autores se basan 4 criterios y para su evaluación proponen uno o más estadísticos, que se describen a continuación.

A. Preservación de los estadísticos marginales

- Comparación de los estadísticos marginales de manera individual mediante test estadísticos como T de igualdad de medias, F de Fisher o Chi-cuadrado. Para hacer una prueba conjunta se define el ASL (Average Significance Value)

$$ASL = 1 - \sqrt[r]{\prod_{i=1}^{r}(1-\alpha_i)}$$

Donde α_i es el p-valor de la i-ésima prueba de hipoesis y r es el número de pruebas de hipótesis efectuadas. Entre más grande es el ASL, más similares son los estadísticos que comparamos.

B. Preservación de la distribución multivariante de los datos

- Comparación de la correlación entre las variables específicas *Y*.
- Comparación de la correlación entre las variables específicas y las variables comunes mediante el ACDe (Average Correlation Difference). Ente más pequeño el ACDe, mejor.

$$ACDe = \frac{\sum_i^t \left| cor\left(y_j^0, x_{j'}^0\right) - cor\left(\hat{y}_j^0, x_{j'}^0\right) \right|}{t}$$

- Comparación del patrón de variabilidad multivariante imputado y el patrón multivariante observado, basado en la proximidad entre las dos distribuciones a partir del posicionamiento de sus ejes de inercia. Esto se hace por medio de un análisis factorial descriptivo con la matriz Y_0, de donde se extraen los valores propios y las direcciones de inercia. Para dar a cada dirección de inercia un peso proporcional a su importancia, ponderamos las correlaciones por su respectivo valor propio calculando

$$wc = \frac{\sum_\alpha^a \lambda_\alpha \left| u'_{Y_0} u_{\hat{Y}_0} \right|}{\sum_\alpha^a \lambda_\alpha}$$

Cuanto más alto y próximo a 1 es el valor de wc, mejor.

C. Preservación de las distribuciones imputadas

- Comparación de la distribución de las variables imputadas y las observadas. Se trata de calcular el matching noise que compara la distribución de una variable observada y_j^0 con

su respectiva imputada $\hat{y}_j^0$. Dado que las distribuciones son desconocidas, la comparación se debe hacer a partir de las respectivas distribuciones empíricas.

Se puede realizar esta comparación mediante el estadístico Smirnov, que mide la distancia máxima entre las dos distribuciones empíricas y sigue una distribución normal para un cierto tamaño de muestra.

Para dar una visión global de la discrepancia, se calcula la media de la distancia de Smirnov de todas las variables específicas

$$ASD = \frac{\sum_j^q sd_j}{q}$$

Entre más pequeño sea el ASD, mejor reproducidas estarían las distribuciones reales en los datos imputados.

D. Precisión de la imputación

- Cálculo del error de predicción individual. Se trata de evaluar que tan cerca están los valores imputados $\hat{y}_{ij}^0$ de los observados y_{ij}^0. Este error se mide mediante $E[y_0 - \hat{y}_0]^2$.

Para poder interpretar esta medida de error, se da en términos relativos al error que se produciría si se utilizara la imputación por la media de la variable, usando el estadístico τ_j

$$\tau_j = \frac{\sum_i^n \left(y_{ij}^0 - \hat{y}_{ij}^0\right)^2}{\sum_i^n \left(y_{ij}^0 - \bar{y}_j^0\right)^2}$$

Finalmente, se hace la media de este estadístico para todas las variables específicas

$$\tau = {}^1\!/_r \sum_j^r \tau_j$$

Cuanto más pequeño sea este estadístico, más precisión tiene la imputación.

- Evaluación de la relevancia predictiva. Consiste en evaluar la hipótesis de ruido blanco para los residuos producidos $\varepsilon_{ij} = y_{ij}^0 - \hat{y}_{ij}^0$. Si se rechaza la hipótesis, sería indicación de que el modelo de imputación empleado $Y = i(X) + \varepsilon$ no explica toda la variabilidad de las variables específicas. Eso implica que el modelo de imputación no tiene relevancia predictiva y que las estimaciones que produce tienen sesgo.

3.5. Tasa de riesgo de pobreza

La tasa de riesgo de pobreza se calcula a partir de la renta total neta del hogar. Primero, es necesario calcular los ingresos por unidad de consumo del hogar para tener en cuenta economías de escala en los hogares. Estos se obtienen dividiendo los ingresos totales del hogar entre el número de unidades de consumo, que se calculan utilizando la escala de la OCDE modificada, que concede un peso de 1 al primer adulto, un peso de 0,5 a los demás adultos y un peso de 0,3 a los menores de 14 años. Una vez calculado el ingreso por unidad de consumo del hogar, se le asigna este valor a cada miembro del hogar.

En segundo lugar, se calcula el umbral de pobreza a partir de la distribución de los ingresos del año anterior al de la encuesta. Este umbral se fija en el 60% de la mediana de los ingresos por unidad de consumo de las personas.

Finalmente, la tasa de riesgo de pobreza es el porcentaje de personas que está por debajo del umbral de pobreza.

4. TRATAMIENTO DE LOS DATOS

Teniendo en cuenta que la reducción de las desigualdades territoriales en la distribución del ingreso se ha convertido en una prioridad para la Unión Europea en los últimos años, se hace necesario contar con estimaciones confiables de los indicadores de pobreza sobre los cuales basar las políticas que se han de adoptar (Fabrizi et al., 2005).

La fusión de datos entre la ECV y la EPA que se propone, tiene como objetivo el enriquecimiento de la base de datos de la ECV, que actúa como donante, mediante el uso de información adicional proporcionada por la EPA, que actúa como receptor, para estimar indicadores regionales de pobreza en Cataluña. El hecho de que ambas encuentras tengan la misma población objetivo y compartan algunas variables, hace que el uso complementario de la EPA sea una buena opción a considerar para tratar de mejorar las estimaciones realizadas a partir de la ECV.

4.1. Fuentes de datos

En España, la ECV es una operación estadística elaborada por el Instituto Nacional de Estadística (INE), armonizada en el ámbito europeo mediante Reglamento Comunitario. Es una encuesta anual dirigida a los hogares y cubre algunos de los principales aspectos que determinan las condiciones de vida de la población como la composición de los hogares, el trabajo, la salud y la vivienda, con el fin de recoger datos transversales y longitudinales para analizar las principales características sociales y económicas de la sociedad, entre los que se incluyen la renta, la pobreza y la exclusión social. Los principales indicadores objetivo de la ECV son: 1) Tasa de hogares riesgo de pobreza, 2) Tasa de hogares con carencia material severa, y 3) Tasa de hogares con baja intensidad en el trabajo (Instituto Nacional de Estadística (a)).

La muestra que se utiliza en este trabajo comprende los datos recogidos en Cataluña entre abril y julio de 2012 de 1390 hogares y 3535 personas, a través de un diseño de panel rotante en el que la muestra la forman cuatro submuestras independientes, cada una de las cuales es un panel de cuatro años de duración. Cada año se renueva la muestra en uno de los paneles. Para cada submuestra se realiza un muestreo bietápico con estratificación de las unidades de primera etapa y sin submuestreo en las unidades de segunda etapa. La primera etapa la conforman las secciones censales y la segunda las viviendas familiares principales.

Para la ECV, se cuenta con 4 ficheros de datos: datos básicos del hogar, datos detallados del hogar, datos básicos de la persona (adultos y menores) y datos detallados de los adultos (con 16 años o más).

La EPA es una encuesta trimestral dirigida a las familias y constituye la principal fuente de información sobre el mercado de trabajo. Se lleva a cabo en todos los países de la comunidad Europea siguiendo los estándares establecidos por la normativa comunitaria y cubre diversos aspectos sobre la participación en el mercado laboral como la situación profesional, ocupación, sector de actividad, tipo de contrato y tipo de jornada, entre otros. Entre sus principales objetivos está el de proporcionar estimaciones sobre el número de personas y las tasas de actividad en las principales categorías poblacionales en relación con el mercado de trabajo, es decir, ocupados, parados e inactivos (Instituto Nacional de Estadística (b)).

Al igual que en la ECV, en la EPA es un panel rotante, con la diferencia que se renueva por sextas partes, permaneciendo las viviendas seleccionadas en la muestra seis trimestres consecutivo. Se utiliza el muestreo bietápico con estratificación de las unidades de primera etapa (secciones censales) y no se realiza submuestreo dentro de las unidades de segunda etapa (viviendas).

Los datos que se utilizan en este trabajo corresponden a la EPA del segundo trimestre del año 2012 de Cataluña, es decir, los datos recogidos entre los meses de abril y junio, y contienen información de 6740 hogares y 17324 individuos, en un solo fichero de datos de personas (adultos y menores).

4.2. Armonización

La primera parte de la armonización de las fuentes de datos está relacionada con la población de referencia, las unidades estadísticas y los periodos de referencia.

- Población de referencia: en la ECV la población objeto de investigación son las personas miembros de hogares privados que residen en viviendas familiares principales[2]. De la misma manera, la EPA va dirigida a la población que reside en viviendas familiares principales.
- Unidad de análisis: en ambas encuestas se toman como unidades de análisis las viviendas familiares utilizadas como residencia principal y las personas residentes en las mismas, aunque existe una pequeña diferencia de concepto. En la ECV se hace referencia a los hogares privados residentes en las viviendas familiares principales utilizando la definición de presupuesto común, mientras que en la EPA se analizan las viviendas a partir de la definición de residencia. Sin embargo, análisis llevados a cabo para el caso de otros países, parece indicar que estas diferencias no tienen efectos significativos sobre la comparabilidad y, por tanto, se podrían considerar que las dos poblaciones se superponen en gran medida.

[2] Aunque todas las personas forman parte de la población objetivo, sólo los que tengan 16 años o más a 31 de diciembre del año anterior al de la entrevista, son investigadas exhaustivamente.

- Unidades de muestreo: tanto en la ECV como en la EPA, la unidad primaria de muestreo es la sección censal y la unidad última de muestreo la vivienda familiar principal.
- Períodos de referencia: el período de referencia de la información es el mismo en las dos encuestas: es la semana inmediatamente anterior (de lunes a domingo) a la de la entrevista según el calendario. En algunas preguntas se hace alusión a otros periodos de tiempo específicos como en las relacionadas con la búsqueda de empleo (las cuatro semanas anteriores a la de la encuesta) y la disponibilidad para trabajar, que se refiere a las dos semanas siguientes al domingo de la semana de referencia. En la ECV se presenta el caso especial de la variable renta, que podría generar inconsistencias ya que está definida como la renta obtenida en el año anterior al de la encuesta.

