

**ENCUESTA FINANCIERA
DE LAS FAMILIAS ESPAÑOLAS (EFF):
DESCRIPCIÓN Y MÉTODOS
DE LA ENCUESTA DE 2002**

2004

Olympia Bover

**Documentos Ocasionales
N.º 0409**

BANCODE ESPAÑA



**ENCUESTA FINANCIERA DE LAS FAMILIAS ESPAÑOLAS (EFF): DESCRIPCIÓN
Y MÉTODOS DE LA ENCUESTA DE 2002**

ENCUESTA FINANCIERA DE LAS FAMILIAS ESPAÑOLAS (EFF): DESCRIPCIÓN Y MÉTODOS DE LA ENCUESTA DE 2002*

Olympia Bover

BANCO DE ESPAÑA

(*) La EFF no hubiera sido posible sin la participación en el proyecto del Instituto Nacional de Estadística y de la Agencia Tributaria. En los aspectos técnicos hay que mencionar el trabajo de algunas personas en concreto, Carmen Anido, Carlos de la Fuente, Juan Luis Pérez, Juana Porras, Jorge Saralegui y Carmen Ureña. El ejemplo del SCF de Estados Unidos y la constante voluntad de ayuda de Arthur Kennickell han sido también fundamentales. De gran utilidad ha sido, asimismo, la generosa colaboración de Giovanni d'Alessio y su relato de la experiencia de la Banca d'Italia. De forma destacada, han dedicado mucho tiempo y esfuerzo al proyecto Cristina Barceló y Pilar Velilla, como parte del equipo de la EFF. Doy también las gracias a Manuel Arellano por su apoyo y consejo constantes a lo largo de todo el proyecto. La empresa encargada del trabajo de campo pudo hacer frente al desafío que la EFF suponía gracias a la determinación de Chelo Perera, Amelia Rochela, Lourdes Vázquez, José Ignacio Wert y de todos los entrevistadores. Por último pero no por ello menos importante, se agradece especialmente la colaboración brindada por los hogares que se mostraron dispuestos a responder a la encuesta.

La serie de Documentos Ocasionales tiene como objetivo la difusión de trabajos realizados en el Banco de España, en el ámbito de sus competencias, que se consideran de interés general.

Las opiniones y análisis que aparecen en la serie de Documentos Ocasionales son responsabilidad de los autores y, por tanto, no necesariamente coinciden con las del Banco de España o las del Eurosistema.

El Banco de España difunde sus informes más importantes y la mayoría de sus publicaciones a través de la red Internet en la dirección <http://www.bde.es>.

Se permite la reproducción para fines docentes o sin ánimo de lucro, siempre que se cite la fuente.

© BANCO DE ESPAÑA, Madrid, 2004

ISSN: 1696-2222 (edición impresa)
ISSN: 1696-2230 (edición electrónica)
Depósito legal: M. 51919-2004
Imprenta del Banco de España

Resumen

Este documento describe la elaboración de la primera ola de la Encuesta Financiera de las Familias (EFF). Esta encuesta es la única fuente estadística en España que permite relacionar las rentas, los activos, las deudas y el gasto de cada unidad familiar. En primer lugar se presenta el diseño muestral, explicando cómo se llevó a cabo el sobremuestreo de los hogares más ricos. Los estratos se construyeron a partir del impuesto del patrimonio mediante un sistema ciego de colaboración entre el Instituto Nacional de Estadística y la Agencia Tributaria que respeta los estrictos requisitos de confidencialidad de la AT a la vez que mantiene un marco único de población para el muestreo. Seguidamente se describen las tasas de no respuesta a la encuesta y los ajustes a los factores de elevación. Finalmente, se explica el método de imputación múltiple utilizado para tratar la no respuesta a determinadas preguntas.

Palabras clave: encuesta de riqueza, sobremuestreo de hogares más ricos, métodos de imputación

Clasificación JEL: C81, D31

1 Introducción

Para el análisis económico y el diseño de políticas económicas y regulatorias es cada vez más relevante conocer la situación y las decisiones financieras de las familias en vista de la creciente complejidad y variedad de estas decisiones. En España se dispone de información sobre la cuantía total de la riqueza financiera de las familias y sus deudas (aunque no así de sus activos reales) a través de las Cuentas Financieras. Sin embargo esta información es demasiado agregada como para permitir el estudio de aspectos relevantes de la situación financiera de las familias.

Efectivamente, los niveles agregados no bastan para evaluar la situación de los distintos tipos de familias, que requiere conocer la distribución de los activos reales y financieros de los hogares, su endeudamiento y su relación con otras variables. Por ejemplo, es fundamental saber cómo el reciente aumento de la deuda de los hogares se distribuye entre los distintos tipos de familias. ¿Oculta la situación agregada, aparentemente cómoda, una fragilidad significativa de algunos hogares o pueden la inversión y el endeudamiento de las familias absorber fluctuaciones adversas de los mercados? ¿Son los hogares que han contraído deudas en los últimos años los mismos que han adquirido activos? Por poner otros ejemplos, hasta ahora no era posible saber qué familias tienen planes de pensiones, o cómo financian las familias jóvenes la compra de su vivienda. Hay también otras muchas cuestiones de importancia para las que se precisa disponer de información a nivel de hogares que permita relacionar las rentas, los activos, las deudas y las características familiares.

Por ello, el Banco de España decidió en 2001 realizar una encuesta sobre la situación financiera de las familias [“Encuesta Financiera de las Familias” (EFF)], similar a las que se llevan a cabo en otros países. La encuesta de la Banca d’Italia [*Survey on Household Income and Wealth (SHIW)*] y el *Survey of Consumer Finances (SCF)* de la Reserva Federal de Estados Unidos fueron los modelos en los que se inspiró esta encuesta.

Una característica distintiva de la EFF, siguiendo el ejemplo del SCF, es que se lleva a cabo un sobremuestreo de los hogares con un mayor nivel de riqueza. La distribución de la riqueza es muy asimétrica y, además, algunas clases de activos solo están en poder de una pequeña fracción de la población. Por lo tanto, se consideró muy importante contar con una muestra que no solo fuera representativa del conjunto de la población, sino también de la riqueza agregada de la economía, y que facilitara el estudio del comportamiento financiero en el tramo superior de la distribución de la riqueza. Esta muestra se logró gracias a la colaboración de la Agencia Tributaria y del Instituto Nacional de Estadística (INE).

La EFF es la única fuente estadística existente en España que permite relacionar las rentas, los activos, las deudas y el gasto de cada unidad familiar. Es habitual comparar cocientes entre los pagos por deuda (o la deuda pendiente) agregados y la renta disponible (o los activos) agregados. Gracias a la EFF será posible calcular medidas genuinas de ratios de endeudamiento agregados como promedio o mediana de los ratios de los hogares. En general un ratio de agregados es distinto de un agregado de ratios, pero esta diferencia puede ser especialmente relevante en el caso de magnitudes con una distribución muy dispersa y asimétrica, como es el caso de las magnitudes que nos ocupan.

El propósito de este documento es describir las principales características de la elaboración de la EFF. En la sección 2 se describe el cuestionario. En la sección 3 se expone el diseño muestral. En la sección 4 se presentan las distintas etapas del trabajo de campo y se proporciona un análisis de los hogares que no responden. En la sección 5 se describe cómo se ajustan los factores de elevación para tener en cuenta dichos hogares y se comenta brevemente otras estrategias de corrección. Finalmente, en la sección 6 se aborda el problema suscitado por

los hogares que responden a la encuesta pero dejan sin contestar algunas preguntas, y se presentan los métodos de imputación utilizados.

2 Cuestionario

Contenido

El cuestionario se divide en nueve secciones principales, que reflejan los objetivos de la EFF, y que son las siguientes:

1. Características demográficas¹
2. Activos reales y deudas asociadas
3. Otras deudas
4. Activos financieros
5. Seguros y pensiones
6. Situación laboral e ingresos relacionados (de todos los miembros del hogar)
7. Rentas no relacionadas con la actividad laboral percibidas en el año natural anterior (2001)
8. Uso de instrumentos de pago
9. Consumo y ahorro.

