¿Es posible construir una base de datos municipal completa y consistente a partir del Censo de 2011?

Francisco J. Goerlich Gisbert

Ivie

¿Es posible construir una base de datos municipal completa y consistente a partir del Censo de 2011?

Francisco J. Goerlich Gisbert

Universitat de València e Instituto Valenciano de Investigaciones Económicas (Ivie)

Versión (R0): Septiembre/2016

DOI: http://dx.medra.org/10.12842/MUNICIPIOS_CENSO_2011

Correspondencia y solicitud de información:

Francisco J. Goerlich Gisbert, Universidad de Valencia, Departamento de Análisis Económico, Campus de Tarongers, Av de Tarongers s/n, 46022-Valencia. E-mail: Francisco.J.Goerlich@uv.es.

Agradecimientos: El autor agradece a Jorge Luis Vega Valle, Carmen Teijeiro Breijo, Antonio Argüeso Jimenez e Ignacio Duque Rodriguez de Arellano del Instituto Nacional de Estadística (INE) su disponibilidad a resolver innumerables cuestiones metodológicas en relación a la información censal. También el *feedback* recibido por diversos técnicos del Instituto Valenciano de Investigaciones Económicas (Ivie), en particular Irene Zaera y Carlos Albert, cuyos comentarios contribuyeron a iterar en el proceso de desarrollo de los algoritmos de desagregación que se mencionan en este trabajo.

El autor agradece la ayuda del programa de Investigación FBBVA-Ivie, así como la del proyecto ECO2015-70632-R.

Este trabajo es un resumen de un documento metodológico mucho más amplio donde se detallan los procedimientos seguidos para cada variable desagregada, y como se solucionaron las diversas inconsistencias encontradas.

La base de datos descrita en este trabajo puede solicitarse al autor: Francisco.J.Goerlich@uv.es.

¿Es posible construir una base de datos municipal completa y consistente a partir del Censo de 2011?

Francisco J. Goerlich Gisbert

RESUMEN

¡Pues casi! Este documento describe el proceso seguido para la elaboración de una base de datos municipal completa a partir de la información del Censo de 2011. Por completa se debe entender variables para los 8,116 municipios existentes en la fecha de referencia censal. Además dicha base de datos debe ser consistente, en el sentido de ser acorde con la información censal publicada: los microdatos y el sistema de *Tablas a Medida*. Se repasan algunas cuestiones metodológicas relacionadas con la elaboración del Censo de 2011, y sus efectos sobre la disponibilidad de información para pequeñas áreas.

Palabras Clave: Censo 2011; Municipios; Áreas pequeñas.

ABSTRACT

Almost! This document describes the process to obtain a complete municipal data base from the Census 2011 information. By complete, we mean variables for the full sample of 8,116 municipalities in the census reference date. In addition, the database should be consistent with the census public information released by the National Statistical Institute: microdata and *Customized Tables*. In the way, we review some methodological questions related to census elaboration, and their effects on the availability of small sample information.

Key Words: Census 2011; Municipalities; Small area.

1. Introducción

El Censo de 2011 ha supuesto una ruptura metodológica importante respecto a la tradición censal histórica española. Ha implicado el paso desde una metodología censal clásica, basada en un recorrido de campo exhaustivo, hacia un sistema mixto, en el que el recuento de la población, así como sus características demográficas más básicas derivan de un registro administrativo –el Padrón–, mientras que el resto de características de la población proceden de una encuesta a gran escala de alrededor del 10% de la población (INE 2011).

Por el contrario, el Censo de 2011 si preveía un recorrido exhaustivo de campo en lo que hace referencia a los edificios en los que hay alguna vivienda principal, al objeto de completar su geo-referenciación, e ir introduciendo de esta forma los Sistemas de Información Geográfica en la información censal (INE 2011).

Sin duda alguna esta ruptura no tiene vuelta atrás, y en este sentido el Censo de 2011 ha sido un censo de transición. El Censo de 2021 está previsto que se sea un censo basado totalmente en registros administrativos (CCHS 2015), de forma que este apoyo de la estadística oficial en general, y de los censos en particular, en los registros administrativos y el llamado *big-data* continuará en el futuro.

El cambio metodológico es consecuencia natural de los grandes avances en las tecnologías de la información y las comunicaciones, que permiten el cruce masivo de información entre las diferentes administraciones públicas. Es un cambio metodológico que, en última instancia, representa un cambio de paradigma importante, y que tiene *pros* y *cons* que conviene tener presentes al analizar la información censal finalmente difundida por el INE (2013a). Aunque el cambio metodológico no implica, necesariamente, una pérdida en la calidad de la información ofrecida (Goerlich, Ruiz, Chorén y Albert 2015), esta afirmación deber ser cualificada en varias direcciones, no solo a la luz de la información finalmente puesta a disposición del público por parte del INE en sus diferentes productos censales: Personas, Hogares, Viviendas y Edificios, sino también en relación al ámbito territorial al que hagamos referencia: Conjunto Nacional, Comunidades Autónomas, Provincias, Municipios o Secciones Censales.

