Encuesta Social 2017: Novedades metodológicas y resultados

María Escudero Tena

Instituto de Estadística y Cartografía de Andalucía

Resumen: Las importantes transformaciones que la sociedad andaluza ha experimentado en los últimos años, hacen necesario contar con instrumentos que permitan comparar esta realidad con las de otros territorios y realizar un seguimiento de los cambios. En este sentido, el Instituto de Estadística y Cartografía de Andalucía (IECA) realiza la Encuesta Social con el propósito de recoger información de carácter específico en cada edición, así como un grupo de variables sociales básicas de forma continuada.

El objetivo de esta ponencia es presentar las principales novedades metodológicas de las dos últimas ediciones de la Encuesta Social realizadas en 2018: la “Encuesta de Movilidad Social” y la “Encuesta de Educación y transiciones al mercado laboral”.

Como novedad destacamos que para ambas encuestas se ha completado la información recogida con información de registros o sistemas de información gestionados por otros organismos de la administración pública. Esto ha posibilitado la obtención de información de contacto con la muestra, el enriquecimiento de la información de análisis y la reducción de la carga de respuesta de los informantes acortando el número de preguntas de los cuestionarios.

Además podemos destacar el uso de software libre (R) para la realización de la elevación y reponderación de la muestra, la imputación de datos faltantes basada en modelos y el cálculo de los errores de muestreo utilizando el método Bootstrap.

Palabras claves: encuestas a hogares, registros administrativos, reponderación, errores de muestreo, imputación, movilidad social, educación

Introducción

La Ley 4/1989, de 12 de diciembre, de Estadística de la Comunidad Autónoma de Andalucía establece que una de las competencias atribuidas al Instituto de Estadística y Cartografía de Andalucía (en adelante, IECA) es "impulsar y fomentar la investigación estadística que contribuya a mejorar el conocimiento de la realidad social y económica de Andalucía". En el ejercicio de esta competencia, el IECA, desde el mismo inicio de su actividad, ha venido desarrollando acciones innovadoras, tanto en el ámbito de su organización como en el de sus procedimientos operativos, a fin de dotarse de instrumentos e infraestructuras que contribuyan a mejorar la calidad de la información estadística producida y transformarla en conocimiento útil para la generación de valor

En este contexto y con ese fin, el IECA ha impulsado desde 2007 una actividad estadística oficial, la Encuesta Social. En el Plan Estadístico 2013-2020, esta Encuesta social se enmarca en el objetivo específico de suministrar información estadística sobre las condiciones de vida y bienestar de la población. Este modelo de encuesta está diseñado con el propósito de recoger información social de carácter específico en distintas ediciones de la misma, así como un grupo de variables socioeconómicas básicas estandarizadas que se han desarrollado en el seno de un grupo de trabajo de Eurostat (Oficina de Estadística de la Comunidad Europea). Así pues, la Encuesta Social, tal como se concibe inicialmente, es una encuesta modular anual dirigida a individuos u hogares, compuesta de módulos fijos y variables, que tiene como objetivo recoger de manera ágil información sobre distintos temas de interés socioeconómico y permitir su seguimiento a lo largo de sus diversas ediciones gracias a la estructura común articulada en torno a las variables sociales básicas comunes.

El marco operativo determinado en los distintos Planes Estadísticos desde 2007 ha propiciado que la Unidad Central de Encuesta, el equipo responsable de esta operación, haya realizado en los últimos años una serie de encuestas que han contribuido a dar respuesta a los mandatos de los distintos planes estadísticos:

- Encuesta Social 2007: una visión de Andalucía
- Encuesta de necesidades de formación y cualificación en Andalucía 2007
- Encuesta Social 2008: hogares y medio ambiente en Andalucía
- Encuesta Social 2010: educación y hogares en Andalucía
- Encuesta Social 2011: movilidad en las regiones urbanas de Andalucía.
- Encuestas Sociales 2013, 2014 y 2015: Encuesta sobre equipamiento y uso de Tecnologías de la Información y Comunicación en los hogares de Andalucía
- Encuesta Social 2017: movilidad social en Andalucía
- Encuesta Social 2017: educación y transiciones al mercado laboral

La experiencia acumulada en estos años de Encuesta Social, la labor de concepción y desarrollo de estas encuestas, así como la economía de recursos, la mayor disponibilidad de información administrativa, el desarrollo de softwares y plataformas específicas de recogida, han posibilitando una transformación significativa en la metodología y en los procesos de trabajo desarrollados por la Unidad Central de Encuesta en la realización de la Encuesta Social.

La consecución de una serie de hitos han ido incorporando paulatinamente al esquema operativo de la Unidad Central de Encuesta nuevos y mejores procedimientos estandarizados e infraestructuras de base, necesarias para la consecución de sus objetivos en unas condiciones de calidad, celeridad, precisión y rigor técnico acordes con los principios del Código de Buenas Prácticas en las Estadísticas Europeas.

En este documento enumeramos las principales características de las encuestas más recientes, en concreto la Encuesta Social 2017 “Encuesta de movilidad social” y la Encuesta Social 2017: “Educación y transiciones al mercado laboral” y nos detendremos en las innovaciones que se han incorporado en dichas encuestas.

Características generales de las Encuesta Sociales 2017

Como se ha comentado a lo largo de 2018 se han llevado a cabo dos Encuestas Sociales, la encuesta sobre movilidad social en Andalucía y la encuesta sobre educación y transiciones al mercado laboral. A continuación se resumen las características principales de estas operaciones:

Encuesta Social 2017: Movilidad social en Andalucía

Objetivo: *Analizar cómo ha evolucionado la sociedad andaluza en términos de clase social para poder observar hasta qué punto los orígenes de los progenitores han configurado el destino de sus hijos/as en las últimas generaciones*

Población de estudio: *Población de 35-60 años que reside en viviendas principales en la Comunidad Autónoma de Andalucía.*

Tamaño de la muestra: *3.000 encuestas.*

Diseño muestral: *trietápico con estratificación de las unidades de primera etapa. Las unidades de primera etapa son las secciones censales, las de segunda etapa, las viviendas familiares principales. En tercera etapa se selecciona aleatoriamente en cada vivienda una persona de 35 a 60 años.*

Sistema de entrevista: *Encuesta multicanal con método de encuesta telefónica como prioritario y encuesta web como método auxiliar.*

Encuesta Social 2017: Educación y transiciones al mercado laboral en Andalucía

Objetivo: *Conocer la evolución de los jóvenes nacidos en 1994, especialmente en lo relativo a la educación y acceso al empleo. Esta encuesta se trata de una continuación de la Encuesta Social 2010. Educación y hogares en Andalucía, en la que se encuestó a 2.584 personas nacidas en 1994 y a sus respectivos padres.*

Población de estudio: *Alumnos nacidos en 1994 a los que se encuestó en 2010.*

Tamaño de la muestra: *2.584 encuestas.*

Diseño muestral: *Se trata de un panel puro. Es decir, la muestra la componen los alumnos nacidos en 1994 seleccionados en la encuesta de 2010. Las unidades muestrales son de dos tipos:*

1. *Muestra titular: aquellos alumnos que respondieron a la encuesta en el año 2010.*
2. *Muestra de reemplazo: personas nacidas en 1994 que formaban parte de la muestra reserva en 2010 y no respondieron al cuestionario. Esta muestra sólo se utilizaría en el caso en el que no se pudieran completar las encuestas a la muestra titular.*

Sistema de entrevista: *Encuesta multicanal con método de encuesta telefónica como prioritario y encuesta web como método auxiliar.*

Novedades metodológicas Encuestas Sociales 2017

La actividad Encuesta Social 2017 ha incorporado novedades impulsadas por diversos motivos: cambios organizativos, mayor disponibilidad de información administrativa, avances en operaciones del IECA que sirven como input en el proceso de encuesta…. Una sucesión de oportunidades y retos que han revertido en mejoras en las distintas fases del proceso de encuesta, desde el pre-campo hasta el tratamiento de los datos y su difusión.