El cuestionario de la EFF contiene un elevado número de preguntas similares a las contenidas en las encuestas de riqueza de otros países (en especial, Italia, Países Bajos y Estados Unidos), aunque existen algunas diferencias. Por ejemplo, dado que la obtención de información sobre el endeudamiento es una de las motivaciones principales de la EFF, se prestó una atención a los pasivos de los hogares superior a la SHIW italiana. Por otra parte, en línea con esta última encuesta, la categorización de los activos es menos detallada que en el SCF. Además, habida cuenta de que en España (al igual que en Italia, y a diferencia de en Estados Unidos) es relevante considerar las familias extensas que viven en el mismo hogar, fue necesario aumentar el número de preguntas sobre la situación laboral y los ingresos.

Por último, a diferencia del SCF, el cuestionario contiene algunas preguntas sobre el gasto en bienes no duraderos y en alimentación, dado el interés de la relación entre consumo, renta y los distintos tipos de riqueza [véase Browning et al. (2002), sobre las formulaciones recomendadas para estas preguntas y una evaluación de sus resultados].

En todo caso, una consideración de suma importancia al elaborar el cuestionario fue tratar de que la entrevista no se prolongara más allá de una hora, en promedio, tiempo sensiblemente inferior al necesario para contestar el SCF, pero que, en la práctica europea, se percibe como la duración máxima factible de una entrevista. De hecho, después de que se llevara a cabo la encuesta piloto, donde la duración media fue de 1 hora y 25 minutos por entrevista, se decidió eliminar algunas preguntas. En el cuadro 1 se recogen datos relativos al número de preguntas en la EFF, y se compara con el SCF. La mediana del número de preguntas planteadas en el SCF es más del doble que la de la EFF, mientras que si se considera el número de preguntas en las que se solicitan importes monetarios en euros, la mediana del SCF es 1,6 veces la mediana de la EFF.

Otros aspectos

Se decidió realizar “entrevistas personales asistidas por ordenador” [*Computer Assisted Personal Interview (CAPI)*]. Además de evitar la fase de introducción de datos, el uso del cuestionario CAPI

1. Las preguntas de demografía se redactaron de modo que pudieran compararse con preguntas similares de otras encuestas a hogares realizadas por el Instituto Nacional de Estadística.

ayudó a los entrevistadores a seguir las rutas correctas en este complejo cuestionario y permitió la programación de algunas verificaciones básicas para detectar, en presencia de los entrevistados, posibles incoherencias o errores en sus respuestas. También permitió incorporar fácilmente la conversión de pesetas a euros (y viceversa), muy necesaria para facilitar las respuestas, ya que el trabajo de campo comenzó cuando solo habían transcurrido unos meses desde la introducción del euro.

El trabajo de campo duró varios meses, desde octubre de 2002 a mayo de 2003. Se añadieron algunas preguntas en las entrevistas realizadas a partir de enero de 2003 para controlar los posibles incrementos de renta vinculados simplemente a la inflación.

3 Diseño de la muestra

Siguiendo el ejemplo del SCF, se consideró importante que hubiera un sobremuestreo de los hogares con un mayor nivel de riqueza. Efectivamente, hay muchos activos en poder de una pequeña fracción de la población y, por lo tanto, una muestra aleatoria estándar contendría insuficientes observaciones para la realización de análisis relevantes sobre las decisiones de inversión de las familias. Gracias a la colaboración del INE y la Agencia Tributaria fue posible la construcción de una muestra que incorporase una sensible sobrerrepresentación de los hogares con mayor riqueza.

Base para el sobremuestreo

La existencia en España de un impuesto sobre el patrimonio, ha permitido que el sobremuestreo en la EFF de los hogares con un mayor nivel de riqueza se basara en declaraciones individuales de riqueza. Esta es una diferencia importante con el SCF, en el que se construye un índice de riqueza integrando la información sobre las rentas del capital incluidas en las declaraciones individuales del impuesto sobre la renta, ya que, en Estados Unidos, no existe un impuesto sobre el patrimonio. Las personas que estaban obligadas a presentar declaración de patrimonio en España en 1999, que es el ejercicio fiscal utilizado para seleccionar la muestra, fueron aquellas con un patrimonio superior a 104.000 euros. En 1999, en torno a 980.000 individuos (correspondientes a, aproximadamente, 700.000 hogares) presentaron declaración de patrimonio.

La definición de los estratos de riqueza se basó en los intervalos del SCF y en la distribución de los hogares por percentiles del impuesto sobre el patrimonio en España. Se definieron ocho estratos con un sobremuestreo creciente con el nivel de riqueza de los estratos. Los intervalos se muestran en el cuadro 2. Los estratos 2 y 3 incluyen algo menos de la mitad de la distribución de la riqueza declarada. Los estratos 4, 5 y 6 recogen el tercer cuartil, salvo por el último percentil y medio, aproximadamente, que está representado por los dos últimos estratos.

Por último, no hubo sobremuestreo de los hogares con un mayor nivel de riqueza en Navarra y el País Vasco, porque la Agencia Tributaria no dispone de datos fiscales de carácter personal para esas dos Comunidades Autónomas.

Garantías de confidencialidad

La Agencia Tributaria está sometida a requisitos de confidencialidad muy estrictos y no puede dar a conocer, ni siquiera al INE, ninguna información fiscal de carácter personal (ni siquiera en forma de intervalos). Para solventar este problema y permitir el sobremuestreo de los hogares con un patrimonio más elevado, preservando al mismo tiempo la confidencialidad, la Agencia Tributaria se ofreció para efectuar ella misma la selección aleatoria de la muestra, siguiendo los requisitos de diseño muestral establecidos por el Banco de España y el INE.

Gracias a la colaboración del INE y de la Agencia Tributaria existe un único universo de población para la muestra². El marco muestral fue el Padrón Continuo de mediados de 2001, en el que las unidades son hogares definidos por su domicilio. Con esta información, enviada por el INE a la AT, esta construyó, para cada uno de los domicilios, tres variables basadas en la información procedente del impuesto sobre la renta y del impuesto sobre el patrimonio, constituyendo estos datos el punto de partida del muestreo.

La primera variable, el indicador del estrato de riqueza, se basa en el patrimonio total declarado por el hogar, que se obtuvo sumando las declaraciones de todos sus miembros, en los casos pertinentes. La segunda variable, para aquellos que presentan declaración del impuesto sobre la renta, pero no sobre el patrimonio, indica a qué cuartil pertenece el hogar en la

2. A diferencia del SCF, para el que se utilizaron dos muestras (una aleatoria y otra con sobremuestreo).

distribución nacional de la renta imponible. Por último, la tercera variable contiene la renta per cápita declarada del hogar. Las variables de renta fueron de gran utilidad para elegir las reservas de la muestra (como se verá más adelante) y garantizar, en el caso de los grandes municipios, que se seleccionaran hogares pertenecientes a todos los niveles de renta. Esto último se obtuvo mediante la aplicación, dentro de cada estrato de riqueza, de un muestreo sistemático con arranque aleatorio a las viviendas ordenadas según su renta per cápita. Además, se utilizó el indicador de los cuartiles de renta para corregir la no respuesta en las grandes ciudades. La información fiscal disponible en el momento era de 1999, lo que supuso la existencia de algunos desajustes de carácter limitado entre las dos fuentes.

Muestreo

Se llevó a cabo un diseño muestral diferente en cada uno de los tres casos siguientes:

- (i) municipios con más de 100.000 habitantes. Para ellos, el muestreo fue aleatorio dentro de los ocho estratos de riqueza.
- (ii) municipios con 100.000 habitantes o menos. En este caso, se realizó un muestreo bietápico, para el que se seleccionaron en primer lugar las secciones censales, siendo la probabilidad proporcional a su población. Luego, dentro de cada sección censal, la selección de hogares fue diferente, dependiendo del número de declarantes de patrimonio.
- (iii) Navarra y el País Vasco. Dado que, en estas comunidades, no fue posible efectuar un sobremuestreo de los hogares con mayor nivel de riqueza, la muestra se seleccionó según un muestreo bietápico estratificado, con seis estratos definidos de acuerdo con el tamaño del municipio.