La segunda parte de la reconciliación de las fuentes está relacionada con las variables. La selección de variables comunes entre las dos encuestas se ha realizado teniendo como base las variables sociales nucleares definidas en el informe "Task Force on Core Social Variables" (Eurostat, 2007). Estas variables recogen información demográfica, geográfica y socioeconómica y se listan en la Tabla 1.

Sexo	Región de residencia
Edad	Grado de urbanización
País de nacimiento	Situación laboral
País de nacionalidad	Situación profesional
Estado civil legal	Ocupación
Estado civil de facto	Sector de actividad económica
Composición del hogar	Nivel educativo alcanzado
País de residencia	Renta neta mensual del hogar

Tabla 1. Variables sociales nucleares de Eurostat.

De estas 16 variables, se descartaron como variables comunes el país y la región de residencia porque el trabajo se centra en Cataluña, el grado de urbanización y el estado civil de facto por no aparecer entre las variables de la EPA, y la renta por ser la variable a imputar.

Además de las 11 variables restantes en esta lista, se encontraron otras variables que aparecen en ambos ficheros de datos, directa o indirectamente, y podrían afectar a los ingresos de los individuos: tamaño del hogar, persona de referencia en el hogar, tipo de jornada laboral, tipo de contrato, si está cursando estudios actualmente y cuál es el nivel de esos estudios, si ha trabajado alguna vez, si ha buscado empleo en las últimas 4 semanas y si está disponible para empezar en un nuevo trabajo en las próximas dos semanas, unidades de consumo. Estas variables comunes definidas en términos individuales, se describen en la Tabla A1 del Anexo I.

En ciertas ocasiones algunas variables necesitan ser modificadas con el fin de que sean consistentes. Con estas variables se llevó a cabo un análisis detallado sobre formulación de las preguntas en ambas encuestas, la definición de los conceptos, escalas de medición y categorías en las posibles respuestas. La variable *nivel de los estudios que está cursando* se descartó por tener categorías no armonizables y *persona de referencia del hogar* por perder sentido al realizar el análisis en términos de hogares, ya que siempre se tiene una sola persona por hogar en esta categoría. Otras variables comunes se transformaron para que fueran consistentes en términos de definiciones y/o tuvieran los mismos niveles, y algunas más se calcularon ya que no aparecían directamente en alguno de los ficheros.

Dado que la variable a predecir renta neta mensual está en términos de hogar, las variables se transforman de manera que cada categoría de respuesta se convierte en una variable de tipo "número de … en el hogar". Por ejemplo, la variable individual *sexo* se convierte en dos variables de hogar: *número de mujeres en el hogar* y *número de hombres en el hogar*. Con estas nuevas variables comunes en términos de hogares, que se describen en la Tabla A2 y A3 del Anexo I, se conforman nuevos ficheros de hogares para la ECV y para la EPA, con los que se realiza la fusión.

Para la comparación de las distribuciones marginales de las variables comunes, la literatura, especialmente la italiana, se decanta por el uso de la distancia de Hellinger (HD) para comprar las distribuciones y hacen énfasis en que si el valor de la HD es mayor de 0.05 la variable se debe descartar, aunque también reconocen que ese umbral un valor arbitrario (Leulescu et al., 2011; D'Orazio et al., 2006). Por otro lado, algunos autores consideran que la igualdad en las distribuciones no debería ser un requisito para la selección de las variables comunes ya que se trata de datos de encuestas provenientes de muestreo y se espera que existan diferencias (Aluja y Daunis, 2013). Teniendo en cuenta este último enfoque, que las distribuciones marginales sean estadísticamente iguales no se tomará como requisito, ni para utilizar las variables en el análisis ni para validar la fusión. En todo caso, se presenta en la Tabla 2 y Tabla 3 la comparación de las distribuciones de las variables comunes originales de los ficheros, que dan indicios de que los datos de ambos provienen de una misma población.

Las variables numéricas no presentan diferencias en la media pero si en la varianza, mientras que la mayoría de las categóricas tienen una distancia e Hellinger por debajo de 0.05. La variable educación es la que presenta mayor diferencia entre los ficheros con 0.096.

Variable	Diferencia medias	T test p-value	Ratio Varianzas	F test p-value
Tamaño	0.005873	0.8476	1.1028	0.00056
Ucon	-0.003500	0.7993	1.1336	0.00001

Tabla 2. Comparación de la distribución de las variables comunes numéricas originales.

	tvd	overlap	Bhatt	D. Hellinger
Sexo	0.000741	0.999259	1.000000	0.000524
Edad	0.005583	0.994417	0.999970	0.005452
País de nacimiento	0.003441	0.996559	0.999941	0.007682
País de nacionalidad	0.018771	0.981229	0.999658	0.018487
Estado civil	0.014417	0.985583	0.999734	0.016320
Situación profesional	0.015926	0.984074	0.998971	0.032080
Ocupación	0.032539	0.967461	0.998616	0.037197
Sector	0.033947	0.966053	0.999002	0.031584
Educación	0.071424	0.928576	0.990804	0.095896
Situación con relación a la actividad	0.024288	0.975712	0.999626	0.019338
Cursando estudios	0.045522	0.954478	0.997537	0.049624
Busqueda de empleo	0.019160	0.980840	0.999723	0.016635
Trabajó antes	0.020028	0.979972	0.999722	0.016685
Composición del hogar	0.070311	0.929689	0.995567	0.066583

Tabla 3. Comparación de la distribución de las variables comunes categóricas originales.

Teniendo en cuenta que las variables de ingresos se caracterizan por el truncamiento y la asimetría y están afectadas por efectos de redondeo, se le aplica la transformación logarítmica a la variable *renta neta mensual del hogar* con tal de aproximarse a la normalidad, como usualmente se hace.

Antes de aplicar el logaritmo, fue necesario sumar a toda la variable su valor mínimo mas 1 ya que presentaba valores negativos en 21 hogares.

$$renta\ hogar = renta\ neta\ mensual\ del\ hogar + (\min(neta\ mensual\ del\ hogar) + 1)$$

$$renta = log(renta\ hogar)$$

4.3. Selección de variables puente y aplicación de las técnicas de fusión

Hasta ahora se ha considerado que las variables comunes tienen capacidad explicativa potencial sobre la *renta*. Para evaluar esa capacidad se utilizó el fichero donante y se verificó que las variables estuvieran exentas de errores y valores faltantes. Para probar la hipótesis de independencia entre la variable dependiente Y y la única variable común categórica, *composición del hogar*, se realizó un ANOVA y se calculó el estadístico V de Cramer, cuyos resultados se enseñan en la Tabla 2. Habitualmente valores de V superiores a 0,30 indican que hay una posible relación entre las variables. Para buscar las variables comunes numéricas que están estadísticamente relacionadas con Y, se calculó el coeficiente de correlación de Spearman, con su respectiva prueba de hipótesis. Los resultados se resumen en la Tabla 3. La variable *n_sprof5* se ha omitido en la tabla ya que al no presentar ningún caso, no es posible calcular la correlación.

En el ANOVA, si Pr(>F) es menor que 0.05 se rechaza la hipótesis nula y se concluye que existe evidencia de una posible relación de dependencia entre las variables. En el caso de los coeficientes

de correlación, si el p-value es menor que 0.05, se rechaza la hipótesis nula de que el coeficiente es estadísticamente igual a cero. Para visualizar cuáles son las variables comunes con mayor correlación con la variable a imputar, se presenta la Figura 3.

Variable	F value	Pr(>F)	V de Cramer
Composición	7.522	0.00000	0.9273632

Tabla 2. Prueba de independencia y V de Cramer X y Y en el fichero donante.

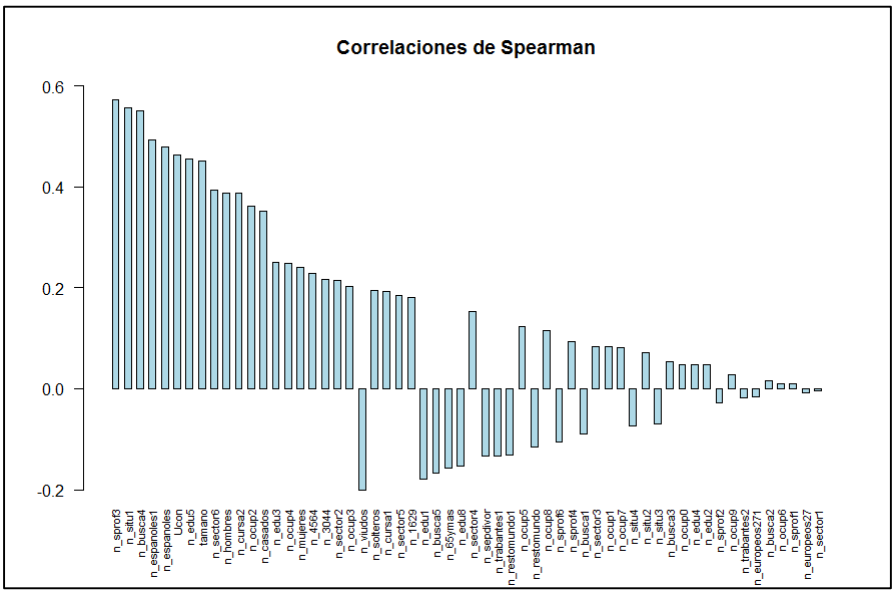


Figura 3. Correlación de Spearman entre X y Y en el fichero donante.

La V de Cramer y el ANOVA confirman que existe relación entre las variables *composición* y *renta*. La correlación de Spearman indica que 47 de las 59 variables comunes numéricas, están relacionadas con la variable específica. Aunque esta medida de asociación tiene la ventaja de ser apropiada para variables discretas y no normales, al evaluar la relación de Y con las variables comunes X una por una, no tiene en cuenta los problemas de colinealidad. Además se comprobó que los resultados de una regresión con las variables seleccionadas de esta manera no son los mejores.

Para evaluar esa capacidad explicativa de las variables comunes de manera conjunta y seleccionar las variables que se van a incluir en el modelo, se llevó a cabo una regresión stepwise. Se introdujeron las 60 variables comunes en el stepwise y el proceso seleccionó 16. El coeficiente $R^2 = 0.196$ es consistente con el de otros estudios similares en los que se ha aplicado el stepwise para seleccionar variables para un modelo con datos del ECV (Parra, 2014).