Fase de Pre-Campo

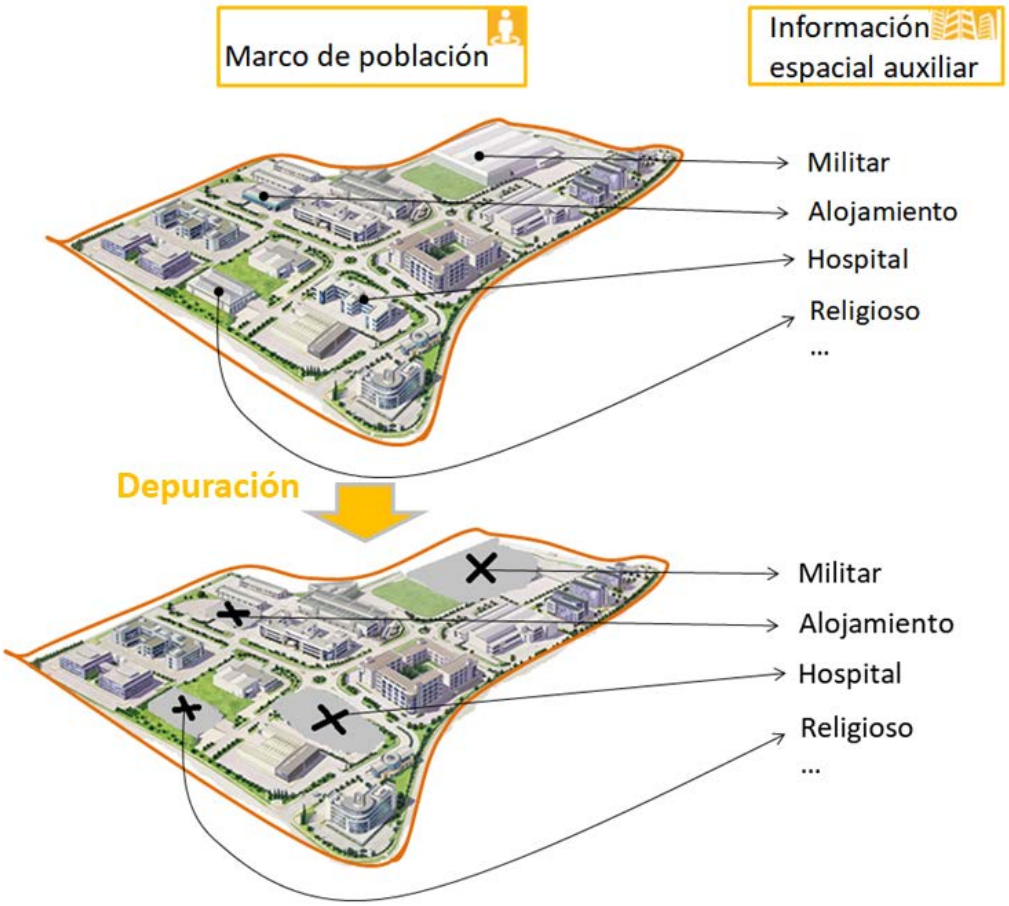
1. **Mejora de la calidad del marco de población objeto de estudio: Construcción de un marco de población y viviendas georreferenciado**

   Desde el año 2011 el IECA desarrolla una intensa labor en el terreno de la integración de la información estadística y cartográfica. La publicación anual desde 2013 del producto "Distribución espacial de la población en Andalucía" es un claro ejemplo de esta línea de trabajo. En esta publicación se muestran los resultados de la georreferenciación de la Base de Datos longitudinal de población de Andalucía. Este esfuerzo de georreferenciación anual de la población a 1 de enero de cada año, permite disponer a la unidad de encuesta de un marco poblacional mejor informado, conteniendo las coordenadas y ubicación en el espacio (dos dimensiones) de las viviendas susceptibles de componer el marco.

   En este caso, la integración de la estadística y la cartografía no sólo ofrece mayores posibilidades de análisis sino que repercute en mejoras de procedimientos intrínsecamente estadísticos como es el proceso de depuración del marco de la población objeto de estudio. Este proceso se ha reforzado y perfeccionado al disponer de la información espacial complementaria procedente del proceso de georreferenciación de la población, pues ha permitido ampliar las fuentes e información de contraste en la identificación de viviendas colectivas. En las Encuestas Sociales precedentes la detección de viviendas colectivas se fundamentaba esencialmente en la información administrativa recabada para la vivienda, esta información podía indicar una ocupación anormalmente elevada y/o una composición demográfica atípica de los habitantes de la vivienda. A partir de la Encuesta Social 2017 se ha incorporado al proceso de identificación de viviendas colectivas, el contraste de la capa espacial del marco de viviendas georreferenciado con capas que contienen información espacial de establecimientos colectivos como edificios religiosos, residencias de ancianos, establecimientos militares, o alojamientos turísticos (campings, balnearios, albergues, paradores, villas turísticas…).

   El contraste se ha realizado mediante uniones espaciales de la capa espacial del marco de viviendas georreferenciado y las capas auxiliares. Estos enlaces se realizan con software GIS, PostGis, empleando un buffer de un metro de distancia entre los portales georreferenciados del marco y los portales correspondientes a establecimientos/viviendas colectivas cuya capa es de tipo "point". Esta holgura tiene en cuenta los desajustes o diferencias que puedan existir entre

capas y que pueden dificultar la unión exacta de los portales. Para la capa de espacios militares, que es de tipo polígono, el criterio para realizar el enlace espacial es determinar aquellos edificios contenidos en el espacio militar y además que tengan un número suficientemente elevado de personas (>=9 miembros).



Incorporar este input en el proceso de depuración del marco poblacional tiene un doble efecto sobre la producción estadística y cartográfica, enriquece y perfecciona el proceso de identificación de viviendas colectivas y a su vez permite identificar incidencias o falta de información en la información cartográfica auxiliar.

La disponibilidad de un marco de viviendas georreferenciado también permite la actualización inmediata del seccionado censal, de cara a un proceso de selección de muestra habitual, multietápico, en el que las secciones censales son las unidades de primera etapa para posteriormente seleccionar entre las viviendas de la sección.