Reservas

Otro aspecto relevante del diseño muestral de la EFF fue el sistema de hogares de reserva elegido. Para tratar de preservar el sobremuestreo en la medida de lo posible, se diseñó un procedimiento muy cuidadoso de reposición. Este sistema consiguió mitigar las consecuencias de no disponer de información sobre el estrato de riqueza al que pertenece cada hogar de la muestra, lo que impedía supervisar durante el trabajo de campo las tasas de respuesta por estratos y expandir la muestra según conviniera.

De este modo, se proporcionaron inicialmente hasta cuatro hogares de reserva por cada hogar titular. Esas reservas eran los dos hogares inmediatamente anteriores y los dos inmediatamente posteriores de un fichero ordenado por cuartiles de renta (para los que no presentan declaración de patrimonio), estrato de riqueza y renta per cápita del hogar. Las reposiciones debían pertenecer al mismo cuartil de renta (para los que no presentan declaración de patrimonio) o al mismo estrato de riqueza que el hogar de la muestra. Esto se hizo así dentro de los municipios, en el caso de las grandes ciudades, y dentro de secciones censales, en el caso de las ciudades pequeñas, para que el domicilio de las reservas no estuvieran demasiado distantes del hogar titular. Ello supuso que, en ocasiones, se dispusiera de menos de cuatro sustituciones (y en algunas circunstancias, de ninguna). En el caso de Navarra y del País Vasco se adoptó un sistema más convencional de un grupo de ocho hogares que podían ser posibles reservas para ocho hogares titulares de la muestra (dentro de la misma sección censal).

4 El trabajo de campo

El trabajo de campo fue realizado por la empresa Demoscopia.

Una de las características de las encuestas sobre la riqueza y la renta es el elevado porcentaje de hogares que no responden a la encuesta, debido a la naturaleza o a la dificultad de las preguntas. En esta sección se describe la situación de la EFF, y en la siguiente se abordan las correcciones adoptadas para minimizar los problemas de autoselección muestral.

Tras realizar el trabajo de campo, el número total de entrevistas válidas conseguidas fue de 5.143. El número de visitas y de intentos de conseguir una entrevista figura en el cuadro 3.

Esfuerzos para reducir la no respuesta

Para esta primera ola de la encuesta, los hogares recibieron una carta solicitando la participación del Gobernador del Banco de España, otra de la empresa encargada de realizar el trabajo de campo y un folleto. La carta del Gobernador se adjuntó en sobre aparte, pero dentro del enviado por la empresa, para poner de manifiesto que el Banco desconocía la identidad y el domicilio de los hogares de la muestra. El objetivo era que fuera evidente para los entrevistados que sus respuestas a la encuesta no se utilizarían con fines fiscales. No obstante, con demasiada frecuencia las cartas se perdieron o los propios entrevistados las tiraron porque no parecían importantes. También se les proporcionó una dirección de Internet y números de teléfono para garantizar la existencia y fines de la encuesta y responder a las preguntas que se les pudieran plantear. Se notificó a las sucursales del Banco de España de la realización de la encuesta, lo que resultó ser importante pues muchas personas llamaron a éstas para pedir información. Por último, se ofreció un regalo simbólico a las familias participantes.

Se decidió comenzar el trabajo de campo a primeros de octubre, tras las vacaciones de verano, porque la probabilidad de no encontrar a las familias en su vivienda principal es muy elevada desde mediados de junio a mediados de septiembre.

Formación de los entrevistadores

Se organizaron tres jornadas de formación de un día de duración en lugares diferentes (Madrid, Barcelona y Bilbao), a las que asistieron los coordinadores de los entrevistadores y algunos encuestadores. Con anterioridad, para llevar a cabo la encuesta piloto, se había celebrado una jornada de formación en Madrid. En estas jornadas, se analizó el cuestionario, examinando casos hipotéticos y familiarizándose con la aplicación CAPI. Un representante del Banco de España participó en estas sesiones para explicar la importancia y la dificultad del proyecto y para aclarar todas las dudas que pudieran surgir durante la explicación del cuestionario. Más adelante, se organizaron por todo el país jornadas de formación de entrevistadores a cargo de los coordinadores. Se entregó a cada encuestador un manual en cuya preparación participó el equipo de la EFF del Banco de España con explicaciones detalladas sobre el cuestionario, definiciones de los conceptos utilizados y ejemplos.

Imposibilidad de establecer contacto (nunca en casa)

El número de hogares en los que el entrevistador no pudo contactar con nadie (tras confirmar in situ que la dirección era la correcta) es muy elevado, a pesar de haberlo intentado, por lo menos, en cinco ocasiones. El número de estos intentos en proporción al número total de visitas por estratos de riqueza tiene algún componente no aleatorio, como puede observarse en el cuadro 4³. Se consideró que el posible motivo por el que no se logró entrar en contacto con personas con un elevado nivel de riqueza durante el trabajo de campo fue la existencia de múltiples residencias. En general, la dificultad de que se cumplimentara el cuestionario de la encuesta (tanto por no poder entrar en contacto con el hogar como por su negativa a participar) convenció a la empresa encargada de realizar el trabajo de campo de la necesidad de duplicar el precio por encuesta cumplimentada y de cambiar el sistema de incentivos a los entrevistadores, incluyendo una dieta diaria fija en las áreas más difíciles. Lamentablemente, esta decisión se tomó después de que algunos de los buenos entrevistadores abandonaran el proyecto.

Negativa

Como se puede apreciar en el cuadro 4, existe un claro componente no aleatorio en las tasas de cooperación [definidas como cumplimentadas/(cumplimentadas+negativas a participar en la encuesta)]. Estas disminuyen a medida que se asciende en los estratos de riqueza, fluctuando entre el 53,6% y el 29,4%, un resultado en línea con los obtenidos en muestras similares en otros países. Resulta evidente que, en general, las tasas de cooperación o de respuesta no son muy informativas en el caso de sobremuestreo, ya que dependen del grado en que éste se practique. Para poder realizar comparaciones, se construyen tasas de cooperación por estratos. En el SCF de 1992 las tasas de cooperación oscilaron entre el 52,6% para el estrato 1 y el 20,1% para el estrato 7⁴.

Mecanismos de supervisión y respuestas descartadas

Los supervisores de la empresa encargada del trabajo de campo revisaron en primer lugar todas las encuestas que habían sido cumplimentadas. Se volvió a contactar con alrededor del 67% de los entrevistados (principalmente por teléfono, y en algunos casos personalmente). Los motivos principales fueron los siguientes: (i) comprobar posibles incoherencias, (ii) verificar los valores extremos, y (iii) reducir la no respuesta a ciertas preguntas.

Se desarrolló un programa para detectar incoherencias lógicas entre las preguntas. Los hogares proporcionaron a veces explicaciones plausibles para estas incoherencias. Por ejemplo, en algunos casos, el motivo de que la persona de referencia del hogar hubiera nacido después de que se adquiriera la vivienda principal se debía a que la había heredado. No obstante, en numerosas ocasiones, el programa fue de gran utilidad para detectar errores. Este programa no se incluyó en el cuestionario CAPI porque hubiera significado ir hacia atrás y hacia delante en el programa del cuestionario en medio de la entrevista. En cambio, sí se incluyó en el programa CAPI un código para la detección de incoherencias o errores básicos relativos a preguntas individuales.

Además de los motivos anteriores, para controlar el trabajo de los entrevistadores se volvió a contactar con un gran número de entrevistados de forma aleatoria.

El equipo de la EFF del Banco de España también examinó las encuestas cumplimentadas para comprobar, en general, la coherencia individual. Como resultado de este proceso, se decidió descartar las encuestas en las que se hubiera contestado menos del 30% de las preguntas relativas a importes monetarios en euros, así como aquellas en las que no se

3. La Agencia Tributaria proporcionó las cifras que figuran en el cuadro 4.

4. Para el año 1992 se dispone de información lo suficientemente desagregada sobre el tipo de respuesta por intento de lograr una entrevista [en Kennickell y McManus (1993)]. Consideramos que, dado el diferente diseño muestral (en particular, la ausencia de reservas en el SCF) y la información disponible, las tasas de cooperación tienen más sentido y son más comparables que las tasas de respuesta general disponibles.

facilitaba información sobre renta (ni rentas del trabajo ni del capital ni de prestaciones de cualquier tipo)⁵. El número final de entrevistas descartadas se muestra en el cuadro 3.