Variable	ρ Spearman	p-value
Tamaño	0.451515212	0.00000
n_hombres	0.388336402	0.00000
n_mujeres	0.241192990	0.00000
n_1629	0.180849571	0.00000
n_3044	0.216722697	0.00000
n_4564	0.229137104	0.00000
n_65ymas	-0.157192252	0.00000
n_espanoles	0.479368064	0.00000
n_europeos27	-0.008241405	0.75885
n_restomundo	-0.114782085	0.00002
n_espanoles1	0.492501558	0.00000
n_europeos271	-0.016071961	0.54937
n_restomundo1	-0.132070775	0.00000
n_solteros	0.195214823	0.00000
n_casados	0.351234899	0.00000
n_viudos	-0.200285983	0.00000
n_sepdivor	-0.133654835	0.00000
n_sprof1	0.010377571	0.69908
n_sprof2	-0.028624686	0.28621
n_sprof3	0.572463203	0.00000
n_sprof4	0.092734283	0.00054
n_sprof6	-0.106093784	0.00007
n_ocup0	0.048660413	0.06973
n_ocup1	0.084045211	0.00171
n_ocup2	0.361109291	0.00000
n_ocup3	0.203068755	0.00000
n_ocup4	0.248667462	0.00000
n_ocup5	0.122407214	0.00000
n_ocup6	0.010756741	0.68865
n_ocup7	0.080846934	0.00256
n_ocup8	0.114715811	0.00002
n_ocup9	0.027318634	0.30878
n_sector1	-0.003175448	0.90584
n_sector2	0.215025053	0.00000
n_sector3	0.084326905	0.00165
n_sector4	0.152329661	0.00000
n_sector5	0.185153332	0.00000
n_sector6	0.393174007	0.00000
n_edu1	-0.179180435	0.00000
n_edu2	0.047249054	0.07824
n_edu3	0.250569166	0.00000
n_edu4	0.047731129	0.07525
n_edu5	0.455143703	0.00000
n_edu8	-0.152656141	0.00000
n_situ1	0.555857002	0.00000
n_situ2	0.071061639	0.00804
n_situ3	-0.070590088	0.00847
n_situ4	-0.072877435	0.00656
n_cursa1	0.193462263	0.00000
n_cursa2	0.388165838	0.00000
n_busca1	-0.089839002	0.00080
n_busca2	0.015524965	0.56304
n_busca3	0.054549414	0.04201
n_busca4	0.549941089	0.00000
n_busca5	-0.166411756	0.00000
n_trabantes1	-0.133303699	0.00000
n_trabantes2	-0.017216729	0.52129
Ucon	0.462553423	0.00000

Tabla 3. Coeficiente de correlación entre X y Y en el fichero donante.

Para entender mejor las relaciones entre las variables del fichero donante, se presenta en la Figura 4 los gráficos del análisis de conglomerados jerárquico sobre las variables. Primero se realizó sobre todas las variables (a), con el fin de evaluar la colinealidad y redundancia de las variables. Y luego, sobre aquellas seleccionadas por el stepwise (b) para contrastar la robustez de la elección de los predictores.

Asumiendo el supuesto de relevancia predictiva, en primer lugar se planteó como método de modelización explicita, una regresión múltiple, teniendo en cuenta los factores de elevación, para hallar un modelo con el que se pudiera predecir la renta de los hogares en el fichero de la EPA. Los resultados son presentados en la Tabla 4.

Weighted Residuals:

Min	1Q	Median	3Q	Max
-11.2705	-0.0813	-0.0111	0.0770	1.2600

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.96411	0.02442	448.956	< 2e-16 ***
n_1629	-0.04331	0.01852	-2.339	0.019501 *
n_4564	-0.03889	0.01333	-2.917	0.003596 **
n_europeos27	-0.19264	0.02744	-7.021	3.44e-12 ***
n_restomundo	-0.06239	0.01219	-5.119	3.51e-07 ***
n_sprof2	-0.19988	0.03332	-5.998	2.55e-09 ***
n_sprof3	0.06569	0.02155	3.049	0.002344 **
n_sprof6	0.09453	0.01446	6.539	8.70e-11 ***
n_sector1	0.08831	0.05877	1.503	0.133168
n_sector2	0.08183	0.03104	2.636	0.008473 **
n_sector3	0.15220	0.04484	3.394	0.000708 ***
n_sector4	0.04885	0.02756	1.773	0.076505 .
n_sector6	0.11507	0.02598	4.429	1.02e-05 ***
n_edu3	0.05722	0.01572	3.639	0.000284 ***
n_edu4	0.15498	0.09660	1.604	0.108869
n_edu5	0.06018	0.01422	4.233	2.46e-05 ***
n_situ1	0.05462	0.0227	2.404	0.016347 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

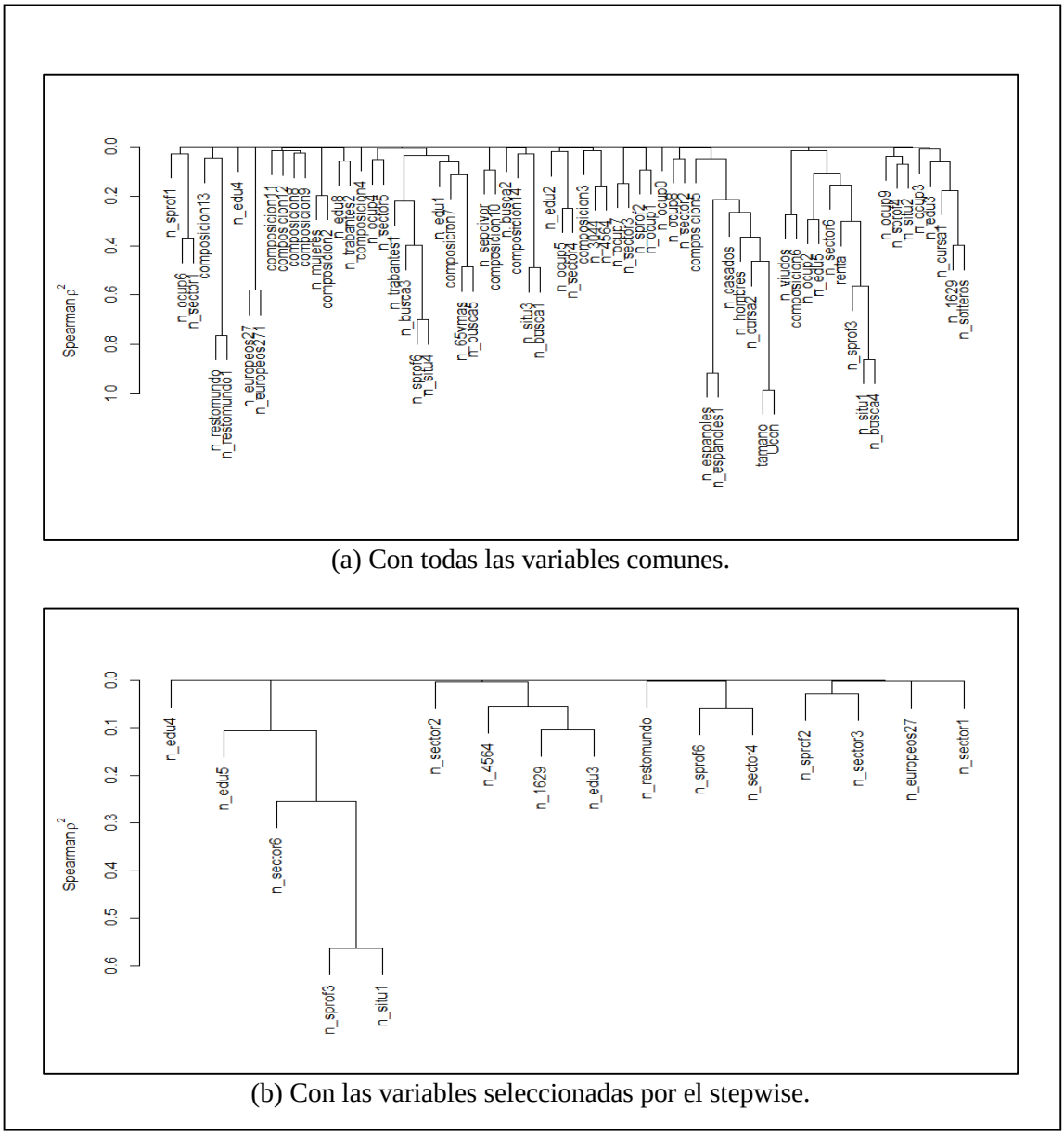
Residual standard error: 0.348 on 1373 degrees of freedom
Multiple R-squared: 0.2053, Adjusted R-squared: 0.196
F-statistic: 22.17 on 16 and 1373 DF, p-value: < 2.2e-16

Tabla 4. Regresión de la *renta* sobre las variables puente seleccionadas

Con el modelo de regresión, se realizó la predicción de la variable *renta* con los datos de los hogares de la EPA. La descripción de la renta observada en la ECV y la renta imputada en la EPA se muestra en la Tabla 5.

	Mean	SE	Min	1st. Qu.	Median	3rd. Qu.	Max
renta observada ECV	11.256	0.0080	0.0000	11.1099	11.2516	11.3874	12.3875
renta imputada EPA	11.249	0.0019	10. 5307	11.1159	11.2320	11.3795	12.0759

Tabla 5. Estadísticos descriptivos de renta observada e imputada por *regresión.*



(a) Con todas las variables comunes.

(b) Con las variables seleccionadas por el stepwise.

Figura 4. Análisis de conglomerados jerárquico sobre las variables

Para comparar la renta observada vs. la imputada, es necesario utilizar el fichero de las personas y no el de los hogares, ya que en la EPA no se cuenta con factores de elevación para los hogares. Para ello, fue necesario asignar a cada persona la renta de su respectivo hogar en el fichero de individuos, tanto en la ECV (la renta observada) como en la EPA (la renta imputada). En la Figura

5 se muestra la comparación de la distribución de la renta observada en la ECV (D) y la imputada en la EPA (R) mediante un boxplot.

Se realizaron los test estadísticos T para la igualdad de medias y F de Fisher para la igualdad de varianzas que se presentan en la Tabla 7. El test T no rechaza la hipótesis nula, sugiriendo que la media de la renta imputada en la EPA no es diferente a la media de la renta observada en la ECV. Sin embargo, el test F sí rechaza la hipótesis nula, esto quiere decir que el verdadero ratio entre las varianzas no es igual a 1 y por tanto la varianza de la renta observada difiere de la varianza de la renta imputada. Posiblemente esta diferencia se deba a la presencia de un outlier cuya renta del hogar observada es cero, mientras que en la EPA el rango de valores es menor ya que no se tiene cero en ninguna imputación.