2. Mejora de la calidad del marco con información de contacto telefónico de origen administrativo

En ediciones previas de la encuesta, como se ha mencionado anteriormente, se contó con información de contacto, principalmente teléfonos, procedentes de diversas fuentes administrativas para completar la información del marco de población mediante enlace de registros (combinación de información que proviene de registros disponibles en bases de datos informatizadas (Winkler 2006; pp.1)). Desde la Encuesta Social 2013, en la que se cambió el modelo de encuesta multicanal a un modelo bicanal CATI-CAWI, la disponibilidad y actualización de estos datos de contacto es un elemento crucial para el correcto desarrollo y la calidad de la encuesta.

En las ediciones de 2013, 2014 y 2015 no fue posible la actualización anual de esta información, contando con la información telefónica disponible en la Base de Datos de Usuarios del Sistema Sanitario Público de Andalucía (BDU) para 2012. Esto implicó que en la edición de la Encuesta Social 2015 la muestra recogida presentara la menor tasa de unidades principales de las tres ediciones. De cara a minimizar esta circunstancia en la Encuesta Social 2017 y dado el acceso limitado o parcial a la BDU, se optó por realizar el enlace exclusivamente de los elementos seleccionados de la muestra, es decir, se seleccionó la muestra a partir de un marco sin teléfonos actualizados y posteriormente se enlazó la muestra seleccionada con los teléfonos disponibles en la BDU. Los resultados de este proceso han sido los siguientes en cada una de las encuestas:

Encuesta Social 2017: Movilidad social en Andalucía

El marco de población utilizado para extraer la muestra procede de la Base Longitudinal de Datos de Población de Andalucía a fecha 1 de enero de 2017. De este marco, formado por viviendas y personas residentes en Andalucía con edades comprendidas entre 35 y 60 años, se extrajo la muestra. Posteriormente, esta muestra se enlazó con la información procedente de la Base de Datos de Usuarios (BDU) del Sistema Sanitario Público de Andalucía para obtener los números de teléfono de las unidades muestrales seleccionadas. Como se observa en la siguiente tabla, tras el enlace quedaron 316 personas sin teléfono, a las que se intentó contactar mediante correo postal.

	Principales	%
Sin teléfono	316	**10,5%**
El teléfono es de la persona solicitada	2.546	**84,9%**
El teléfono es de un hijo de la persona solicitada	98	**3,3%**
El teléfono es de la pareja (matrimonio) de la persona solicitada. Se exige convivencia a 1 de Enero	40	**1,3%**
Total	**3.000**	**100%**

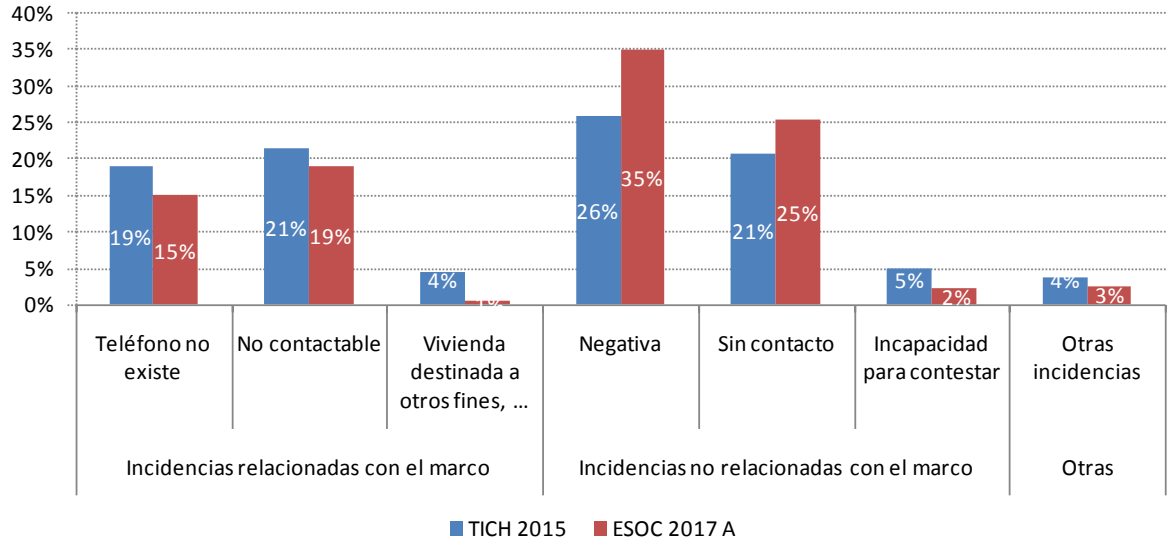
Encuesta Social 2017: Educación y transiciones al mercado laboral en Andalucía

Al tratarse de una encuesta de tipo panel, el marco lo constituyen las 2.584 personas que respondieron a la encuesta previa realizada en el año 2010. Este marco se enlazó con la información procedente de la Base de Datos de Usuarios (BDU) del Sistema Sanitario Público de Andalucía para obtener los números de teléfono actualizados de las unidades muestrales. Como se observa en la siguiente tabla, tras el enlace sólo quedaron 5 personas sin teléfono, a las que se intentó contactar mediante correo postal.

	Principales	%
Sin teléfono	5	**0,2%**
El teléfono procede de la BDU del SSPA actual	2.366	**91,6%**
El teléfono es el que disponíamos de la familia en la encuesta de 2010	208	**8,0%**
La persona ha fallecido	5	**0,2%**
Total	**2.584**	**100%**

Este esquema de actualización de datos de contacto, junto con la depuración de viviendas colectivas, ha supuesto un incremento notable en la calidad del marco muestral para la recogida telefónica que tiene efectos positivos en la calidad final de la muestra recogida. El análisis de las incidencias que han provocado el reemplazo de unidades muestrales en las últimas Encuestas Sociales, muestra una disminución de aquellas relacionadas con errores del marco, como "No contactable" (cuando el teléfono no corresponde a la persona en muestra)[1], "Teléfono no existe"[2] o "Vivienda destinada a otros fines, …"[3]. Por otro lado, la incidencia "Negativa a colaborar", se presenta como la principal causa de sustitución en encuestas telefónicas. Esta incidencia siempre tiene relevancia (Díaz de Rada, 2010; pp.58)

[1] La incidencia "No Contactable" incluye igualmente situaciones que son errores propios del marco no relacionados con el teléfono (por ejemplo, cuando la persona ha cambiado de residencia).
[2] En la incidencia "Teléfono no existe" sólo se han incluido aquellas unidades muestrales que de partida disponían de un teléfono de contacto, pero al intentar realizar la llamada telefónica se nos informa que dicho teléfono no existe.
[3] La incidencia "Vivienda destinada a otros fines,… también incluye otras situaciones minoritarias como desavenencias por causa de divorcio, larga estancia en hospital, estancia en centro penitenciario,…etc.