Análisis descriptivo de los hogares que no responden

No se dispone de demasiada información sobre los hogares que no responden. Sin embargo se conoce la provincia y el tamaño del municipio. Además, se pidió a los entrevistadores que contestaran cuatro preguntas generales sobre cada uno de los hogares, tanto para los que respondieron como para los que se negaron⁶. Las preguntas eran las siguientes: (i) tipo de vivienda, (ii) tipo de barrio, (iii) tipo de edificio, y (iv) clase social.

En el cuadro 5 se recoge un análisis de descomposición de la varianza sobre la base de un modelo de probabilidad lineal que incluye como factores geográficos variables ficticias para las 48 provincias y el tamaño de los municipios y como factores socioeconómicos el tipo de vivienda, el tipo de barrio y la clase social⁷. También se facilita una medida de la bondad del ajuste para la estimación logit. Como puede apreciarse, las variables socioeconómicas de las que informan los entrevistadores no añaden mucho a los factores geográficos. Los entrevistadores están acostumbrados a cumplimentar estas variables para otras encuestas, pero, sin embargo, éstas carecen de poder explicativo en este caso. En general, los factores geográficos y la información procedente de los entrevistadores no parecen explicar demasiado las decisiones de cooperar. De hecho, el pseudo- R^2 es del 11%, y más del 30% de los pares posibles en nuestra muestra tienen probabilidades predichas que no están en consonancia con las respuestas observadas. Se esperaba que fueran incluso menos relevantes en un análisis condicional a las variables del diseño muestral. Como se verá más adelante, las correcciones por la no respuesta obtenida se basan únicamente en los diferenciales de no respuesta por cuartiles de renta, estratos de riqueza y factores geográficos llevados a cabo por la Agencia Tributaria al elaborar los factores de elevación. Se planteó la posibilidad de realizar nuevos ajustes por regiones, pero no se efectuaron por ser las celdas demasiado pequeñas.

En el cuadro 6, se presentan estimaciones de los parámetros de un modelo logit de la decisión de cumplimentar la encuesta frente a la de negarse. En aras de la simplicidad, se incluyeron variables ficticias para las 17 comunidades, en lugar de para las 48 provincias. Los hogares de Asturias, Madrid y Andalucía son los que tienen una probabilidad más elevada de cooperar (la probabilidad predicha es de 59% para los hogares con las características de la moda en Asturias; 55% en Madrid)⁸. Por el contrario, los hogares de Cantabria y Galicia son los que tienen menos probabilidad de cooperar, con probabilidades predichas del 23% y 24%, respectivamente. El efecto del tamaño del municipio es el esperado: es más probable que los hogares se nieguen a cooperar conforme aumenta el tamaño del municipio. Por ejemplo, en la Comunidad Autónoma de Madrid, la probabilidad de cooperar llega al 67% en los municipios de diez a cincuenta mil habitantes.

5. 30% parecía el porcentaje más adecuado tras haber revisado el contenido informativo de las encuestas cumplimentadas.

6. Obsérvese que esta información no se registró para los 247 hogares que no respondieron y que se incluyeron inicialmente en "otros motivos de no respuesta", pero que finalmente se codificaron como negativas a responder tras recodificar la categoría de "otros motivos". Por razones obvias, tampoco se dispone de esta información, cuando no se logró entrar en contacto con el hogar.

7. También se trató de añadir la variable del tipo de inmueble, pero no afecta a los resultados y es una variable difícil de interpretar.

8. Las características de la moda son: municipio de entre 100.000 y 500.000 habitantes, viviendas de alta calidad, barrio de tipo medio y clase social media-media.

El grado de sobremuestreo de la muestra final

Por último, se ofrecen algunos datos sobre el grado de sobremuestreo de la muestra final, proporcionados por la Agencia Tributaria debido a las restricciones de confidencialidad. En general, algo más del 40% de los hogares que realizaron la entrevista corresponde a personas que presentan declaración de patrimonio, lo que está en consonancia con el SCF, donde entre el 35% y el 37% de las encuestas cumplimentadas proceden de la submuestra estratificada por riqueza. Además, la información agregada de las declaraciones tributarias indica que un 4 por mil de los hogares posee el 40% de la riqueza imponible total. Así pues, en una muestra aleatoria de 5.000, debería esperarse contar como máximo con 20 hogares representando a éstos, un límite superior ya que supone una tasa de respuesta uniforme. En cambio la muestra de la EFF incluye más de 500, lo que ilustra el grado de sobrerrepresentación de los hogares más ricos logrado.

Factores de elevación

Los factores de elevación del diseño muestral para cada hogar se obtienen de forma natural de nuestro marco único de población como la inversa de la probabilidad de ser incluido en la muestra. El Padrón Continuo de referencia está fechado en el último trimestre del 2002. En un primer paso, estos factores de elevación iniciales se han ajustado por la no respuesta dentro de las celdas definidas por las distintas variables del diseño muestral (que difieren según el tamaño del municipio). En particular, éstas incluyen: tamaño del municipio, estrato de riqueza, cuartil de renta de los no declarantes del impuesto de patrimonio en las grandes ciudades, proporción de declarantes del impuesto sobre el patrimonio en cada sección censal, tamaño de la sección censal. No fue posible realizar un ajuste adicional por regiones dentro de esas celdas, por el tamaño insuficiente de la muestra. De acuerdo con las restricciones de confidencialidad mencionadas anteriormente, la Agencia Tributaria elaboró estos primeros factores de elevación surgidos del diseño muestral y de la no respuesta siguiendo instrucciones detalladas del INE.

Basado en los factores de elevación anteriores, el INE realizó un análisis de los estimadores muestrales de varias características de la población: edad, educación, situación laboral, sexo y tamaño del hogar. Las estimaciones obtenidas por educación y situación laboral son satisfactorias, pero se observó que la muestra estaba sesgada hacia hogares de menor tamaño e individuos de mayor edad. Por lo tanto, los primeros factores de elevación se ajustaron (por una función de distancia lineal utilizando el procedimiento de Calmar) para adaptarse a las estructuras de edad, tamaño del hogar y sexo del Padrón⁹.

A fin de mejorar los pesos, se consideraron para este ajuste las distintas estructuras de edad que proporciona el Censo por tamaño de municipio, distinguiendo entre municipios grandes (más de 100.000 habitantes) y el resto. Es de esperar que las ratios de tasas de respuesta entre grupos de edad varíen por tamaño del municipio. Especialmente teniendo en cuenta que el sobremuestreo por nivel de riqueza se realizó principalmente en los municipios grandes y que las ratios de tasas de respuesta entre grupos de edad son probablemente distintas según el estrato de riqueza.

Posibles correcciones adicionales por no respuesta

Como se deduce de lo anterior, un supuesto de trabajo esencial es que la probabilidad por estratos de que un hogar no responda a la encuesta no depende de las variables de interés [véase Rubin (1976) o Little y Rubin (1987)]. En general, el supuesto de ausencia aleatoria de respuesta puede escribirse como $f(Y|X) = f(Y|X, D=1)$, donde Y es una variable de interés, X es el vector de las variables del diseño muestral y $D=1$ indica la participación en la encuesta.

Supongamos que este supuesto no se cumple pero se dispone de algunas variables adicionales Z , para las que se verifica $f(Y|X, Z) = f(Y|X, Z, D=1)$. En ese caso los factores de elevación (evaluados sobre la base de X) tendrían que multiplicarse por $Pr ob(D=1|X, Z) / Pr ob(D=1|X)$, por ejemplo por $\Phi(\alpha + \beta'X + \gamma'Z) / \frac{1}{N} \sum_{i=1}^N \Phi(\alpha + \beta'X + \gamma'Z_i)$, si $Pr ob(D=1|X, Z)$ se modeliza utilizando una especificación probit¹⁰.

Estas correcciones por no respuesta se basan en supuestos similares a los que se llevan a cabo en los métodos de emparejamiento, donde, condicional a X y Z , D se asigna aleatoriamente. Estas correcciones difieren de las correcciones λ de Heckman realizadas en el

9. En Sautory (1993) se puede encontrar información sobre el procedimiento de Calmar, desarrollado por el INSEE francés. Una característica de utilidad de este procedimiento es que permite realizar simultáneamente ajustes a niveles distintos, en particular de hogares y de individuos.