Otra forma de comparar la distribución de la renta observada en la ECV y la imputada en la EPA es mediante las funciones de distribución acumulada y la diferencia máxima entre las ellas. En la Figura 6 se puede ver las gráficas de la función de distribución acumulada de la ECV y de la EPA al lado izquierdo, y el gráfico de las diferencias a la derecha. La máxima diferencia es de 0.0321 y se presenta cuando la variable *renta* toma el valor 11.52.

T test			**F test**			
t	df	p-value	F	num df	denom df	p-value
-0.8041	17233	0.4214	3.7939	2920	14313	< 2.2e-16
Difference in mean:			Ratio of variances:			
-0.0066279			3.793863			

Tabla 6. Comparación de medias y varianzas de la renta en ECV y EPA por Regresión.

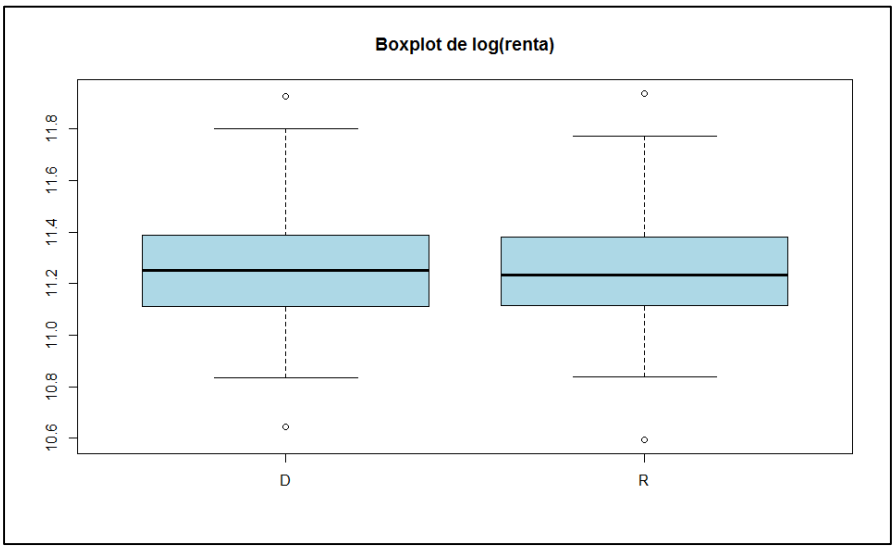


Figura 5. Boxplot de variable específica *renta* en el fichero fusionado por *regresión*.

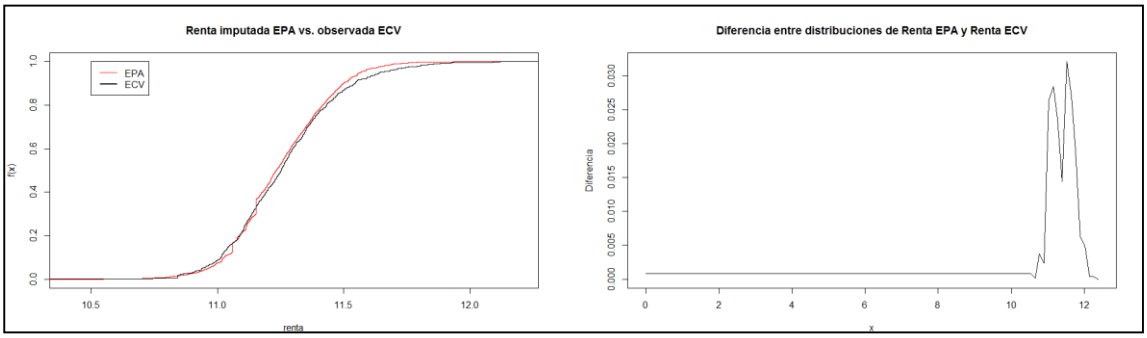


Figura 6. Comparación de la renta observada en ECV y la renta imputada en EPA por *regresión*.

Es importante hacer notar que para realizar la imputación de la renta, el método de regresión calcula valores a partir de la predicción, lo que implica tener la desventaja de la regresión hacia la media y la sensibilidad a los errores en la especificación del modelo (Leulescu y Agafite, 2013).

El segundo método propuesto para la imputación de la renta en la EPA es de tipo hot deck. El método *knn* implementado buscó en el fichero de la ECV el vecino más cercano para cada registro en la EPA de acuerdo a la distancia de Manhattan calculada sobre las variables puente seleccionadas y le asignó el valor de la renta observado en dicho registro de la ECV. Dado que el fichero de la EPA contiene más observaciones que el de la ECV, los registros de la ECV son usados como donantes más de una vez. Este método tiene la ventaja de imputar valores realmente observados pero también tiene una desventaja y es que el uso múltiple de los mismos donantes tiende a influenciar la varianza (Leulescu y Agafitei, 2013).

De la misma manera que se hizo para el método de regresión, se muestran ahora los estadísticos descriptivos de la renta observada y la renta imputada por el *knn* en la Tabla 7, y se compara la renta observada con la imputada por medio del boxplot de la Figura 8 y las funciones de distribución acumuladas en la Figura 9. La distancia máxima entre la renta de la ECV y de la EPA estimada por *knn*, se da en el punto donde la renta toma el valor de 11.27 y es de 0.0174, menor a la de la regresión. Los resultados de los test de medias y varianzas se presentan en la Tabla 8.

	Mean	**SE**	**Min**	**1st. Qu.**	**Median**	**3rd. Qu.**	**Max**
renta observada ECV	11.256	0.008	0.0000	11.1099	11.2516	11.3874	12.3875
renta imputada EPA	11.259	0.002	10.64545	11.1052	11.2440	11.3764	12.3875

Tabla 7. Estadísticos descriptivos de renta observada e imputada por *knn*.

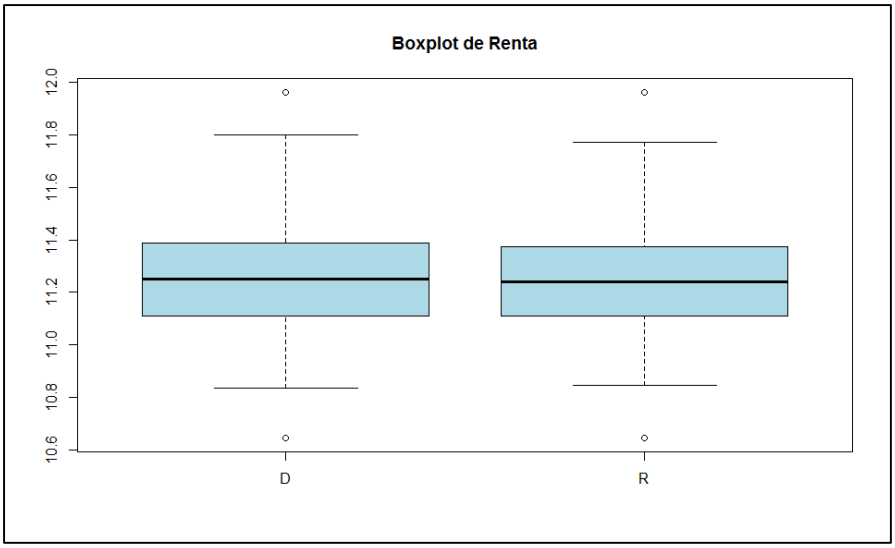


Figura 8. Boxplot de variable específica *renta* en el fichero fusionado por *knn*.

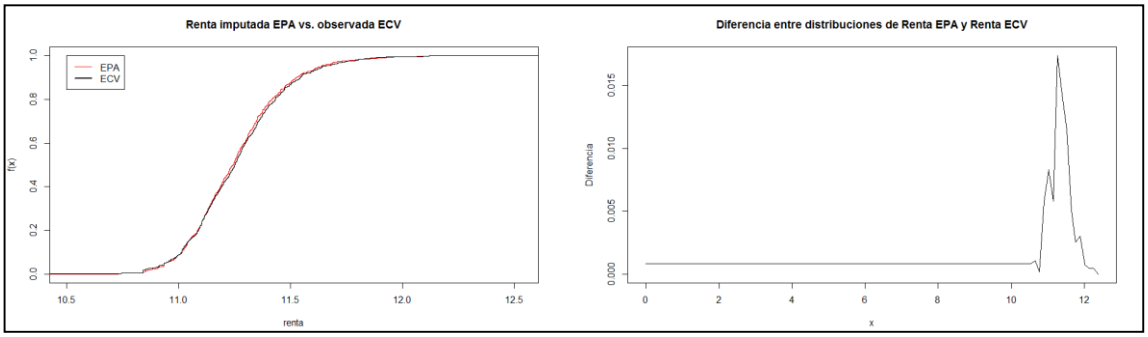


Figura 9. Comparación de la renta observada en ECV y la renta imputada en EPA por *regresión*.

T test			F test			
t	df	p-value	F	num df	denom df	p-value
0.3481	17233	0.7278	2.96	2920	14313	2.2e-16
Difference in mean:			Ratio of variances:			
0.0028854			2.960021			

Tabla 8. Comparación de medias y varianzas de la renta en ECV y EPA por *knn*.

5. EVALUACIÓN DE LA CALIDAD DE LA FUSIÓN

En este apartado se analizan los resultados obtenidos a través de los dos métodos presentados anteriormente y se evalúa su desempeño mediante el proceso de validación adaptado a la particularidad de los datos tratados y el tipo de variables utilizadas.

Para realizar la validación empírica de la fusión, se utilizó el modelo hallado y se hizo la predicción de la renta con los mismos datos de los hogares de la ECV, para comparar los datos estimados con los realmente observados. De esta forma se evalúan los métodos utilizados.

5.1. Preservación de los estadísticos marginales

La comparación de los estadísticos marginales para la variable específica, que permite evaluar el insesgamiento de la imputación, se hace mediante la comparación entre la renta observada y la renta imputada en la misma ECV, en el fichero de hogares. Los estadísticos descriptivos de la variable *renta* observada e imputada mediante ambos métodos, se presentan en la Tabla 11. Para visualizar mejor los resultados, se presentan los gráficos de la Figura 10 para el caso de la regresión y la Figura 11 para el *knn*.

	Mean	**SE**	**Min**	**1st. Qu.**	**Median**	**3rd. Qu.**	**Max**
renta observada	11.218	0.0116	0.0000	11.0843	11.2020	11.3515	12.3875
renta Regresión	11.218	0.0055	10.3797	11.0957	11.1920	11.3291	12.0481
renta knn	11.225	0.0048	10.8773	11.1094	11.1976	11.3200	11.8648

Tabla 11. Estadísticos descriptivos de renta observada e imputada en ECV.