La experiencia en el manejo de la información censal difundida por el INE (2013a) (Goerlich, Ruiz, Chorén y Albert 2016; Reig, Goerlich y Cantarino 2016) indica que el Censo de 2011 ofrece una escasa información para ámbitos territoriales reducidos, incluso municipios, que constituyen la unidad administrativa básica de división del territorio, y para los que los censos constituían la única posibilidad de disponer de información homogénea y comparable más allá de la información puramente demográfica.

La situación es mucho más dramática en lo que hace referencia a las estadísticas infra-municipales. La información a nivel de sección censal es prácticamente inexistente con generalidad, ya que para la totalidad de las secciones censales solo se dispone de una única variable: la población total, restringida a la población residente en viviendas principales, y redondeada al múltiplo de 5 más cercano. Por otra parte, el Censo de 2011 ha sido el primero de la historia moderna en la que dicho censo no lleva asociado un nomenclátor de entidades poblacionales, que sin embargo si está disponible en relación al Padrón de 2012, cuya fecha de referencia es solo 2 meses posterior al del Censo de 2011: 1 de enero de 2012, para el Padrón, frente a 1 de noviembre de 2011, para el censo. Lo mismo se aplica a determinadas características demográficas básicas de la población, que aunque sí están disponibles en el Padrón de 2012, no lo están en el Censo de 2011.

La misma experiencia indica que la calidad de la información no es uniforme para los diferentes ámbitos censales, y que a pesar de que –al menos en teoría– el recorrido de campo ha sido exhaustivo en lo que hace referencia a los edificios con alguna vivienda principal, la información finalmente disponible para el usuario final no ha sido todo lo rica que cabía esperar inicialmente a partir del contenido del proyecto censal (INE 2011).

Este trabajo examina métodos sencillos que permitan obtener estimaciones para la gran mayoría de variables censales, y el conjunto completo de los 8,116 municipios del Censo de 2011. Dichas estimaciones deberán ser consistentes con los microdatos de la encuesta de dicho censo. Es decir, el marco de referencia para la obtención de variables a nivel municipal son los microdatos, que es la fuente de información para la totalidad de las características no-demográficas de la población. Esta puntualización es importante por las razones que se expondrán más adelante. El objetivo último es disponer de una base de datos municipal completa para un conjunto amplio de variables.[1]

La estructura del trabajo es la siguiente. A continuación se exponen, muy brevemente, los elementos básicos de la metodología censal. Ello es necesario, en parte, para entender el esquema seguido en la desagregación de la información a nivel municipal, que se describe en el apartado 3. No todos los tipos de variables pueden encajar en el mismo esquema de desagregación, por lo que se describirán los métodos finalmente seguidos en el apartado 4. Un último apartado ofrece unas breves conclusiones finales. El listado de variables desagregadas aparece en el apéndice.

[1] Un trabajo paralelo aplica técnicas estadísticas de estimación en pequeñas áreas (Elbers, Lanjouw y Lanjouw 2003) para añadir información sobre variables monetarias al Censo de 2011. Los métodos de este trabajo son más sencillos y no implican modelización estadística.

2. Estructura de la información en el Censo 2011

2.1. Información sobre Personas y Hogares

Desde el punto de vista de la información sobre personas y sus características el censo de 2011 se apoya en dos pilares fundamentales: El Padrón, para la información puramente demográfica, y una gran encuesta, diseñada en principio para ser representativa a nivel de municipio, para el resto de características de la población (INE 2011). Estos dos pilares proporcionan información diferente y es necesario conocer sus interrelaciones para entender ciertos detalles del proceso de desagregación a escala municipal.

En primer lugar parece natural preguntarse por qué las características demográficas básicas de la población se basan en el Padrón y no son **exactamente** las que proceden del Padrón para la fecha de referencia del censo, 1 de noviembre de 2011. La razón deriva de la propia naturaleza del Padrón como registro administrativo con regulación legal, lo que hace que cualquier modificación del mismo debe tener un respaldo normativo, es decir, no pueden hacerse ajustes estadísticos. Desde el establecimiento del Padrón continuo en 1998[2] las cifras de población derivadas del Padrón se disociaron de las cifras de población censales, de forma que la población del Censo de 2001 no coincidió con la población que se derivaba del Padrón en la fecha de referencia más cercana.[3] Es conocido que la propia gestión del Padrón produce en el mismo una sobre-estimación de la población, fundamentalmente ligada al registro de los extranjeros, aunque también se detectan problemas en los extremos de la distribución por edades (Goerlich 2007, 2012).

El resultado es que para conocer la 'cifra de población' de España y sus territorios, como mejor estimación estadística de la población residente, el Padrón tenía que ser ajustado para hacerlo más fiel a la realidad. Por tanto, a partir del Padrón el INE construyó, para la fecha de referencia censal, un Fichero Precensal (FPC) que ajustaba convenientemente las altas y bajas del Movimiento Natural de la Población (MNP), y en el que cada registro disponía de una factor de recuento que era igual a 1 si era posible comprobar que esa persona era efectivamente residente en España, mediante el cruce con diversos registros administrativos, por ejemplo los de Seguridad Social, o era desconocido si no era posible comprobar con certidumbre que esa persona era residente en nuestro país. A estos registros se les denomina dudosos. El 97.2% de los registros del FPC obtuvieron un factor de recuento igual a 1. Este FPC constituye el primer pilar fundamental de la información censal referida a personas y hogares. La

[2] La última revisión padronal fue la de 1996.