10. Sería necesario asegurarse de que los ajustes adicionales de los factores de elevación no entrañaran aumentos de la volatilidad.

modelo de selección clásico. En ese contexto, D depende de variables no observables que no son independientes de X ¹¹. Esta solución clásica es más ambiciosa, ya que no requiere exogeneidad condicional de D . Sin embargo, es más restrictiva porque depende de la imposición de un modelo estructural específico.

En el caso de la EFF no se dispone de más información de interés con el suficiente apoyo en la muestra como para explicar la no participación. En el SCF se dispone de una estimación de la probabilidad de no responder a partir de una rica información sobre participantes y no participantes procedente de la submuestra estratificada por riqueza, aunque se sabe que aumenta la volatilidad de los factores de elevación sin añadir demasiada información a los ajustes por estratos de riqueza, ubicación y medida de la renta financiera [Kennickell (2000)].

11. Se tiene $E(Y | X, Z) = X\beta$ donde $E(Y | X, Z) \neq E(Y | X, Z, D = 1)$ y $E(Y | X, Z, D = 1) = X\beta + \sigma\lambda(X\delta + Z\mu)$ donde $Y = X\beta + u, D = 1$ si $X\delta + Z\mu + v$.

6.1 Hogares que responden a la encuesta, pero dejan sin contestar algunas preguntas

Este tipo de no respuesta se produce cuando un hogar acepta responder a la encuesta, pero deja sin contestar una o varias preguntas. Junto con un nivel elevado de negativas a participar en ella, la no respuesta a determinadas preguntas es una característica inherente de las encuestas de riqueza. Además, estos dos tipos de no respuesta están estrechamente relacionados. De hecho, la no respuesta a algunas preguntas dependerá, en parte, de lo estricto de las condiciones que se hayan definido (en términos de cantidad de preguntas relevantes que se han de contestar) para dar por válida una entrevista, y esto a su vez afectará la tasa de respuesta general a la encuesta. Debe tenerse en cuenta que criterios de aceptabilidad estrictos ofrecen los incentivos adecuados a los entrevistadores, pero pueden dar lugar a una autoselección adicional a la provocada por las negativas totales a responder. Además, la imposición de un mínimo de respuestas demasiado exigente puede reducir la fiabilidad del trabajo de los entrevistadores.

En general, se suele responder fácilmente a las preguntas relativas a si el hogar posee un activo concreto. Sin embargo, a los hogares puede resultarles más difícil proporcionar información sobre el valor de los activos mantenidos o sobre el importe de las fuentes de ingresos. Al diseñar el cuestionario se consideró la posibilidad de recoger sistemáticamente datos por rangos mediante preguntas sucesivas con intervalos abiertos en los casos de respuestas “no sabe/no contesta” (NS/NC) a las preguntas sobre importes monetarios en euros, como ocurre en el SCF o en la Encuesta de Salud y Envejecimiento de Estados Unidos [*Health and Retirement Study (HRS)*]. Sin embargo, se estimó que plantear a los entrevistados una batería de preguntas con rangos abiertos podría inducirles a no cooperar en la encuesta, ya que, a priori, se sospechó que la mayor parte de las respuestas “no sabe/no contesta”, estarían reflejando en realidad una negativa a facilitar información. En ese sentido, la evidencia del SCF sobre las tasas de respuesta a algunas preguntas por tipo de información suministrada (es decir, valor preciso, tarjeta o rango) indica que la información obtenida utilizando preguntas con rangos abiertos es mínima [véanse los cuadros en Kennickell (2000)]. Por lo tanto, a los entrevistadores solo se les entregó tarjetas de rangos cerrados para ayudar a los entrevistados a proporcionar información. Sin embargo, mientras se realizaba el trabajo de campo, quedó patente que los entrevistadores apenas utilizaban estas tarjetas, alegando que no parecían ayudar a los entrevistados.

En el cuadro 1 se presentan algunos estadísticos sobre el número de preguntas planteadas y contestadas por hogar. Además, en el cuadro 7, figuran las tasas de no respuesta a determinadas preguntas clave. Para comparar, se examinan cuadros similares correspondientes a los SCF de 1995 y de 1998 incluidos en Kennickell (2000). Las tasas de respuestas de la EFF relativas a los valores de distintos activos son muy cercanas a las obtenidas en el SCF (sumando en este último caso las tasas de valores precisos y las de tarjetas de rangos). Por otra parte, las tasas de respuesta para los distintos tipos de rentas relacionadas con la actividad laboral son sensiblemente más elevadas en la EFF. Las tasas de respuesta para las rentas procedentes de activos financieros son sin embargo reducidas en la EFF, pero no tenemos información comparable del SCF.

6.2 Métodos de imputación y su motivación¹²

¿Por qué se realiza la imputación?

Dadas las tasas de no respuesta a algunas preguntas aisladas, no sería sensato trabajar solo con los casos en que se dispone de información e ignorar esta no respuesta. En primer lugar, esto supondría que los casos completos son un submuestra aleatoria de la muestra original, lo que probablemente no es válido (como puede observarse en el ejemplo del caso de la negativa a participar en la encuesta), y, por lo tanto, tal análisis podría dar lugar a importantes sesgos en los resultados. En segundo lugar, en los análisis multivariantes, trabajar sólo con las observaciones para las cuales todas las variables de interés se han completado produciría muestras demasiado pequeñas.

- *Imputación para permitir el análisis de la EFF con métodos de datos completos.* Se pueden realizar inferencias correctas de un conjunto de datos incompletos utilizando, por ejemplo, métodos de máxima verosimilitud basados en modelos estadísticos. Sin embargo, esto no está técnicamente al alcance de todos los usuarios potenciales de los datos. Por lo tanto, resulta conveniente proporcionar a los usuarios de los datos algunas imputaciones de (es decir, “rellenar”) los datos que faltan, que, por supuesto, los analistas pueden ignorar si lo desean¹³.

El objetivo de la imputación no es generar información artificialmente ni dar la impresión de que el conjunto de datos contiene más información de la que realmente tiene, sino explotar exhaustivamente la existente, de tal manera que se puedan efectuar los distintos análisis posibles utilizando herramientas de datos completos.

- *Imputación como responsabilidad del proveedor de los datos.* La imputación es un proceso que requiere muchos recursos, que no están a disposición de la mayoría de los usuarios y que se considera, correctamente, responsabilidad de los proveedores de datos [véase Rubin (1996)]. Un motivo adicional, muy importante en el caso de la EFF, para que el Banco de España realice una imputación es que esta institución puede acceder a determinada información (como algunas variables de estratificación y ubicación) relevante para la imputación de valores razonables que no estará incluida en el fichero de datos públicos, por motivos de confidencialidad.

Elección del método de imputación

Antes de explicar la elección del método de imputación, debe señalarse que todos los métodos dependen del supuesto de “ausencia aleatoria de respuesta” [definida en Rubin (1976) y en Little y Rubin (1987)], que requiere que los valores sin respuesta se comporten como una muestra aleatoria de todos los valores, pero dentro de grupos definidos por datos observados. La bondad de este supuesto dependerá de la disponibilidad de variables observadas que puedan explicar plausiblemente la ausencia de respuesta y que permitan efectuar el análisis condicionando en ellas.

Una de las principales motivaciones para realizar la EFF fue conocer la distribución de los activos reales y financieros de los hogares, su endeudamiento y sus relaciones con otras variables. Para preservar la distribución observada de las variables y las covarianzas entre ellas, deberán utilizarse métodos de imputación estocásticos. De hecho, métodos sencillos como la imputación de la media (condicional o no condicional) tienden a producir distribuciones concentradas de las variables y subestimaciones de las varianzas.

Un método muy popular de imputación estocástica es el método *hot deck*, con distintas variantes. En general, con un procedimiento *hot deck*, el valor sin respuesta de un determinado hogar se sustituye por el valor de la pregunta contestada por algún hogar con características similares. Sin embargo, en el caso de la EFF el número de características/variables sobre las que

¹². Las referencias para esta sección (excepto la última parte) son Little y Rubin (1987), Rubin (1987) y Schafer (1997).