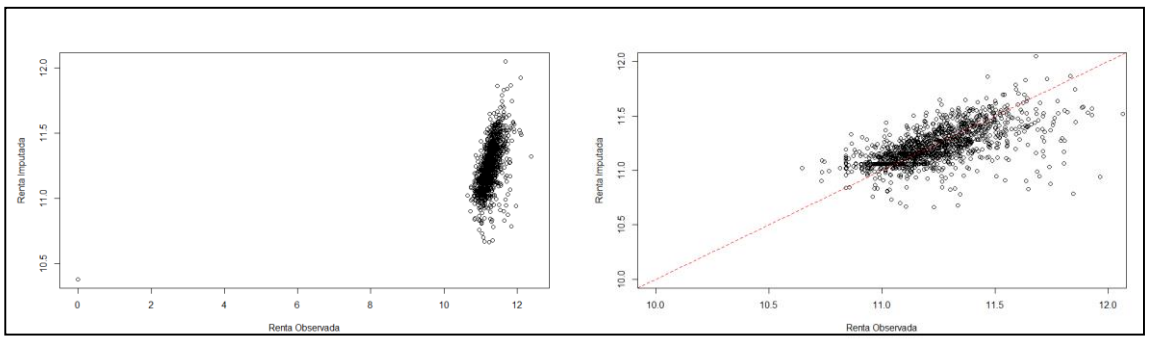


Figura 10. *renta* observada vs. *renta* imputada por *regresión*.

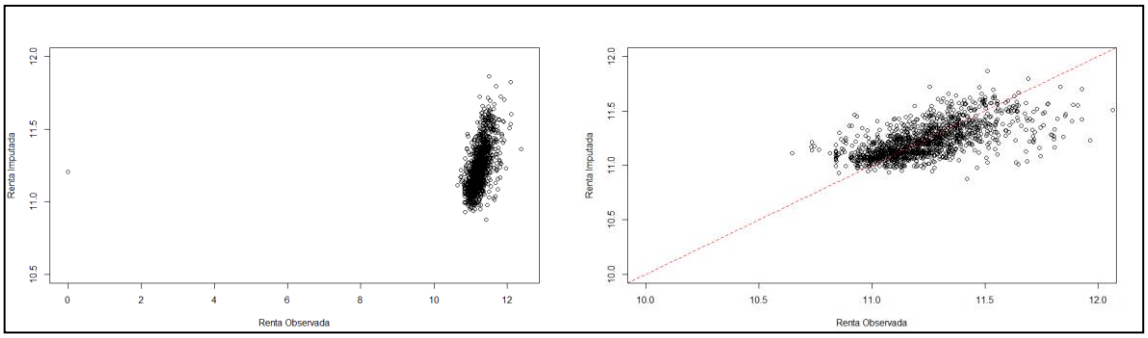


Figura 11. *renta* observada vs. *renta* imputada por *knn*.

Los resultados que se muestran en la Tabla 12, sobre los test de igualdad de medias y de varianzas, indican que no existen diferencias entre la media observada e imputada para ninguno de los dos métodos, pero si hay diferencias en las varianzas, aunque el ratio de varianzas es menor por *regresión*.

	T test		F test	
	p-value	**Diff. in mean**	**p-value**	**Ratio of variances**
Regresión	1	0.00000	0.0000	4.476408
Knn	0.5183	-0.00702	0.0000	5.606804

Tabla 12. Comparación de medias y varianzas de la renta observada y la imputada en la ECV

5.2. Preservación de la distribución multivariante de los datos

Para realizar la comparación de la correlación entre las variables específicas y las variables comunes, se calcula la correlación de cada una de las variables comunes con la variable *renta*. En la sección de armonización ya se había calculado el coeficiente de correlación entre las variables comunes numéricas y la renta observada. Ahora, se calcula este coeficiente entre las variables comunes numéricas y la renta imputada en la ECV. Las diferencias entre las primeras correlaciones calculadas y las segundas, se utilizan para calcular el indicador ACDe.

El estadístico denominado ACDe resume las correlaciones de las variables comunes con la variable específica, para poder decir en términos generales si la distribución multivariante de los datos se mantiene para los datos imputados. El ACDe se calcula a partir del valor absoluto de las diferencias en las correlaciones. Para incluir en el estadístico a todas las variables comunes, se proponen dos opciones respecto a la variable categórica: 1) "cuantificarla" haciendo variables dummy con cada categoría y calculando las correlaciones, o 2) calcular el valor de la V de Cramer que toma valores entre 0 y 1, y utilizarlo como los valores absolutos de las correlaciones utilizados para calcular el ACDe. Los resultados de los diferentes cálculos se presentan en la Tabla 13.

Por el método de *regresión*, la relación entre la variable composición y la renta medida por la V de Cramer se conserva mejor al hacer imputación por *knn*. Esto no se mantiene al medir la relación por medio de las correlaciones de la variable categórica "cuantificada".

Sea cual sea la forma de introducir la variable categórica en el ADCe que incluye todas las variables, el valor del indicador es menor por el método de *regresión*, lo que indica que se preserva mejor la distribución multivariante de los datos al imputar por el método explícito.

	Var. numéricas	**Var. Categórica V de Cramer**	**Var. Categórica "cuantificada"**	**Todas las variables (con V de Cramer)**	**Todas las variables (con cuantificación)**
Regresión	0.111335	0.044530	0.062843	0.110203	0.101906
Knn	0.131789	0.017080	0.076024	0.129845	0.120946

Tabla 13. ACDe

5.3. Preservación de las distribuciones imputadas

La comparación de la distribución de la variable imputada y la observada se hace mediante el cálculo del matching noise, que compara la distribución de una variable observada y_j^0 con su respectiva imputada $\hat{y}_j^0$. Dado que la renta no se distribuye Normal el uso de un test no paramétrico es conveniente. El test de Kolmogorov-Smirnov prueba si las dos variables, la renta observada y la renta imputada en ECV, tienen la misma distribución. El resultado del test en ambos métodos presentado en la Tabla 14, sugiere que la distribución de la renta observada en la ECV difiere de la distribución de la renta imputada con los mismos datos de la ECV.

	D	**P-value**
Regresión	0.1043	5.394e-07
Knn	0.1187	6.234e-09

Tabla 14. Test Kolmogorov-Smirnov.

Se presenta, además, la comparación de las funciones de distribución acumuladas y su máxima diferencia en las Figuras 12 y 13 para la imputación por *regresión* y por *knn*, respectivamente. La distancia máxima por el método explícito es de 0.0626 y 0.1141 en el *knn*, ambos en el punto donde la renta alcanza el valor de 11.025.

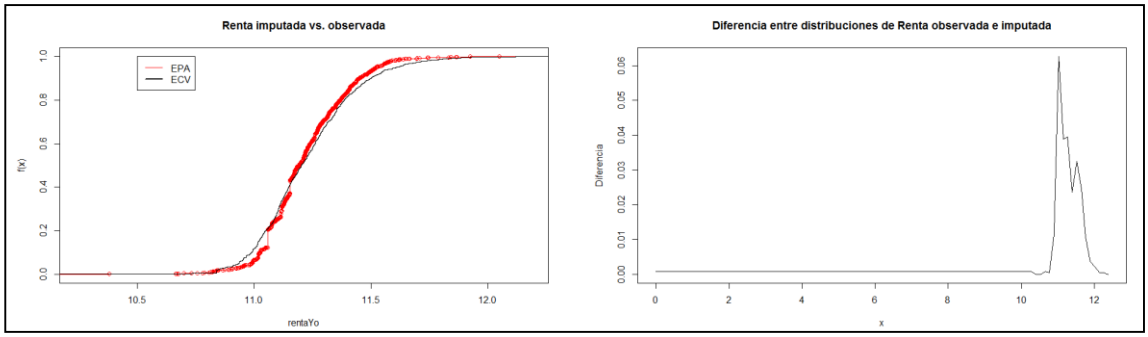


Figura 12. Comparación de *renta* observada en ECV y *renta* imputada en ECV por *regresión*.

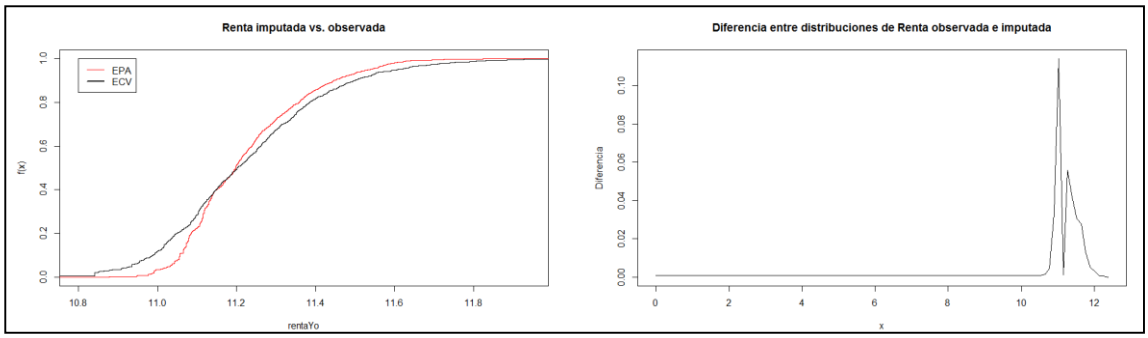


Figura 13. Comparación de *renta* observada en ECV y *renta* imputada en ECV por *knn*.

5.4. Error de predicción individual

El error de predicción individual, que mide qué tan cerca están los valores imputados $\hat{y}_{ij}^0$ de los observados y_{ij}^0, es $\tau_R = 0.794695$ en la regresión y $\tau_{KNN} = 0.887592$ para el *knn*. Esta medida está dada en términos relativos al error que se produciría si se utilizara la imputación por la media de la variable. El resultado en ambos casos menor que 1, indica que el error de la imputación realizada es menor que si se hubiera imputado por la media, pero el valor de τ más cercano a 1 por el método explícito, indica que la imputación a partir del modelo de regresión es más precisa que la del *knn*.

5.5. Evaluación de la relevancia predictiva

Se evalúa la hipótesis de ruido blanco para los residuos producidos $\varepsilon_{ij} = y_{ij}^0 - \hat{y}_{ij}^0$ mediante el test de normalidad de Shapiro-Wilk. El resultado que se muestra en la Tabla 15, al rechazar la hipótesis nula, pone en evidencia que el modelo de imputación empleado $Y = i(X) + \varepsilon$ no explica toda la variabilidad de las variables específicas y los residuos al no comportarse como ruido blanco indican que la renta imputada, tanto por el método de regresión como por el *knn*, tiene sesgo.