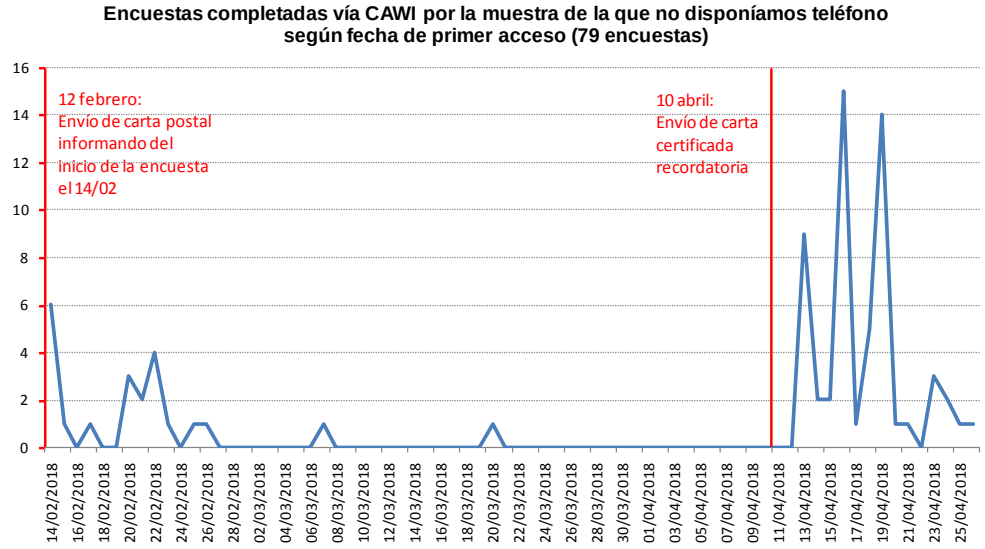


Fase de campo

En ambas encuestas se desarrollaron procesos específicos encaminados a mejorar la calidad de la información recogida, en concreto

3. Refuerzo de contacto vía postal con la unidades sin teléfono identificado

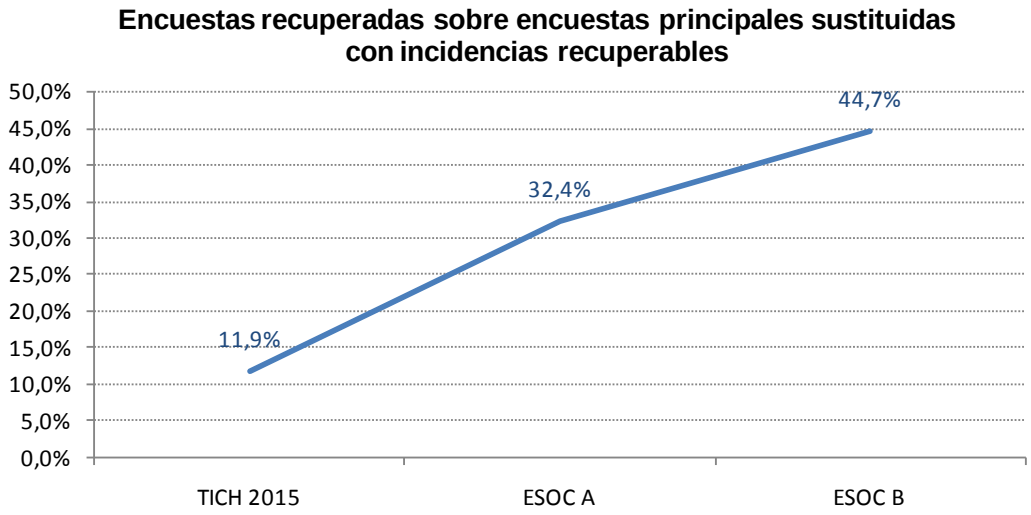
Aquellas unidades muestrales para las que no se disponía de teléfono de contacto en la Encuesta de movilidad social en Andalucía, fueron contactadas vía postal el 12 de febrero informando del comienzo de la encuesta el día 14 de febrero y ofreciendo la posibilidad de cumplimentarla vía web. Para el conjunto de las 316 viviendas principales sin teléfono, ante la imposibilidad de contacto telefónico, se envió un recordatorio vía carta certificada el 10 de abril. En el siguiente gráfico se observa que el impacto que tuvo esta segunda carta fue muy superior. En total, de las 316 viviendas sin teléfono, completaron la encuesta vía CAWI 79 (25%).



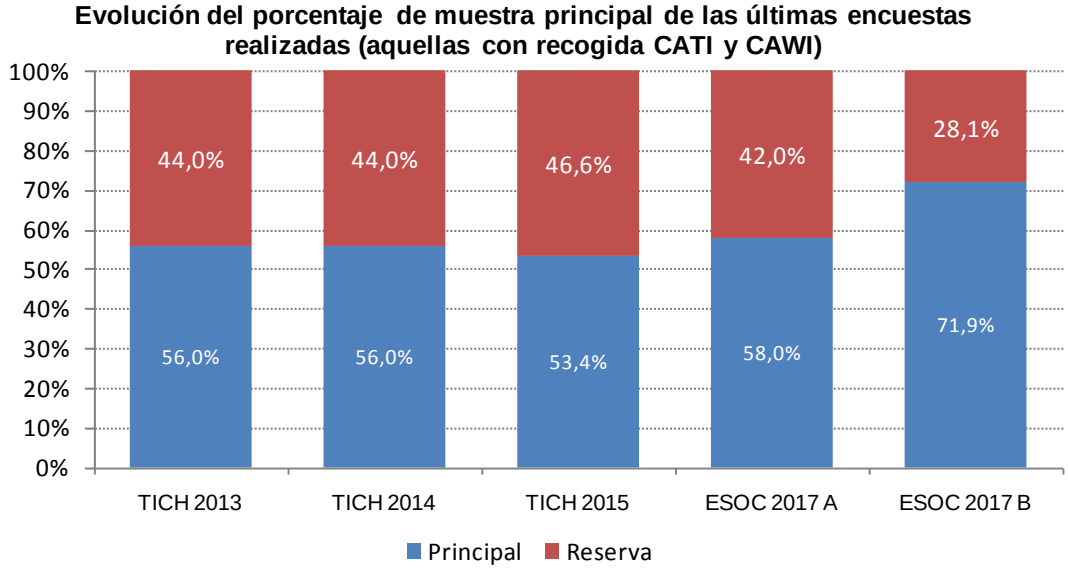
4. Antelación y refuerzo en labores de recuperación

En las Encuestas Sociales 2017 las labores de recuperación, principalmente de negativas a colaborar, han sido muy productivas. En esta ocasión comenzaron en las primeras semanas de campo y en el caso de la Encuesta de educación y transiciones al mercado laboral se extendieron más allá del periodo de recogida. Dadas las características de esta encuesta, panel puro con vocación de recontacto en futuras oleadas, se ha realizado un esfuerzo adicional, en términos de personal y tiempo dedicado a estas labores de recuperación. El resultado de estos esfuerzos de recuperación ha sido positivo, alcanzando tasas de recuperación de negativas y otras incidencias recuperables[4] superiores a las alcanzadas en campos previos como puede verse en el siguiente gráfico

[4] Se consideran incidencias recuperables las siguientes: "Negativa", "Comunica", "Nadie atiende el teléfono", "Nos atienden al teléfono pero no está en casa la persona con la que tenemos que contactar" o "El teléfono marcado tiene las llamadas restringidas"



Las innovaciones y mejoras incorporadas en la operativa por la unidad de encuesta durante las fases de pre-campo y campo, se han reflejado en la calidad de los datos recogidos como se puede apreciar en el siguiente gráfico evolutivo de Encuestas Sociales



Fase de post-campo

5. El uso de registros administrativos para completar la información recogida de la encuesta

Desde las primeras décadas del siglo XX se han depurado en gran medida los métodos para vincular fuentes de datos, lo que ha permitido realizar en muchos países estudios fundamentados en registros administrativos, de gran calado analítico.