¹³. Todos los valores imputados están señalizados de forma correspondiente.

convendría condicionar el análisis para que fuera plausible suponer que la falta de información está “ausente de forma aleatoria” es demasiado elevado para obtener celdas de un tamaño razonable y poder extraer de ellas la imputación *hot deck*. Por lo tanto, la mayoría de las imputaciones de la EFF, como se verá más adelante, se basan en modelos de regresión aleatoria.

Por último, cualquiera que sea el método estocástico que se elija, uno de los problemas que se plantea con el hecho de proporcionar una única imputación (es decir, asignar un valor a cada valor sin respuesta) es que los valores imputados son tratados como si se hubieran observado realmente. Al actuar así no se tiene en cuenta la incertidumbre acerca de la imputación en el modelo considerado ni la posible incertidumbre adicional acerca de la elección de modelo para efectuar la imputación. Así pues, proporcionar múltiples imputaciones (MI), como propone Rubin (1987), es un objetivo deseable en el proceso de imputación.

La idea que subyace a la imputación múltiple es que para cada valor sin respuesta, se proporcione varios valores imputados (digamos m) en lugar de uno solo¹⁴. Un conjunto de datos imputados de forma múltiple da lugar a m conjuntos de datos completos. La manera de utilizar el conjunto de datos MI es (i) analizar en primer lugar, por separado, los m conjuntos de datos imputados, utilizando herramientas de datos completos, y (ii) dos, combinar los resultados. Una preocupación inicial fue en nuestro caso que los datos imputados de forma múltiple pudieran desincentivar a los posibles usuarios de los datos. Aunque los pasos que se han de seguir para utilizar un conjunto de datos MI son, en principio, sencillos, la combinación de resultados podría desalentar al usuario menos técnico, para el que, en parte, se ha realizado el proceso de imputación, lo que frustraría parcialmente el objetivo de la imputación. No obstante, hoy en día, software extensamente utilizado, como Stata, puede programarse fácilmente para realizar esas tareas [véase Carlin et al. (2003)].

Software utilizado en la imputación

Se han podido utilizar los programas de imputación múltiple escritos por Arthur Kennickell, de la Reserva Federal de Estados Unidos, para el SCF [véase Kennickell (1991 y 1998)], así como beneficiarse de su asesoramiento. Estos programas se adecuan muy especialmente a nuestro caso, ya que ambas encuestas comparten un número importante de características que no pueden incorporarse en otros paquetes informáticos disponibles. En particular, en conjuntos de datos tan complejos como estos, casi cada observación tiene un patrón diferente en cuanto a la ausencia de respuesta a las variables, por lo que no es posible utilizar patrones monótonos, donde las variables se ordenan según su falta de respuesta. Otra importante característica disponible en los programas de imputación del SCF es la posibilidad de imponer restricciones a los valores que han de ser imputados específicas a cada observación.

El programa de imputación múltiple del SCF (Fritz)

En lo que sigue esbozaremos someramente las características principales del programa de imputación múltiple del SCF¹⁵.

El programa de imputación múltiple (Técnica de imputación Zeta de la Reserva Federal o Fritz, en sus siglas en inglés) tiene una estructura secuencial e iterativa. En una iteración dada, las variables son imputadas secuencialmente, y una variable imputada se toma como “observada” para imputaciones posteriores en la secuencia y en las siguientes iteraciones (pero sujeta a actualización). El orden de imputación de las variables en la secuencia tiene importancia, por lo que hay que imputar en primer lugar las variables con no demasiados casos de ausencia de respuesta que son relevantes como estadísticos suficientes para predecir otras variables. Esta

¹⁴. Habitualmente es suficiente con m entre 2 y 10. Por ejemplo, [véase Schafer (1997)] con el 50% de información no completada, una estimación basada en $m=5$ tenderá a tener un error estándar solamente 1,049 veces mayor que el estimado con $m=\infty$.

¹⁵. En Kennickell (1991, 1998 y 2000) y Kennickell y McManus (1994) puede encontrarse más información.

imputación iterativa y secuencial está relacionada con el desarrollo de algoritmos de simulación de cadenas de Markov [*Markov Chain Montecarlo (MCMC)*], en particular al muestreo de Gibbs.

El programa Fritz permite tres tipos de imputación: continua, binaria y multinomial. Para las variables continuas, las imputaciones se efectúan añadiendo un factor aleatorio a las predicciones mediante regresión. Una característica importante del programa es que permite especificar todas las variables que se querían utilizar como regresores. No obstante, en la primera iteración, en lugar de utilizar solamente los coeficientes de una regresión basada únicamente en casos completos, para cada variable y observación que se ha de imputar, el programa determina las variables con respuesta entre el conjunto completo de regresores y utiliza el subconjunto correspondiente de filas y columnas de la matriz de covarianzas necesario para esta regresión “individual”. En iteraciones posteriores, se dispone de la matriz de covarianzas completa utilizando datos imputados de la iteración anterior.

Para las variables binarias, se utiliza una variante del modelo de probabilidad lineal, para aprovechar la regresión “individual” basada en las covarianzas que se acaba de describir. Por último, en el caso de las variables multinomiales se utiliza un procedimiento de *hot deck* que permite condicionar en una variable discreta y en una variable continua, como máximo¹⁶.

Algunos asuntos prácticos

- *Evaluación de las imputaciones.* El conjunto de variables utilizadas en la regresión deberá ser tan general como sea posible, y tendrá que incluir todos los posibles determinantes de la variable que se ha de imputar, así como, en el sistema Fritz, predictores de estos determinantes, en caso de que estos falten. Para que se satisfaga la hipótesis de la ausencia aleatoria de respuesta deberá incluirse, también, cualquier variable potencialmente relacionada con el patrón de ausencia de respuesta. El deseo de contar con un modelo tan completo como sea posible, tiene que equilibrarse, en la práctica, con la limitación impuesta por los grados de libertad.

No es fácil saber cómo evaluar la bondad de los valores imputados de una variable dada. En primer lugar, no se dispone de un único estadístico de bondad del ajuste del modelo de regresión utilizado en la imputación para la muestra de valores observados de la variable, dada la regresión casi “individual” empleada para cada caso sin respuesta (como se ha explicado anteriormente). Además, un buen ajuste dentro de la muestra (es decir, para los valores observados) no tiene que ser ni necesario ni suficiente para imputaciones convincentes fuera de la muestra (es decir, de los que no responden).

En segundo lugar, para considerar si las variables imputadas son “razonables”, sería deseable poder comparar la distribución de los valores observados de la variable (Y^0) con la distribución de los valores imputados (Y^I). Sin embargo, para una comparación adecuada, las distribuciones relevantes deberían ser condicionales a las variables (al menos a algunas de ellas) utilizadas en la regresión de imputación. El problema con tales comparaciones de las distribuciones condicionales es la ausencia de observaciones suficientes.

Dadas estas consideraciones, se han aplicado procedimientos de los vecinos más cercanos para verificar las imputaciones extremas¹⁷. Para cada valor imputado Y_i^I con un vector de variables asociado X_i , se define un conjunto de observaciones de vecinos que responden (J) como aquellos para los que el vector de variables observadas es próximo a X_i , y se compara Y_i^I con el conjunto de Y_j^0 de los vecinos. Formalmente, los vecinos quedan definidos

¹⁶. La variable continua puede definirse como el resultado de la interacción de más de una variable, permitiendo así una condicionalidad más amplia.

¹⁷. Tras la primera iteración, para tener valores iniciales razonables, y tras la última.

como los que responden con $\|\underline{X}_i - \underline{X}_j\| \leq \varepsilon$ para un ε elegido adecuadamente¹⁸. La norma que se utiliza es

$$\|\underline{X}_i - \underline{X}_j\| = \sqrt{(X_{i1} - X_{j1}, \dots, X_{ik} - X_{jk}) \Omega^{-1} (X_{i1} - X_{j1}, \dots, X_{ik} - X_{jk})'}$$

donde Ω es la matriz de covarianzas de las variables correspondientes. Una vez que se selecciona un conjunto de vecinos, se verifica si Y_i^l se encuentra entre el máximo Y_j^0 más una desviación típica residual (y el mínimo Y_j^0 menos una desviación típica). En caso de que esto no se satisfaga, el valor imputado se redefine como el máximo (o el mínimo) de los vecinos.