	W	P-value
Regresión	0.3245	<2.2e-16
Knn	0.2966	<2.2e-16

Tabla 15. Test Shapiro-Wilk.

5.6. Estimación de la tasa de riesgo de pobreza

Para calcular la tasa de riesgo de pobreza, es necesario volver a los ficheros de todas las personas, adultos y menores. La tasa se calculó para los datos observados en la ECV y los imputados, por *regresión* y *knn*, en la ECV, la EPA y el fichero fusionado.

Los resultados de las estimaciones se presentan en la Tabla 16 y las Figuras 14 y 15. Por el método de *regresión*, el fichero fusionado proporcionó una tasa de riesgo de pobreza más cercana a la observada. De igual manera, por el *knn* la tasa estimada a partir del fichero fusionado fue la más cercana a la observada, siendo iguales.

		Tasa de riesgo de pobreza	SE
	Observada	0.201	0.0080
Regresión	Imputada en ECV	0.156	0.0070
	Imputada en EPA	0.178	0.0035
	Fichero fusionado	0.190	0.0044
Knn	Imputada en ECV	0.150	0.0070
	Imputada en EPA	0.202	0.0036
	Fichero fusionado	0.201	0.0044

Tabla 16. Tasa de riesgo de pobreza a partir de renta observada e imputada.

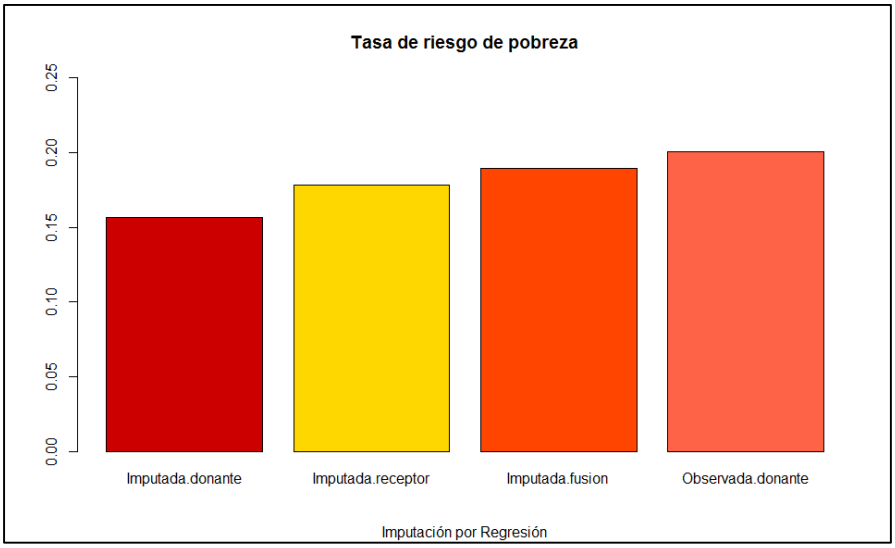


Figura 14. Tasa de riesgo de pobreza a partir de renta observada e imputada por *regresión*.

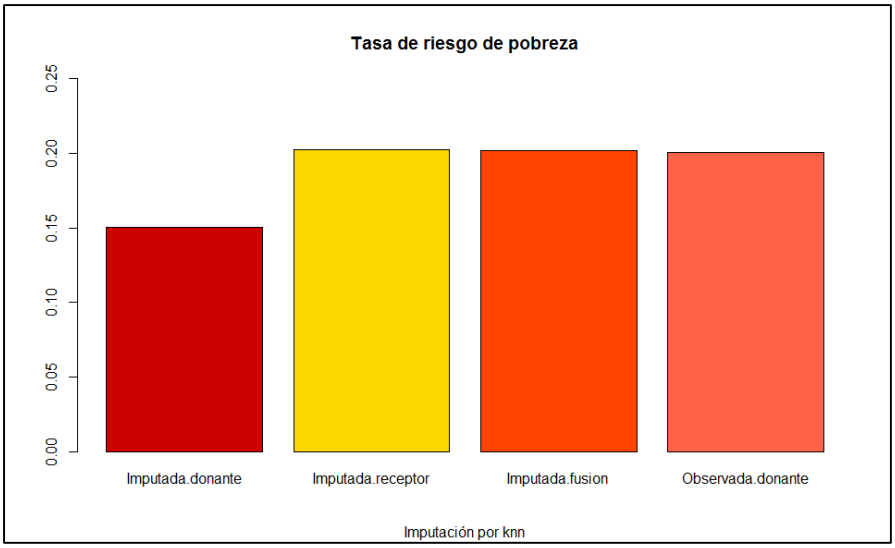


Figura 15. Tasa de riesgo de pobreza a partir de renta observada e imputada por *knn*.

Para ambos métodos el error estándar de la estimación fue menor cuando se utilizaron los datos de la renta imputada en la EPA y el fichero fusionado, posiblemente debido al gran tamaño muestral. También hay que tener en cuenta que el modelo de *regresión* deja a un lado el outlier para hacer la predicción de la renta y el *knn* no utiliza esta observación como donante para ningún registro en la EPA. Este es un hecho que cuenta a favor de la estimación por medio de la fusión, ya que no existe un "outlier reciproco" que profundice la afectación de las estimaciones.

A partir de la imputación de la renta realizada en la EPA, se calcularon además las tasas de pobreza por provincias que se presentan en la Tabla 17, y se comparan con las calculadas por la Encuesta de Condiciones de Vida y Hábitos de la Población (ECVHP) 2011. Esto no es posible aun con la ECV ya que el fichero proporcionado por el INE no cuenta con la información de las provincias y ciudades.

	Barcelona	**Girona**	**Lleida**	**Tarragona**
Observada	0.2026	0.2553	0.2187	0.2968
Regresión	0.1696	0.2104	0.1941	0.2063
Knn	0.1968	0.2407	0.2014	0.2167

Tabla 17. Tasa de riesgo de pobreza provincial observada en ECVHP e imputada en EPA.

Las tasas estimadas por *knn* en la EPA se aproximan más a las observadas en la ECVHP, que las imputadas por *regresión*. Aunque para la provincia de Tarragona, la diferencia es mucho mayor que para las demás. En cualquier caso, las tasas denominadas observadas son estimaciones también, lo que dificulta hacer comparaciones.

6. CONCLUSIONES

En los últimos años, el Sistema Estadístico Europeo (ESS) se encuentra inmerso en la búsqueda de mecanismos y metodologías que contribuyan a la viabilidad y el fortalecimiento de estadísticas oficiales a través del uso intensivo y combinado de múltiples fuentes de información. Con este fin, la integración de datos a través de statistical matching ha sido el punto de mira que se ha estado investigando en pro de la eficiencia del sistema.

En este contexto, el presente trabajo llevó a cabo una fusión de datos entre la Encuesta de Condiciones de Vida (ECV) y la Encuesta de Población Activa (EPA) para Cataluña, con el fin de enriquecer el fichero de datos de la ECV y hacer posible la estimación de indicadores regionales, consolidando los métodos de reutilización y aprovechamiento de los ficheros para producir estadísticas de calidad, que reduzcan los costes y cargas de los entrevistados, siguiendo las líneas estratégicas europeas.

El ejercicio aquí implementado supone un avance para el sistema estadístico catalán, tanto en la aplicación de estas metodologías como en la obtención de indicadores regionales, que abre las puertas a otras posibles aplicaciones que busquen la eficiencia mediante la aplicación de técnicas de integración de datos. Es la primera vez que se realiza una fusión de datos entre dos encuestas realmente diferentes en el Idescat. En algunos estudios anteriores se trabajó la integración de datos de otras maneras que no requirieron una labor de armonización tan profunda. Por ejemplo, en el estudio llevado a cabo con la Encuesta de Seguridad Pública, se trataron diferentes olas de la misma encuesta. En la fusión de datos entre la Encuesta de Salud de Cataluña (ESCA) y el Examen de Salud, los datos de este último se obtuvieron a través de una submuestra de los respondientes de la propia ESCA.

Esta fusión entre la ECV y la EPA implementada íntegramente en R, tal y como se ha estructurado, permitirá escalar fácilmente el proceso al análisis de otras ediciones de estas encuestas, ampliar los resultados, implementar más métodos de fusión y validación, entre otras posibilidades. Además, permitió hacer una estimación del coste, en horas, de hacer una fusión de datos con encuestas no armonizadas.

Teniendo en cuenta la importancia de la estrategia Europa 2020, cabe destacar que esta fusión incide en unos indicadores de máximo interés, impacto y trascendencia en todos los niveles territoriales. Así pues, siendo una fusión de datos entre dos encuestas complejas, con dos niveles de observación (hogares e individuos) y teniendo en cuenta los condicionantes del diseño muestral mediante el uso de los factores de elevación, puede considerarse como un paso importante que se da en Cataluña en línea con las últimas tendencias de investigación en Europa.

Para realizar la fusión, primero fue necesario armonizar los ficheros de datos. El proceso de armonización de las fuentes de datos es un reto de mucho cuidado que requirió aproximadamente el 70% del tiempo empleado en el proyecto. Para analizar la coherencia de las fuentes de datos no solo hay que buscar documentación sobre las variables y analizar los datos, sino también indagar sobre la metodología, los aspectos relacionados con el muestreo y las definiciones utilizadas.

Tomar algunas medidas para mejorar la preparación, coordinación e implementación de las encuestas, proporcionaría más oportunidades para el aprovechamiento de los datos recogidos a través de las técnicas de integración de datos. Acciones como homogenizar los conceptos en el marco de las encuestas sociales al momento de la planificación, generarían más ocasiones de análisis mediante fusión de datos. Algunas veces el trabajo de armonización se dificulta debido a que hay conceptos que básicamente son equivalentes, pero se han expresado de manera diferente y esto podría afectar la respuesta de los encuestados. También, a causa de categorizaciones diferentes para una misma variable, es posible no hallar una estructura común, obligando al investigador a descartar variables.

El ejercicio de imputación realizado generó valores sesgados posiblemente debido a que se hizo una imputación simple que no considera la incertidumbre del modelo. Debe, por lo tanto, tenerse en cuenta la aplicación de una imputación estocástica en un trabajo futuro. También hay que tener en cuenta que la ECV indaga sobre los ingresos del hogar percibidos el año natural inmediatamente anterior al de la entrevista. Esto puede introducir errores donde las condiciones del hogar hayan cambiado.