En el ámbito estadístico, los países con larga tradición en el uso de registros administrativos fueron los pioneros en incorporar éstos a los estudios estadísticos. Los registros administrativos, como se ha presentado anteriormente, pueden emplearse en la mejora de marcos, diseños muestrales, como sustitución o proxy de variables objeto de estudio o incluso ser la fuente exclusiva de una operación estadística. Actualmente encontramos ejemplos en numerosas operaciones en el ámbito europeo y español del uso de registros administrativos como fuente principal de la estadística o como fuente auxiliar. La Muestra Continua de vidas laborales, la EPA, la encuesta de condiciones de vida o la encuesta anual de estructura salarial son ejemplos de distintos tipos de tratamiento de registros administrativos con fines estadísticos.

La motivación y objetivo principal de la incorporación de la información procedente de registros administrativos a la información muestral es aliviar la carga de respuesta de los informantes acortando el número de ítems y preguntas en los cuestionarios, así como el enriquecimiento de la información de análisis. En el caso de la Encuesta Social 2017: Movilidad social en Andalucía se ha enlazado la muestra efectiva con información procedente de registros administrativos. En el diagrama siguiente se muestran las fuentes empleadas:

Información residencial procedente de la Base de Datos Longitudinal de Andalucía (BDLPA):

- Lugar de nacimiento del entrevistado/a en relación al municipio de residencia actual
- Tiempo que el entrevistado/a lleva residiendo en el municipio actual
- Tiempo que el entrevistado/a ha estado residiendo fuera de Andalucía en los últimos 15 años
- Nº de veces que el entrevistado/a ha cambiado de municipio en los últimos 15 años dentro de Andalucía

Afiliaciones a la Seguridad Social:

- Indicador de si el entrevistado está registrado en el fichero de afiliaciones o no en el último periodo de 2016.
- Régimen en el que está en alta o situación asimilada al alta en la Seguridad Social
- Tipo de contrato. Identifica la modalidad contractual que vincula al trabajador por cuenta ajena y al empleador.
- Grupo de cotización
- Número de días de alta de la demanda de empleo.
- Número de trabajadores de la empresa

Demandantes de empleo:

- Situación de demanda del trabajador.
- Número de días transcurridos desde la última situación registrada de demanda del trabajador.
- Causa de exclusión de paro. La OM de 11 de marzo de 1985 establece la definición del paro registrado como las demandas de empleo pendientes de satisfacer el último día del mes excepto las que se encuentran en determinadas situaciones que detalla en la variable.
- Situación de demanda según el paro registrado.

6. Cálculo de errores de muestreo utilizando el método bootstrap

No existe un acuerdo internacional acerca de la mejor metodología para estimar la varianza en encuestas a hogares, excepto en lo que se refiere al uso de métodos indirectos. Asimismo, los diseños muestrales complejos y multietápicos hacen que el cálculo de los errores de muestreo, su estimador y la implementación no sea una decisión evidente. Los métodos que utilizan las distintas oficinas estadísticas, por tanto, son diversos como puede verse en la siguiente tabla:

País	Método
Alemania	Linealización de Taylor
Canadá	Jacknife / Bootstrap
España	Semimuestras repetidas / Jacknife
Francia	Estimación directa de fórmulas teóricas
Holanda	Bootstrap generando pseudo-poblaciones a partir de la muestra
Italia	Varianza del estimador GREG
Polonia	Bootstrap
Portugal	Jacknife
Reino Unido	Métodos de linealización

Fuentes: Eurostat, StatisticsCanada, Instituto Nacional de Estadística (INE)

Dentro de los métodos bootstrap, basados en el remuestreo, hay distintas opciones de implementación. En esta línea y para la Encuesta Social 2017 valoramos las opciones que se exponen a continuación.

Método bootstrap para diseños multietápicos según Rao and Yue (1992)

Encontramos aplicaciones de este método en Yeo et al (1999) para la Encuesta Nacional de Salud de Canadá o en Azor et al (2011) para la Encuesta de Población Activa (EPA) del Instituto Nacional de Estadística (INE). Sea θ el parámetro que queremos estimar y $\hat{\theta}$ su estimador calculado con la muestra completa. Para estimar la varianza de $\hat{\theta}$, seleccionamos de forma repetida muestras bootstrap de la muestra completa y aplicamos el siguiente procedimiento en cada paso b=1,…,B, donde B es lo suficientemente grande (B=1.000 en nuestro ejercicio),:

1. Independientemente en cada estrato n, h=1…H, seleccionamos una muestra bootstrap mediante una muestreo aleatorio simple de n_h^* unidades primarias de muestreo con reemplazamiento de la muestra de n_h unidades primarias de muestreo (secciones censales en nuestro caso)
2. Para cada unidad secundaria de muestreo (vivienda en nuestro caso) k de cada unidad primaria de muestreo hi (sección censal), se calcula el peso inicial bootstrap como sigue:

   $d_{hik,b}^* = d_{hik}\left\{\left(1-\sqrt{\frac{n_h^*}{n_{h-1}}}\right)+\sqrt{\frac{n_h^*}{n_h-1}}\cdot\frac{n_h}{n_h^*}\cdot t_{hi,b}^*\right\}$, donde d_{hik} es el peso de diseño de la unidad hik, igual a la inversa de la probabilidad de sección $d_{hik} = {1}/{\pi_{hik}}$
3. Para cada unidad secundaria de muestreo (vivienda), el peso final bootstrap $w_{hik,b}^*$ se calcula aplicando, al peso inicial bootstrap $d_{hik,b}^*$, el mismo procedimiento de ajustes (corrección por falta de respuesta y calibración) que se aplicaron a los pesos de la muestra completa d_{hik} para obtener el peso final de la encuesta w_{hik}
4. De la misma forma que se calcula $\hat{\theta}$ para la muestra completa, obtenemos $\hat{\theta}_b^*$ para cada muestra bootstrap, reemplazando w_{hik} por $w_{hik,b}^*$ en la fórmula de $\hat{\theta}$

La varianza bootstrap del estimador $\hat{\theta}$ es dada por

$$v_{boot}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \overline{\hat{\theta}^*}\right)^2,$$

$$\overline{\hat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

Si se toma $n_h^* \leq n_h - 1$ entonces los pesos bootstrap son no negativos. Normalmente se toma $n_h^* = n_h - 1$, ya que simplifica la fórmula para calcular los pesos bootstrap $d_{hik,b}^* = d_{hik}\left\{\frac{n_h}{n_h^*} \cdot t_{hi,b}^*\right\}$

Método bootstrap generando pseudopoblaciones según Booth et al (1994)

Este método es utilizado, por ejemplo, por el Instituto de Estadística de Holanda en Kuijvenhoven and Scholtus (2011). Se basa en la generación de pseudopoblaciones a partir de la muestra, donde cada unidad muestral se repite d_k veces, sean $d_k = {}^{1}/_{\pi_k}$ los pesos de diseño calculados como la inversa de la probabilidad de sección de la unidad muestral. Posteriormente, se aplica el diseño muestral original a las pseudopoblaciones generadas, y se calcula el estimador original a cada muestra bootstrap obtenida.