Si se observa que el número de vecinos es demasiado pequeño para una evaluación sólida, se repite el proceso incluyendo en el conjunto de vecinos potenciales no solo a los que responden, sino también a las imputaciones obtenidas^{19,20}.

• *Convergencia del proceso iterativo.* Otra consideración práctica es cómo controlar si las imputaciones convergen tras un cierto número de iteraciones. Para ello, en la EFF, para cada variable Y a la cual se le han imputado algunos valores, comparamos las estimaciones de su mediana y de su desviación intercuartílica resultantes de iteraciones sucesivas. En particular, el criterio que utilizamos es

$$\sqrt{(Q50_t^Y - Q50_{t-1}^Y, (Q75 - Q25)_t^Y - (Q75 - Q25)_{t-1}^Y) (Q50_t^Y - Q50_{t-1}^Y, (Q75 - Q25)_t^Y - (Q75 - Q25)_{t-1}^Y)'}$$

donde t es el número de la iteración y $Q25, Q50, Q75$ los percentiles 25, 50 y 75, respectivamente. Se lleva a cabo este análisis para la distribución incondicional de cada Y y también condicionando en algunas variables de estratificación.

Dado que en cada iteración se randomiza y se imputa de forma múltiple, se define la convergencia con respecto a medidas de posición y dispersión en vez de a un criterio punto por punto.

18. Se selecciona un valor de ε igual a la raíz cuadrada de la dimensión de \underline{X}_i , salvo cuando se observa el conjunto completo de variables. En tal caso, se puede distinguir fácilmente entre las variables discretas y las continuas y establecer ε en $\varepsilon = \sqrt{n^0 \text{ de variables discretas}} + \sqrt{n^0 \text{ de variables continuas}}$.

19. Se considera que el número mínimo necesario de vecinos es 15.

20. Finalmente, en la quinta y última comparación de vecinos, ε se aumenta en un décimo del número total de variables.

REFERENCIAS

- BROWNING, M., T. F. CROSSLEY, y G. WEBER (2002). *Asking Consumption Questions in General Purpose Surveys*.
- CARLIN, J. B., N. LI, P. GREENWOOD, y C. COFFEY (2003). "Tools for analyzing multiple imputed datasets", *The Stata Journal*, 3, pp. 226-244.
- KENNICKELL, A. (1991). *Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation*.
- (1998). *Multiple Imputation in the Survey of Consumer Finances*.
- (2000). *Wealth Measurement in the Survey of Consumer Finances: Methodology and Directions for Future Research*.
- KENNICKELL, A., y D. McMANUS (1993). *Sampling for Household Financial Characteristics Using Frame Information on Past Income*.
- (1994). *Multiple Imputation of the 1983 and 1989 waves of the SCF*.
- LITTLE, R. J. A., y D. B. RUBIN (1987). *Statistical Analysis with Missing Data*, Wiley, New York.
- RUBIN, D. B. (1976). "Inference and Missing Data", *Biometrika*, 63, pp. 581-592.
- (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley.
- (1996). "Multiple Imputation After 18+ Years", *Journal of the American Statistical Association*, 91, pp. 473-489.
- SAUTORY, O. (1993). *La macro Calmar. Redressement d'un échantillon par calage sur marges*, Document n.º F9310, Institut National de la Statistique et des Études Économiques, Paris.
- SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London.

**Cuadro 1. Número de preguntas planteadas y contestadas
por los hogares, sin ponderar**

	Media	Mediana	Desviación típica	Mínimo	Máximo
Nº de preguntas planteadas	209	206 ¹	56,1	100	462 ²
Nº de preguntas planteadas sobre importes monetarios (€)	23,4	22 ³	9,5	5	73 ⁴
Nº de preguntas contestadas	201,2	198	54,1	91	425
Nº de preguntas contestadas sobre importes monetarios (€)	19,4	18	8,7	3	66
% de preguntas contestadas	96,2	96,7	2,4	75,7	100
% de preguntas contestadas sobre importes monetarios (€)	82,7	85,7	14,4	30,8	100

Los datos siguientes proceden de Kennickell (2000):

1. 434 en el SCF de 1998
2. 784 en el SCF de 1998
3. 35 en el SCF de 1998
4. 107 en el SCF de 1998

Cuadro 2. Definición de los estratos de riqueza

Estrato 1	no presentan declaración de patrimonio
Estrato 2	menos de 20 millones de pesetas ¹ (120.200 €)
Estrato 3	de 20 a 50 millones de pesetas (de 120.200 a 300.500 €)
Estrato 4	de 50 a 100 millones de pesetas (de 300.500 a 601.000 €)
Estrato 5	de 100 a 200 millones de pesetas (de 601.000 a 1.202.200 €)
Estrato 6	de 200 a 500 millones de pesetas (de 1.202.200 a 3.005.500 €)
Estrato 7	de 500 a 2.000 millones de pesetas (de 3.005.500 a 12.022.000 €)
Estrato 8	más de 2.000 millones de pesetas (12.022.000 €)

1. La definición de los intervalos es en pesetas porque esa era la moneda en que estaban expresadas las declaraciones de impuestos del año 1999, antes de la introducción del euro.

Cuadro 3. Número de intentos, por tipo de respuesta

Encuestas cumplimentadas	5.143
Negativas a contestar a la encuesta	5.722
Nunca en casa	6.670
Fuera del ámbito de la encuesta (dirección equivocada, no era un hogar, vivienda vacía, fallecidos, otros fuera de ámbito)	1.797
Encuestas descartadas tras la supervisión	569
Total	19.901

Cuadro 4. Algunas medidas de no respuesta, por estratos de riqueza

	Nunca en casa ¹	Tasa de cooperación ²
Total	33,5%	47,3%
Estrato 1	31%	53,6%
Estrato 2	38,9%	45,3%
Estrato 3	32,9%	44,7%
Estrato 4	35,5%	46,5%
Estrato 5	37%	38,5%
Estrato 6	38%	36,1%
Estrato 7	40,1%	37,8%
Estrato 8	39,8%	29,4%
Navarra y País Vasco	26%	46%

1. Definida como (Nunca en casa/Total intentos)
2. Definida como (Cumplimentadas/Cumplimentadas+Negativas a responder a la encuesta)

**Cuadro 5. Análisis de descomposición de la varianza
y medida de bondad del ajuste**

Variable dependiente = 1 si se cumplimenta, 0 si se se niegan a cumplimentarla Número de observaciones = 10.586				
	Media	Factores geográficos	Factores socioeconómicos	Factores geográficos y socioeconómicos
Desviación típica	0,50	0,48	0,49	0,47
R ²		0,09	0,03	0,11
Asociación de probabilidad predichas y respuestas observadas ¹ :				
Concordancia		65,3%	49,3%	68,5%
Empate		3,1%	18,7%	0,8%

1. Como medida de la bondad del ajuste, se considera la asociación de las probabilidades predichas y las respuestas observadas. Esto mide cuántos pares de observaciones tienen una respuesta que concuerda, es decir, cuántos pares (en nuestra muestra, casi 28 millones de pares) con diferentes respuestas observadas tienen probabilidades predichas que se ordenan de forma coherente con dichas respuestas.