Otro aspecto que no se debe pasar por alto es que uno de los supuestos de los modelos lineales es la independencia entre las observaciones. Los métodos utilizados no tienen en cuenta en primer lugar que las respuestas dentro de los hogares probablemente están correlacionadas creando dependencia, y por otro lado la estructura de dependencia regional ignorando el contexto. Esto abre las puertas a la introducción de modelos multinivel en la aplicación de fusión de datos. Análisis de este tipo serían posibles si se contara con las variables de diseño muestral (secciones censales y estratos) que permitieran mejorar las estimaciones.

La evaluación de la calidad de la fusión contempla diferentes aspectos relacionados con la calidad y coherencia de las fuentes, el poder explicativo de las variables, las técnicas de fusión y los métodos utilizados para calcular estimaciones basadas en los ficheros fusionados. Para definir cuál método debe ser utilizado para un proceso de fusión de datos, se debe tener en cuenta la particularidad de los datos disponibles y el objetivo del análisis que se desea realizar partir de los datos fusionados. En algunos casos será más conveniente utilizar uno u otro dependiendo de cuál sea la prioridad, por ejemplo, la precisión en la imputación o la homogeneidad de los datos. Sin

embargo, no hay que olvidar que estos valores imputados son estimaciones, no valores observados y realizar un análisis de la incertidumbre será necesario.

En la evaluación de la calidad de la fusión, respecto a la preservación de los estadísticos marginales, la distribución imputada y la evaluación de la relevancia predictiva, ambos métodos utilizados tuvieron desempeño similar. Aunque las diferencias en los otros puntos evaluados no son muy grandes, el método de regresión presento mejores estadísticos en la evaluación de la preservación de la distribución multivariante de los datos y el error de predicción individual.

Entre las limitaciones encontradas, por un lado se tiene las inherentes a los métodos usados, como la desventaja de la regresión hacia la media y la sensibilidad a los errores en la especificación del modelo en la *regresión*. El uso múltiple de los mismos donantes que tiende a influenciar la varianza y la escasez de donantes respecto a los receptores, son desventajas en el *knn*. Por otro lado, está la limitación de los ficheros disponibles en los que no se tienen todas las variables deseadas, y el poder explicativo de las variables comunes. Aunque se intentara mejorar el modelo con efectos aleatorios, el hecho de que las variables no expliquen lo suficiente no haría posible que se obtengan mejoras relevantes.

De cualquier forma, el estudio llevado a cabo permite visualizar posibles mejoras en el proceso. En este caso se implementaron dos métodos diferentes pero podrían aplicarse otros. También podría extenderse el análisis para ver hasta qué punto el modelo es capaz de preservar la distribución multivariante de la renta con otras variables no utilizadas en el modelo y, siendo más ambiciosos, buscar la producción de estadísticas conjuntas sobre la renta y el empleo, analizando las relaciones entre la renta y otras variables específicas de la EPA. En el caso particular de la renta, el acceso a registros administrativos sería una buena oportunidad para mejorar de la calidad de los datos.

Aunque el tema aún continúa siendo objeto de investigación, está claro que las técnicas de fusión de datos son una buena oportunidad para ampliar la disponibilidad de datos y las posibilidades de análisis, sin incrementar la carga de los informantes ni los costos. Incluso podría facilitar el desarrollo de nuevos indicadores.

7. BIBLIOGRAFÍA

Alter, H. E. (1974). Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey. *Annals of Economic and Social Measurement*, *3*(2), 373-397.

Aluja, T., Dauinis i Estadella, J., & Ripoll, E. (2009). Assessing the uncertainty in knn Data Fusion. *9th Journées Francophones Extraction et Gestion des Connaissances (EGC 2009)*. Strasbourg, enero de 2009.

Aluja, T., & Daunis i Estadella, J. (2012). Seminari d'integració de dades: enllaç de registres i fusió de dades. Barcelona, noviembre de 2012.

Aluja, T., & Daunis i Estadella, J. (2013). Compendi de fusió de dades. Manual intern de Fusió de dades. Barcelona: Idescat.

Aluja, T., Dauinis i Estadella, J., & Pellicer, D. (2007). GRAFT, a complete system for data fusion. *Computational Statistics & Data Analysis*, *52*(2), 635-649.

Budd, E. C. (1971). The Creation of a Microdata File for Estimating the Size Distribution of Income. *Review of Income and Wealth*, *17*(4), 317-334.

Comisión Europea (2009). Más allá del PIB. Evaluación del progreso en un mundo cambiante. Comunicación de la Comisión al Consejo y al Parlamento Europeo. Bruselas, 20.8.2009. COM(2009) 433. Recuperado el 07 de enero de 2015, de http://www.europarl.europa.eu/meetdocs/2009_2014/documents/com/com_com%282009%290433_/com_com%282009%290433_es.pdf

Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M., & Spaziani, M. (2014). Statistical Matching of Income and Consumption Expenditures. *International Journal of Economic Sciences*, *3*(3), 50-65.

D'Orazio, M., Di Zio, M., & Scanu, M. (2001). Statistical Matching: a tool for integrating data in National Statistical Institutes. *Second International Seminar of Exchange of Technology and Know-How/Fourth New Techniques and Technologies for Statistics Seminar*. Creta, junio de 2001.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Chichester: John Wiley & Sons.

Eurostat. (2007). Task Force on Core Social Variables. *Methodologies and Working papers*. Luxemburgo: Comisión Europea.

Eurostat. (2013). Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation. *Methodologies and Working papers.* Luxemburgo: Comisión Europea.

Fabrizi, E., Ferrante, M. R., Pacei, S. (2005). Estimation of poverty indicators at Sub-national level using multivariate small area models. *Statistics in transition, 7*(3), 587-608.

Gazzelloni, S., Romano, M. C., Corsetti, G., Di Zio, M., Pintaldi, F., Scanu, M., & Torelli, N. (2008). Time Use and Labour Force: a proposal to integrate the data through statistical matching. En *Time Use in Daily Life*, 297-323. Roma: Istituto Nazionale di Statistica.

Instituto Nacional de Estadística (a). Condiciones de Vida. Recuperado el 30 de enero de 2015, de http://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254735976608

Instituto Nacional de Estadística (b). Mercado Laboral. Recuperado el 30 de enero de 2015, de http://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254735976595

ISAD (2008). Statistical Methodology Project on Integration of Survey and Administrative Data. Report of WP1. State of the art on statistical methodologies for integration of surveys and administrative data. Recuperado el 01 de Octubre de 2014, de http://www.cros-portal.eu/content/isad-finished

Juárez Alonso, C. A. (2005). Fusión de Datos: Imputación y Validación. Tesis doctoral, Universitat Politècnica de Catalunya.

Kadane, J. B. (1978). Some Statistical Problems in Merging Data Files. *1978 Compendium of Tax Research*, U.S. Department of the Treasury, 159-171.

Leulescu, A., & Agafitei, M. (2013). Statistical Matching: A model based approach for data integration. *Methodologies and Working papers*. Luxemburgo: Eurostat, Comisión Europea

Leulescu, A., Agafitei, M., & Mercy, J.L. (2011). Statistical matching: a case study on EU-SILC and LFS. ESSnet Data Integration Workshop. Madrid, noviembre de 2011.

Okner, B. (1972a). Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File. *Annals of Economic and Social Measurement, 1*(3), 325-342.

Okner, B. (1972b). Reply and Comments. *Annals of Economic and Social Measurement, 1*(3), 359-362.

Okner, B. (1974). Data Matching and Merging: An Overview. *Annals of Economic and Social Measurement, 3*(2), 347-352.

Paass, G. (1986). Statistical Match: Evaluation of Existing Procedures and Improvements by Using Aditional Information. En *Microanalytic Simulation Models to Support Social and Financial Policy*, 401-422. Amsterdam: Elsevier Science.

Parra Rodríguez, F. (2014). Estimación de la tasa de pobreza en Cantabria en Área pequeña. Recuperado el 12 de Enero de 2015, de http://rpubs.com/PacoParra/52327

Peck, J. K. (1972). Comments. *Annals of Economic and Social Measurement, 1*(3), 347-348.

Rässler, S. (2004). Data fusion: Identification Problems, Validity, and Multiple Imputation. *Austrian Journal of Statistics, 33*(1-2), 153-171.

Rius, R., Aluja, T., & Nonell, R. (1999). File Grafting in Market Research. *Applied Stochastic Models in Business and Industry, 15*(4), 451-460.

Rodgers, W. L. (1984). An Evaluation of Statistical Matching. *Journal of Business & Economic Statistics, 2*(1), 91-102.

Rubin, D. B. (1986). Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputation. *Journal of Business & Economic Statistics, 4*(1), 87-94.

Ruggles, N., & Ruggles, R. A. (1974). A Strategy for Merging and Matching Microdata Sets. *Annals of Economic and Social Measurement, 3*(2), 353-371.

Sims, C. A. (1972). Comments. *Annals of Economic and Social Measurement, 1*(3), 343-345.

Singh, A. C., Mantel, H. J., Kinack, M. D., & Rowe, G. R. (1993). Statistical matching: use of auxiliary information as an alternative to conditional independence assumption. *Survey Methodology, 19*(1), 59-79.

Stiglitz J. E., Sen, A. & Fitoussi J.-P. (2009) Report by the Commission on the Measurement of Economic Performance and Social Progress, Paris. Recuperado el 30 de septiembre de 2014, de http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf

Radner, D. (1980). Report on Exact and Statistical Matching Techniques. Washington D.C.: Federal Committee on Statistical Methodology, U.S. Department of Commerce.

Van der Laan, P. (2000). Integrating administrative registers and household surveys. *Netherlands Official Statistics, 15*(2), 7-15.

ANEXO I

Tabla A1. Distribución de variables comunes en ficheros de individuos.