Estos son los 4 pasos que componen el algoritmo.

1. Dado que los pesos de diseño pueden contener decimales, una forma estocástica de redondearlos sería a partir de $d_k = \lfloor d_k \rfloor + \varphi_k \ (con\ \varphi_k \in [0,1))$. Por tanto se redondearía hacia abajo $\delta_k = \lfloor d_k \rfloor$ con probabilidad 1-φ_k y hacia arriba $\delta_k = \lfloor d_k \rfloor + 1$ con probabilidad φ_k. Posteriormente, se genera una pseudopoblación $\hat{\mathcal{U}}$ replicando δ_k veces cada elemento k de la muestra original s
2. Se extrae una muestra s^* de $\hat{\mathcal{U}}$ con el diseño muestral original, y se obtienen los pesos bootstrap siguiendo el mismo procedimiento que para la muestra completa: calculando los pesos iniciales como inversa de la probabilidad de selección $d_k^* = {}^{1}/_{\pi_k^*}$ y aplicando el mismo procedimiento de ajustes (corrección por falta de respuesta y calibración) que se aplicaron a los pesos de la muestra completa para obtener los pesos bootstrap $w_{k,b}^*$. Finalmente obtenemos el estimador $\hat{\theta}_b^*$ para cada muestra bootstrap con los pesos obtenidos tras los ajustes $w_{k,b}^*$
3. El paso 2 se repite C veces (C=1.000 veces en nuestro caso) y se obtienen $\hat{\theta}_1^*, \ldots, \hat{\theta}_C^*$. Se calcula:

$$v_{boot}^b = \frac{1}{C}\sum_{c=1}^{C}\left(\hat{\theta}_c^* - \overline{\hat{\theta}^*}\right)^2,$$

$$\overline{\hat{\theta}^*} = \frac{1}{C}\sum_{c=1}^{C}\hat{\theta}_c^*$$

4. Los pasos 1 a 3 se repiten B veces (B=1 en nuestro caso) y se obtienen $v_{\text{boot}}^{1}, \ldots, v_{\text{boot}}^{B}$. Finalmente la varianza bootstrap del estimador $\hat{\theta}$ es dada por:

$$v_{boot}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} v_{boot}^{b}$$

Los resultados obtenidos con ambos métodos para una de las tablas de la Encuesta de Movilidad Social son los siguientes.

	Errores relativos según Bootstrap para diseños multietápicos Rao and Yue (1992)			Errores relativos según algoritmo Bootstrap generando pseudo-poblaciones Booth et al (1994)		
Clase social individuo	**Ambos sexos**	**Hombres**	**Mujeres**	**Ambos sexos**	**Hombres**	**Mujeres**
I+II	**4,4**	6,1	5,5	**3,7**	5,7	5,0
III	**3,2**	5,4	3,8	**2,9**	5,3	3,6
IV ab	**5,1**	6,9	8,1	**4,7**	6,0	7,6
IV c	**16,1**	16,4	43,8	**14,4**	15,5	42,3
V + VI	**5,0**	5,0	15,4	**4,5**	4,8	13,5
VII a	**5,8**	11,0	6,6	**4,7**	10,6	5,5
VII b	**7,6**	10,0	10,0	**5,8**	8,8	8,2

Las cifras se muestran en términos de errores relativos en porcentaje (coeficiente de variación):

$$\widehat{CV}(\hat{\theta}) = \frac{\sqrt{v_{boot}(\hat{\theta})}}{\hat{\theta}} \cdot 100$$

Como se puede observar en general se obtienen resultados similares con ambos métodos, siendo los errores relativos resultantes del método bootstrap generando pseudopoblaciones ligeramente inferiores.

7. El uso de software libre para la reponderación

Los estimadores utilizados en esta encuesta son estimadores basados en el diseño de la muestra, corregidos por la falta de respuesta a nivel de estrato y por información auxiliar facilitada por fuentes externas. Es decir, los pesos de diseño obtenidos a partir del tipo de muestreo utilizado, corregidos por la falta de respuesta, se han calibrado posteriormente mediante técnicas de reponderación. El objetivo de la utilización de estas técnicas es ajustar las estimaciones de la encuesta a la información demográfica procedente de fuentes externas. En esta encuesta se ha utilizado como fuente externa la Encuesta continua de hogares a 1 de enero de 2017 (INE. Instituto Nacional de Estadística). En concreto, las variables auxiliares que se han utilizado son las siguientes:

- Población por sexo y grupos de edad quinquenales
- Población por grupos de nacionalidad
- Población por provincias

En ediciones anteriores se ha contado con el paquete CALMAR del software SAS para la realización del proceso de calibrado. En esta ocasión, siguiendo la línea de política digital de la Junta de

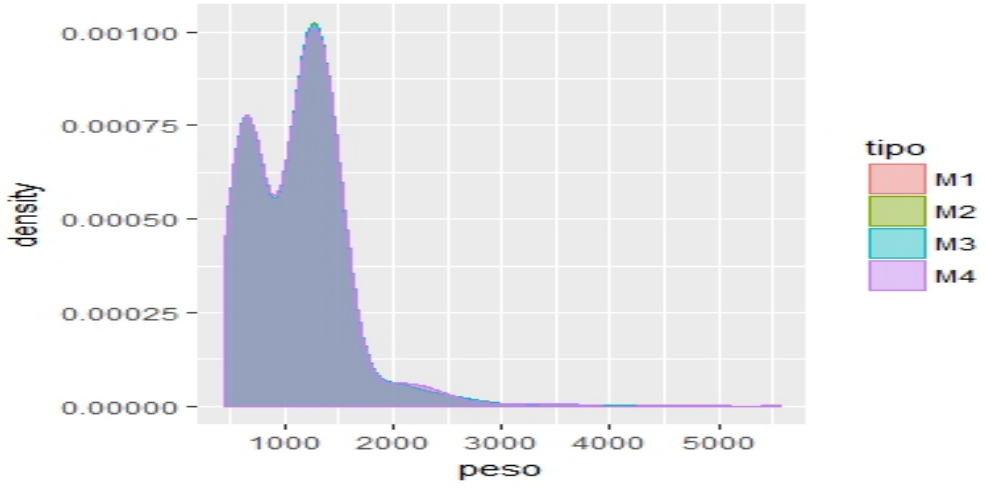
Andalucía, este proceso se ha implementado por primera vez en la Encuesta Social, a partir del software R.