Cuadro 6. Estimaciones logit de los parámetros de la decisión de cumplimentar la encuesta frente a negarse¹

	Coefficiente	Ratio de probabilidades	Ratio-t
Tipo de vivienda			
Lujo	-0,4423742	0,6425092	-2,04
Semilujo	-0,8676473	0,4199384	-3,68
Alta calidad	-1,2784430	0,2784705	-5,33
Baja calidad	-1,5481060	0,2126504	-6,09
Muy mala calidad	-1,2162740	0,2963321	-3,75
Tipo de barrio			
Alta categoría	-0,3941852	0,6742292	-2,00
Por encima de la media	0,1242046	1,1322470	0,60
Media	0,5946163	1,8123350	2,78
Por debajo de la media	0,6797912	1,9734660	2,96
Baja	1,0055460	2,7333990	3,66
Muy baja	1,2245340	3,4025810	2,50
Pobre	-0,7514802	0,4716679	-1,76
Clase social			
Media-alta	0,2801777	1,3233650	2,45
Media-media	0,1091127	1,1152880	0,83
Media-baja	0,5073223	1,6608380	3,39
Baja	1,0015720	2,7225590	4,61
Tamaño del municipio			
2.000<hab=<10.000	-0,7049064	0,4941548	-5,21
10.000<hab=<50.000	-0,7412587	0,4765137	-5,63
50.000<hab=<100.000	-0,9232242	0,3972362	-6,61
100.000<hab=<500.000	-1,2313290	0,2919044	-9,60
500.000<hab=<1.000.000	-1,4425380	0,2363273	-9,90
hab>1.000.000	-1,3295370	0,2645998	-9,44
Comunidad Autónoma			
Aragón	-0,7990338	0,4497633	-6,54
Asturias	0,1853343	1,2036210	1,33
Islas Baleares	-0,7270669	0,4833245	-4,21
Islas Canarias	-0,4181743	0,6582475	-3,53
Cantabria	-1,3713790	0,2537567	-9,46
Castilla-La Mancha	-0,2665401	0,7660253	-1,92
Castilla-León	-0,2544959	0,7753073	-2,36
Cataluña	-0,7376957	0,4782146	-8,84
Valencia	-0,4092502	0,6641480	-4,90
Extremadura	-0,2212221	0,8015386	-1,26
Galicia	-1,3394490	0,2619900	-12,74
Madrid	0,0187715	1,0189490	0,19
Murcia	-0,0833291	0,9200483	-0,53
Navarra	-1,2323730	0,2915997	-5,79
País Vasco	-0,5408224	0,5822692	-4,81
La Rioja	-0,4666641	0,6270907	-2,14
Constante	1,9994390		9,42

1. Categorías omitidas: vivienda de gran lujo, barrio de categoría muy elevada, clase social alta, municipios con 2.000 habitantes o menos, Andalucía.

Cuadro 7. Tasas de respuesta (%) a algunas preguntas, muestra sin ponderar

	Tienen		Valor para los que tienen		
	Sí	No se sabe	Valor	No sabe	No contesta
Vivienda principal en propiedad	84,5	0,0	86,5	13,0	0,5
Importe pendiente, primer préstamo, vivienda principal	15,0	0,0	88,6	11,2	0,3
Pago mensual, 1er préstamo, vivienda principal	15,0	0,0	96,2	3,5	0,1
Vivienda principal en alquiler	9,9	0,0	97,4	1,0	1,6
Otras propiedades inmobiliarias, primera propiedad	41,7	0,0	82,0	16,4	1,0
Importe pendiente, primer préstamo, primera propiedad	5,0	0,0	91,1	6,6	0,8
Cuentas para realizar pagos	96,9	1,5	74,3	11,7	14,0
Cuentas que no pueden utilizarse para realizar pagos	20,8	2,2	81,8	6,5	11,8
Acciones cotizadas	20,7	0,3	76,6	15,9	7,4
Acciones no cotizadas	6,9	0,2	51,3	34,6	14,2
Fondos de inversión, primer fondo	14,7	0,2	76,6	12,8	7,5
Valores de renta fija	3,3	0,2	81,4	11,0	7,6
Planes de pensiones, primer plan	25,8	0,0	62,3	34,6	3,0
Seguro de vida (primera póliza)	8,9	0,0	63,9	33,5	2,6
Valor de mercado del negocio (persona de referencia)	13,1	0,0	64,3	32,3	3,4
Renta salarial (persona de referencia, 2001)	36,9	0,0	97,6	1,2	1,3
Renta como autónomo (persona de referencia, 2001)	13,4	0,0	89,6	5,2	5,2
Prestación por desempleo (persona de referencia, 2001)	1,5	0,0	94,7	5,3	0,0
Pensiones (persona de referencia, 2001)	31,8	0,0	99,2	0,2	0,6
Rentas de los activos reales (2001)	11,1	0,1	92,0	3,3	4,7
Ingresos por dividendos, cupones, etc. (2001)	9,3	0,9	60,7	33,4	5,9
Ingresos por intereses en cuentas bancarias (2001)	65,1	3,6	34,1	60,5	5,4
Gasto en alimentos	100,0	0,0	93,8	5,8	0,4
Gasto en bienes no duraderos	100,0	0,0	95,9	3,6	0,5

PUBLICACIONES DEL BANCO DE ESPAÑA

DOCUMENTOS OCASIONALES

- 0301 GIANLUCA CAPORELLO Y AGUSTÍN MARAVALL: A tool for quality control of time series data. Program
TERROR.
- 0302 MARIO IZQUIERDO, ESTHER MORAL Y ALBERTO URTASUN: El sistema de negociación colectiva en España:
un análisis con datos individuales de convenios. (Publicada una edición en inglés con el mismo número.)
- 0303 ESTHER GORDO, MARÍA GIL Y MIGUEL PÉREZ: Los efectos de la integración económica sobre
la especialización y distribución geográfica de la actividad industrial en los países de la UE.
- 0304 ALBERTO CABRERO, CARLOS CHULIÁ Y ANTONIO MILLARUELO: Una valoración de las divergencias
macroeconómicas en la UEM. (Publicada una edición en inglés con el mismo número.)
- 0305 ALICIA GARCÍA HERRERO Y CÉSAR MARTÍN MACHUCA: La política monetaria en Japón: lecciones a extraer
en la comparación con la de los EEUU.
- 0306 ESTHER MORAL Y SAMUEL HURTADO: Evolución de la calidad del factor trabajo en España.
- 0307 JOSÉ LUIS MALO DE MOLINA: Una visión macroeconómica de los veinticinco años de vigencia
de la Constitución Española.
- 0308 ALICIA GARCÍA HERRERO Y DANIEL NAVIA SIMÓN: Determinants and impact of financial sector FDI to
emerging economies: a home country's perspective.
- 0309 JOSÉ MANUEL GONZÁLEZ-MÍNGUEZ, PABLO HERNÁNDEZ DE COS Y ANA DEL RÍO: An analysis of the
impact of GDP revisions on cyclically adjusted budget balances (CABS).
- 0401 J. RAMÓN MARTÍNEZ-RESANO: Central Bank financial independence.
- 0402 JOSÉ LUIS MALO DE MOLINA Y FERNANDO RESTOY: Evolución reciente del patrimonio de empresas y
familias en España: implicaciones macroeconómicas. (Publicada una edición en inglés con el mismo número.)
- 0403 ESTHER GORDO, ESTHER MORAL Y MIGUEL PÉREZ: Algunas implicaciones de la ampliación de la UE para
la economía española.
- 0404 LUIS JULIÁN ÁLVAREZ GONZÁLEZ, PILAR CUADRADO SALINAS, JAVIER JAREÑO MORAGO E ISABEL
SÁNCHEZ GARCÍA: El impacto de la puesta en circulación del euro sobre los precios de consumo.
- 0405 ÁNGEL ESTRADA, PABLO HERNÁNDEZ DE COS Y JAVIER JAREÑO: Una estimación del crecimiento potencial
de la economía española.
- 0406 ALICIA GARCÍA-HERRERO Y DANIEL SANTABÁRBARA: Where is the Chinese banking system going with the
ongoing reform?
- 0407 MIGUEL DE LAS CASAS, SANTIAGO FERNÁNDEZ DE LIS, EMILIANO GONZÁLEZ-MOTA Y CLARA MIRA-
SALAMA: A review of progress in the reform of the International Financial Architecture since the Asian crisis.
- 0408 GIANLUCA CAPORELLO Y AGUSTÍN MARAVALL: Program TSW. Revised manual. Version May 2004.
- 0409 OLYMPIA BOVER: Encuesta Financiera de las Familias españolas (EFF): descripción y métodos de la encuesta
de 2002. (Publicado el original en inglés con el mismo número.)

BANCO DE ESPAÑA

Unidad de Publicaciones
Alcalá, 522; 28027 Madrid
Teléfono +34 91 338 6363. Fax +34 91 338 6488
e-mail: Publicaciones@bde.es
www.bde.es