	EPA	ECV
Sexo		
Hombre	0.4832	0.491
Mujer	0.5168	0.509
Edad		
16-29 = edad entre 16 y 29 años	0.16399	0.16378
30-44 = edad entre 30 y 44 años	0.27475	0.25586
45-64 = edad entre 45 y 64 años	0.31824	0.34013
65+ = edad 65 o más años	0.24303	0.24023
País de nacimiento		
Español = nacido en España	0.87480	0.86854
Europeo-EU27= nacido en país de Europa EU-27	0.02118	0.02157
Resto del mundo = nacido en otro país	0.10402	0.10989
País de nacionalidad		
Español = nacionalidad española	0.87197	0.89490
Europeo-EU27 = nacionalidad de país de Europa EU-27	0.02303	0.01951
Resto del mundo = nacionalidad de otro país	0.10500	0.08559
Estado civil legal		
Soltero	0.28380	0.28757
Casado	0.57545	0.55563
Viudo	0.08140	0.09894
Separado o Divorciado	0.05935	0.05786
Situación profesional		
1 = Empleador	0.02741	0.02645
2 = Empresario sin asalariados o trabajador independiente	0.05065	0.04397
3 = Asalariado con contrato indefinido	0.31687	0.30780
4 = Asalariado con contrato temporal	0.07118	0.06012
5 = Ayuda familiar	0.00202	0.00000
6 = No ocupado	0.53187	0.56166
Ocupación		
0 = Ocupaciones militares	0.00074	0.00234
1 = Directores y gerentes	0.05959	0.06386
2 = Técnicos y profesionales científicos e intelectuales	0.17031	0.17991
3 = Técnicos; profesionales de apoyo	0.09006	0.12461
4 = Empleados contables, administrativos y otros empleados de oficina	0.13078	0.14642
5 = Trabajadores de los servicios de restauración, personales, protección y vendedores	0.22158	0.17601
6 = Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero	0.02066	0.01713
7 = Artesanos y trabajadores cualificados de las industrias manufactureras y la construcción (excepto operadores de instalaciones y maquinaria)	0.11963	0.11604
8 = Operadores de instalaciones y maquinaria, y montadores	0.08842	0.06776
9 = Ocupaciones elementales	0.09823	0.10592
Sector de actividad económica		

1 = Agricultura	0.02541	0.02179
2 = Industria	0.17313	0.19144
3 = Construcción	0.06836	0.06148
4 = Comercio al por mayor y al por menor	0.27894	0.25214
5 = Financiero	0.05157	0.06770
6 = Otras actividades de servicios	0.40259	0.40545
Nivel educativo alcanzado		
1 = Educación primaria	0.31135	0.26370
2 =Educación secundaria de 1ª etapa (incluye formación e inserción laboral equivalente)	0.23231	0.20514
3 =Educación secundaria de 2ª etapa (incluye formación e inserción laboral equivalente)	0.20142	0.21438
4 =Formación e inserción laboral que precisa título de segunda etapa de secundaria	0.00007	0.00445
5 = Educación superior	0.23565	0.24829
8 = No ha recibido nunca educación	0.01920	0.06404
Situación con relación a la actividad		
1 = Ocupado a jornada completa	0.40026	0.38872
2 = Ocupado a jornada parcial	0.06791	0.05573
3 = Parado	0.12231	0.12029
4 = Inactivo	0.40952	0.43527
Está cursando estudios		
1 = Sí	0.14562	0.10476
2 = No	0.85438	0.89524
Busqueda de empleo		
1 = No ocupado menor de 74 años que sí buscó trabajo en las últimas 4 semanas y está disponible para empezar a trabajar en las próximas 2 semanas	0.11640	0.12067
2 = No ocupado menor de 74 años que sí buscó trabajo en las últimas 4 semanas y no está disponible para empezar a trabajar en las próximas 2 semanas	0.00410	0.00306
3 = No ocupado que no buscó trabajo en las últimas 4 semanas	0.28560	0.30252
4 = Ocupado	0.46817	0.44460
5 = No ocupado mayor de 74 años	0.12572	0.12916
Ha trabajado antes		
1 = Sí	0.77786	0.74663
2 = No	0.22214	0.25337

Tabla A2. Descripción de las variables comunes numéricas en el fichero de hogares.[3]

	Definición	EPA		ECV	
		Mín.	Máx.	Mín.	Máx.
Tamaño del hogar	Número de personas que conforman el hogar	1	11	1	10
Unidades de consumo del hogar		1	5.2	1	5.1
Sexo					
n_hombres	Número de hombres en el hogar	0	5	0	8
n_mujeres	Número de mujeres en el hogar	0	5	0	5
Edad					
n_1629	Número de personas en el hogar con edad entre 16 y 29 años	0	4	0	4
n_3044	Número de personas en el hogar con edad entre 30 y 44 años	0	4	0	5
n_4564	Número de personas en el hogar con edad entre 45 y 64 años	0	3	0	3
n_65ymas	Número de personas en el hogar con 65 años o más	0	4	0	3
País de nacimiento					
n_espanoles	Número de personas en el hogar en la categoría de país de nacimiento 1	0	8	0	6
n_europeos27	Número de personas en el hogar en la categoría de país de nacimiento 2	0	5	0	4
n_restomundo	Número de personas en el hogar en la categoría de país de nacimiento 3	0	7	0	8
País de nacionalidad					
n_espanoles1	Número de personas en el hogar en la categoría de país de nacionalidad 1	0	8	0	6
n_europeos271	Número de personas en el hogar en la categoría de país de nacionalidad 2	0	5	0	4
n_restomundo1	Número de personas en el hogar en la categoría de país de nacionalidad 3	0	6	0	8
Estado civil legal					
n_solteros	Número de personas en el hogar solteros	0	5	0	4
n_casados	Número de personas en el hogar casados	0	5	0	6
n_viudos	Número de personas en el hogar viudos	0	2	0	2
n_sepdivor	Número de personas en el hogar separados o divorciados	0	3	0	2
Situación profesional					
n_sprof1	Número de personas en el hogar en la categoría de situación profesional 1	0	4	0	2
n_sprof2	Número de personas en el hogar en la categoría de situación profesional 2	0	4	0	2
n_sprof3	Número de personas en el hogar en la categoría de situación profesional 3	0	4	0	4
n_sprof4	Número de personas en el hogar en la categoría de situación profesional 4	0	3	0	2
n_sprof5	Número de personas en el hogar en la categoría de situación profesional 5	0	2	0	0

[3] Las categorías mencionadas en esta tabla están relacionadas con las variables descritas en la Tabla A1.

n_sprof6	Número de personas en el hogar en la categoría de situación profesional 6	0	7	0	8
Ocupación					
n_ocup0	Número de personas en el hogar en la categoría de ocupación 0	0	1	0	1
n_ocup1	Número de personas en el hogar en la categoría de ocupación 1	0	3	0	2
n_ocup2	Número de personas en el hogar en la categoría de ocupación 2	0	5	0	3
n_ocup3	Número de personas en el hogar en la categoría de ocupación 3	0	2	0	2
n_ocup4	Número de personas en el hogar en la categoría de ocupación 4	0	3	0	2
n_ocup5	Número de personas en el hogar en la categoría de ocupación 5	0	4	0	3
n_ocup6	Número de personas en el hogar en la categoría de ocupación 6	0	3	0	3
n_ocup7	Número de personas en el hogar en la categoría de ocupación 7	0	3	0	2
n_ocup8	Número de personas en el hogar en la categoría de ocupación 8	0	3	0	3
n_ocup9	Número de personas en el hogar en la categoría de ocupación 9	0	3	0	2
Sector de actividad económica					
n_sector1	Número de personas en el hogar en la categoría de sector 1	0	3	0	3
n_sector2	Número de personas en el hogar en la categoría de sector 2	0	3	0	2
n_sector3	Número de personas en el hogar en la categoría de sector 3	0	3	0	2
n_sector4	Número de personas en el hogar en la categoría de sector 4	0	4	0	4
n_sector5	Número de personas en el hogar en la categoría de sector 5	0	2	0	2
n_sector6	Número de personas en el hogar en la categoría de sector 6	0	4	0	3
Nivel educativo alcanzado					
n_edu1	Número de personas en el hogar en la categoría de educación 1	0	6	0	5
n_edu2	Número de personas en el hogar en la categoría de educación 2	0	5	0	5
n_edu3	Número de personas en el hogar en la categoría de educación 3	0	6	0	5
n_edu4	Número de personas en el hogar en la categoría de educación 4	0	1	0	1
n_edu5	Número de personas en el hogar en la categoría de educación 5	0	5	0	5
n_edu8	Número de personas en el hogar en la categoría de educación 8	0	4	0	3
Situación con relación a la actividad					
n_situ1	Número de personas en el hogar ocupadas en jornada completa	0	6	0	4
n_situ2	Número de personas en el hogar ocupadas en jornada parcial	0	4	0	2
n_situ3	Número de personas en el hogar en paro	0	6	0	8
n_situ4	Número de personas en el hogar inactivas	0	5	0	5

Está cursando estudios					
n_cursa1	Número de personas en el hogar que sí están cursando estudios	0	5	0	4
n_cursa2	Número de personas en el hogar que no están cursando estudios	0	7	0	8
Búsqueda de trabajo					
n_busca1	Número de personas en el hogar en la categoría de búsqueda de trabajo 1	0	6	0	8
n_busca2	Número de personas en el hogar en la categoría de búsqueda de trabajo 2	0	2	0	1
n_busca3	Número de personas en el hogar en la categoría de búsqueda de trabajo 3	0	5	0	4
n_busca4	Número de personas en el hogar en la categoría de búsqueda de trabajo 4	0	6	0	4
n_busca5	Número de personas en el hogar en la categoría de búsqueda de trabajo 5	0	3	0	3
Ha trabajado antes					
n_trabantes1	Número de personas no ocupadas en el hogar que sí han trabajado antes	0	4	0	8
n_trabantes2	Número de personas no ocupadas en el hogar que no han trabajado antes	0	5	0	5

Tabla A3. Descripción de la variable común categórica en el fichero de hogares.

	EPA	**ECV**
Composición del hogar		
1 = Una persona: hombre de menos de 30 años	0.00653	0.00791
2 = Una persona: hombre de entre 30 y 64 años	0.04362	0.04892
3 = Una persona: hombre de 65 o más años	0.02745	0.02878
4 = Una persona: mujer de menos de 30 años	0.00341	0.00216
5 = Una persona: mujer de entre 30 y 64 años	0.04451	0.04245
6 = Una persona: mujer de 65 o más años	0.08739	0.09712
7 = 2 adultos sin niños dependientes económicamente, al menos una persona de 65 o más años	0.16766	0.15036
8 = 2 adultos sin niños dependientes económicamente, teniendo ambos menos de 65 años	0.15223	0.15899
9 = Otros hogares sin niños dependientes económicamente	0.16306	0.12374
10 = Un adulto con al menos un niño dependiente	0.01469	0.03453
11 = Dos adultos con un niño dependiente	0.09555	0.10719
12 = Dos adultos con dos niños dependientes	0.09644	0.11079
13 = Dos adultos con tres o más niños dependientes	0.02240	0.03022
14 = Otros hogares con niños dependientes	0.07507	0.05683