A continuación se relacionan algunos de los paquetes de R que se utilizan para calibrar:

- Sampling (Yves Tillé 2016)
- Icarus (Antoine Rebecq, 2017). Desarrollado por la Oficina Estadística de Francia (INSEE), al igual que CALMAR
- Calif. Herramienta de la Oficina Estadística de Eslovaquia
- Otros: Laeken, Simpop, Survey

A modo de ejercicio, se replicaron los calibrados de las tres últimas Encuestas Sociales con el paquete CALMAR del software SAS y los paquetes Sampling (Yves Tillé 2016) e Icarus (Antoine Rebecq, 2017) del software R, estudiando las posibles diferencias según los distintos métodos de ajuste (lineal, raking ratio, logit y lineal truncado). Los resultados confirmaban un ajuste idéntico al 5º decimal con los tres paquetes, a excepción del método lineal truncado con el paquete Icarus, que aún no está implementado en la última versión disponible de este paquete (0.3.0 del 04/03/2017).

Los cuatro métodos de ajuste (M1=lineal, M2=raking ratio, M3=logit y M4=lineal truncado) son asintóticamente equivalentes. En el gráfico se puede comprobar cómo la distribución de pesos finales para los 4 métodos, calculados mediante el paquete Sampling para la Encuesta de Movilidad Social, son prácticamente iguales:



Siguiendo recomendaciones del Instituto Nacional de Estadística (INE), al igual que se hizo en anteriores ediciones de la Encuesta Social, se optó por el método Lineal truncado con límites 0,1 y 10. La no disponibilidad actual de este método con el paquete Icarus, nos ha llevado a utilizar el paquete Sampling.

8. Imputación basada en modelos

En los procesos de imputación se pretende tratar diversos tipos de omisión de información:

1. Errores de cobertura y de selección de la muestra: dichos errores se presentan, por ejemplo, cuando hay unidades de la población objetivo que no están representadas en el marco de muestreo, cuando las probabilidades de selección de las unidades están distorsionadas, o cuando se producen otros errores de selección de la muestra.

2. Falta de respuesta de las unidades de muestreo: hace referencia a la ausencia de información en unidades completas (hogares y/o personas) seleccionadas para la muestra.

3. Falta de respuesta en determinadas características: hace referencia a la situación en la que una unidad de la muestra se ha registrado con éxito, pero no se ha obtenido toda la información requerida.

En la Encuesta Social 2017: Movilidad social en Andalucía, el tipo de omisión que se ha tratado en los procesos de imputación ha sido la omisión parcial, es decir, la falta de respuesta en determinadas características, si bien la incidencia de este tipo de omisión ha sido muy reducida como muestra el cuadro adjunto. Esta falta de información afecta a diversas características, alguna de ellas fundamental en el proceso de clasificación social del individuo por lo que se optó por la imputación en estos casos.

	Clase Social Entrevistado	Clase Social Padre	Clase Social Madre
Información completa	2.847 (94,9%)	2.633 (87,8%)	2.162 (49,6%)
No aplica calcular la clase social debido a inactividad d ela persona de referencia	153 (5,1%)	274 (9,1%)	2.180 (50,0%)
Sin información suficiente para imputar (ni se dispone de la ocupación ni de la formación)	-	30 (1,0%)	7 (0,2%)
Registros susceptibles de imputación:	**-**	**63 (2,1%)**	**11 (0,3%)**
1. Falta el nº de asalariados en caso de ser empresario	-	22 (0,7%)	4 (0,1%)
2. Falta el sector de actividad en caso de ser empresario	-	5 (0,2%)	1 (0,0%)
3. Falta la ocupación	-	37 (1,2%)	6 (0,1%)

En la estadística oficial[5] se distingue entre dos grupos de métodos aplicables en el proceso de imputación

- Imputación, en referencia a la generación de la información omitida a partir de relaciones estadísticas internas del conjunto de datos.

[5] REGLAMENTO (CE) No 1981/2003 DE LA COMISIÓN de 21 de octubre de 2003 por el que se aplica el Reglamento (CE) no 1177/2003 del Parlamento Europeo y del Consejo relativo a las estadísticas comunitarias sobre la renta y las condiciones de vida (EU-SILC) en lo que respecta a las características del trabajo de campo y los procedimientos de imputación

- Modelización, utilizando relaciones fundamentales e información externa al conjunto de datos.

El método de imputación habitualmente empleado en ediciones anteriores se podría englobar en la primera categoría indicada, tomando para la generación de la información omitida aquel registro completo que en términos estadísticos fuera más próximo. Es decir se asumía cierta relación interna entre el conjunto de datos de cada registro.

En la última edición de la Encuesta Social, en concreto en la Encuesta de Movilidad Social de Andalucía, la Unidad Central de Encuesta ha implementado un método de imputación basado en modelos, siguiendo las recomendaciones de Eurostat en otras Encuestas Sociales, de manera que el procedimiento aplicado a los datos preserve la variabilidad de las variables y la correlación entre ellas. Los métodos que incluyan un «componente de error» en los valores imputados son preferibles a los que imputan simplemente un valor determinado. Los métodos que tengan en cuenta la estructura de las correlaciones (u otras características de la distribución conjunta de las variables) son preferibles al enfoque marginal o univariante. En este sentido se han probado distintos paquetes disponibles en R, entre otros mice (Stef van Buuren 2018), en línea con softwares como IVEware empleado por el INE[6], para imputación basada en modelos multivariantes con modelos de regresión secuenciales.

Finalmente, debido a la sencillez de implementación y mejor interpretación de los resultados, optamos por el paquete missForest (Daniel J. Stekhoven 2013). Dicho paquete (Non parametric Missing Value imputation using Random Forest) permite imputar valores perdidos para el caso de datos de tipo mixto, es decir imputa datos continuos y/o categóricos utilizando árboles aleatorios. Además proporciona una estimación de error de imputación fuera de bolsa (out of bag (OOB)).
A continuación se muestra un resumen de los datos imputados y los resultados del ajuste:

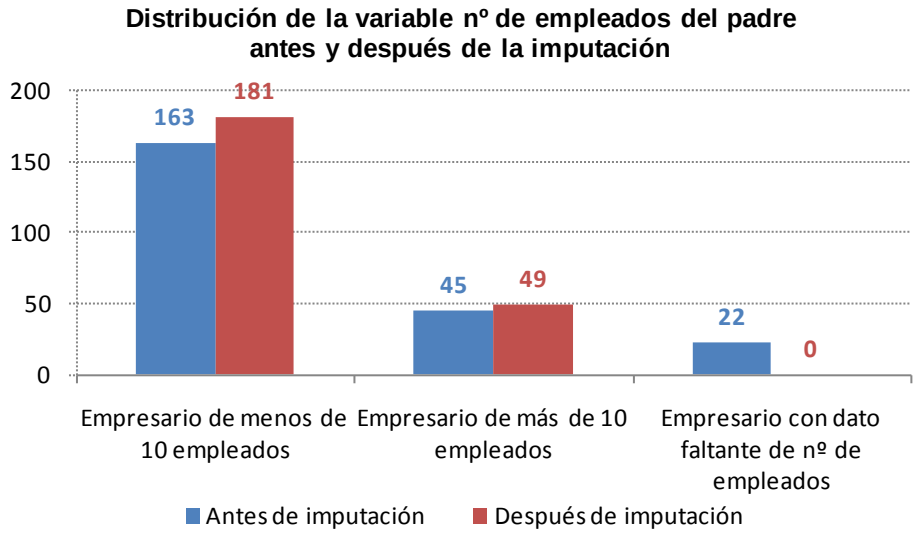
1. Nº de asalariados en caso de ser empresario

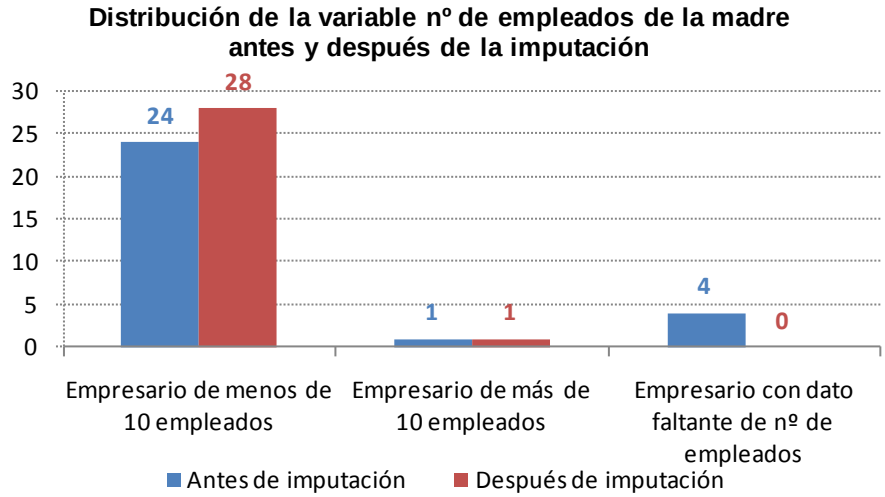
   Las variables empleadas para realizar la imputación han sido las siguientes:
   - Provincia y tamaño de municipio de residencia del entrevistado/ a
   - Formación del padre/ madre
   - Situación laboral del padre/ madre
   - Ocupación del padre/ madre
   - Sector de actividad del padre/ madre
   - Pregunta del cuestionario "¿Cómo era la situación económica del hogar en el que vivía en aquel momento?"
   - Año de nacimiento del padre/ madre

   Los errores OOB han sido del 1,8% en el caso del nº de empleados del padre y del 0,03% para el nº de empleados de la madre.

[6] Estimation and Imputation. SILC 2014. EU-SILC 2014 Quality Report

La distribución de las variables antes y después de la imputación es la siguiente:





2. Sector de actividad en caso de ser empresario
   Para construir la clase social es necesario conocer si el empresario trabaja en el sector de la agricultura o no. En lo que se refiere a padres empresarios, se desconoce el sector de actividad de 4 padres y de 1 madre. Parece más plausible que si el entrevistado recuerda que su padre/ madre era empresario/ a pero no recuerda el sector de actividad, dicho sector de actividad no sea la agricultura. Por este motivo y dado el escaso número de casos, se ha decidido imputarlos con el sector de "no agricultura" y no aplicar un modelo estadístico.

3. Ocupación
   Para la imputación de esta variable se han probado árboles aleatorios con las siguientes variables:
   - Provincia y tamaño de municipio de residencia del entrevistado/ a
   - Formación del padre/ madre
   - Situación laboral del padre/ madre
   - Sector de actividad del padre/ madre

- Pregunta del cuestionario “¿Cómo era la situación económica del hogar en el que vivía en aquel momento?”
- Año de nacimiento del padre/ madre

Sin embargo dado al escaso número de casos y a que los errores OOB obtenidos han sido superiores al 35%, finalmente se optó por no imputar dicha variable.

Posibles replanteamientos metodológicos futuros

En las próximas Encuestas Sociales, tomando ejemplos de otros institutos de estadística y siguiendo recomendaciones metodológicas, estamos valorando incorporar mejoras y modificaciones en el diseño muestral, en concreto:

- Eliminar la muestra reserva: Las recomendaciones en el ámbito de la estadística pública en cuanto al empleo de muestra reserva para sustituir unidades muestrales no encuestadas señalan su preferencia hacia una ampliación del tamaño de muestra teórico para la consecución de una muestra efectiva aceptable, en detrimento del uso de una muestra de unidades de reserva. El INE ha ido integrando sucesivamente esta visión en el diseño de sus encuestas a hogares, siguiendo las recomendaciones de Eurostat. La experiencia del IECA en encuestas telefónicas hasta el momento ha partido de un diseño muestral con unidades principales y unidades reserva, de cara a las próximas Encuestas Sociales la unidad de encuesta está valorando la estrategia de prescindir de la utilización de una muestra de viviendas reserva en cada sección, destinada a realizar sustituciones en el caso de incidencias en la muestra titular. Esta metodología implicaría incrementar el tamaño de muestra teórica para garantizar un tamaño de muestra efectiva adecuado para el propósito de la encuesta.
- Simplificar el diseño muestral, eliminando etapas, en concreto la selección de secciones censales. Actualmente el IECA en las Encuestas Sociales implementa un diseño muestral trietápico, tomando las secciones censales como conglomerados de primera etapa. Este diseño responde a un tipo de encuesta a hogares presencial donde el coste de desplazamiento es un factor importante en el diseño, mientras que la Encuestas Sociales del IECA han sido en estas últimas ediciones bicanal (CATI-CAWI).
- Incorporar en el diseño unidades de selección (celdas) o estratos (grados de urbanización) nuevos a partir de la información del marco de población y viviendas georreferenciado.

Bibliografía

- Azor G., Jiménez J., Pérez C. and Porras J. (2011): Study of variance methods in the Spanish Labour Force Survey (EPA). Working Papers 03/2011. Instituto Nacional de Estadística (INE)
- Booth J. G., Butler R. W. and Hall P. (1994), 'Bootstrap Methods for Finite Populations', Journal of the American Statistical Association 89, pp. 1282–1289
- Díaz de Rada, V. (2010). Comparación entre los resultados proporcionados por encuestas telefónicas y personales: el caso de un estudio electoral. Colección Opiniones y Actitudes, número 66. CIS. Madrid.
- Dunn, H.L. (1946). Record Linkage. American Journal of Public Health, 36 (12), 1412–1416.
- Kuijvenhoven, L and Scholtus, S. (2011): Bootstrapping Combined Estimator based on Register and Sample Survey Data. Discussion paper 201123. Statistics Netherlands.
- Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992). Some recent work on resampling methods for complex surveys. Survey Methodology, 18,209-217.
- Winkler, W.E. (2006). Overview of Record Linkage and Current Research Directions. Bureau of the Census. Research Report Series.
- Yeo D., Mantel H., Liu TP. (1999): Bootstrap variance estimation for the National Population Health Survey. In Proceedings of the Survey Research Methods Section. Baltimore, MD, American Statistical Association, 1999, p. 778–783
- A Point-based Foundation for Statistics – Final report from the GEOSTAT 2 project, https://www.efgs.info/wp-content/uploads/2017/03/GEOSTAT2ReportMain.pdf
- Package 'sampling' (Yves Tillé, 2016) (cran.r-project.org)
- Package 'icarus' (Antoine Rebecq, 2017) (cran.r-project.org)
- Package 'mice' (Stef van Buuren 2018) (cran.r-project.org)
- Package 'missForest' (Daniel J. Stekhoven 2013) (cran.r-project.org)