[3] Esta disociación de fuentes de información demográfica hizo que el INE (*on line*), tuviera que explicar en su *web* los tipos de cifras de población que ofrece al público.

población de referencia del FPC es la **población residente** en España, ya sea en viviendas principales o colectivas.

Paralelamente se realiza una gran encuesta por muestreo con dos objetivos: (*i*) determinar el factor de recuento en los registros dudosos del FPC, y (*ii*) estimar las características de la población. Se fijó una fracción de muestreo teórica del 12.3%, aunque finalmente la muestra recogida fue algo menor, alrededor del 9%. Se seleccionó muestra en todas las secciones censales, unidad primaria de muestreo, siendo la unidad secundaria las viviendas principales. Esta fracción de muestreo aumentaba conforme disminuía el tamaño del municipio, de forma que el muestreo es exhaustivo en los municipios inferiores a los 200 habitantes.[4] La muestra debía ser representativa a nivel municipal. La **población de referencia de la muestra** es la **población residente en viviendas principales**. En consecuencia, de la muestra queda excluida la población en colectivos, que fue objeto de una operación estadística independiente: *Encuesta de Colectivos del Censo de Población y Viviendas 2011* (INE 2013b).[5]

2.2. El encaje entre la muestra y la información del Fichero Precensal

El FPC y la muestra son operaciones independientes que deben ser reconciliadas. Este proceso de reconciliación, llevado a cabo por el INE, descansa en dos actuaciones no totalmente independientes.

Primero, para determinar el factor de recuento de los registros dudosos se efectuó (*i*) una partición en clases a partir de características observables en ambos conjuntos de información –edad, nacionalidad y lugar de residencia–, y (*ii*) un cruce nominal entre la muestra recogida en campo y el FPC, de forma que se pudieran enlazar los registros, y así poder identificar los que, si bien figuran en el FPC como dudosos, fueron encontrados en la realidad al ser recogidos en la muestra.

A partir de esta identificación, por el principio de analogía a nivel de clase, se estimaron los factores de recuento de los registros dudosos. Estos factores de recuento son números reales, inferiores a la unidad en su mayoría, y con un valor medio de 0.424. Este procedimiento de asignación de factores de recuento a registros dudosos

[4] Se podría decir, en cierta forma, que para estos municipios sí se realizó un censo exhaustivo, y en consecuencia no deberían plantarse problemas de representatividad para estos municipios, salvo en el caso de que la falta de respuesta fuera importante.

[5] También se realizó, como módulo del censo, una encuesta específica sobre las personas sin hogar (EPSH 2012), derivada de la reglamentación desarrollada por la Unión Europea en relación al Censo de 2011 (Reglamento 1201/2009 de la Comisión Europea). Los resultados de esta encuesta no son difundidos por el INE en la sección del censo 2011, sino en una sección aparte. En principio estas personas deberían estar en el FPC, puesto que deberían estar empadronadas en algún lugar aunque no tengan domicilio (https://www.boe.es/buscar/doc.php?id=BOE-A-2015-3109).

es el responsable de que las cifras de población derivadas del censo sean reales, y no números naturales.[6]

El procedimiento detallado está descrito en INE (2012) y Goerlich, Ruiz, Chorén y Albert (2015, capítulo 1). Todo lo que nos interesa aquí es que, tras esta operación, cada registro del FPC dispone de un factor de recuento asignado. De esta forma disponemos de un **fichero censal final ponderado** que es el que nos determina la cifra de población censal y sus características demográficas básicas. La población residente derivada del censo por este procedimiento resultó ser de 46,815,916 personas.[7]

La información demográfica asociada al **fichero censal final ponderado** fue difundida por el INE en diciembre de 2012, y va referida al total de **población residente**.

Segundo, la muestra debe ser calibrada a la población, de forma que sea consistente con ella en diversas dimensiones, tanto referidas a características de la población como a ámbitos territoriales. Sin embargo la población de referencia de la encuesta no es la derivada del fichero censal final ponderado, sino la población en viviendas familiares principales, excluyendo la población en colectivos. Dicha población no se conoce a partir del FPC.

La población residente en establecimientos colectivos fue estimada por la *Encuesta de Colectivos* en 444,101 personas. Sin embargo, no toda la población residente en establecimientos colectivos se encuentra empadronada en ellos. De acuerdo con dicha encuesta, solo 241,187 personas residentes en establecimientos colectivos se encuentran empadronados en ellos, mientras que 202,914 personas residentes en establecimientos colectivos no se encuentran empadronados en ellos, sino en viviendas familiares principales, y se están contando en las viviendas familiares a efectos de la muestra, que es donde figuran empadronados. En consecuencia, la población residente en viviendas familiares principales es

$$46{,}815{,}916 - 241{,}187 = 46{,}574{,}729 \quad \text{personas}$$

Esta es la población a la que hay que calibrar la encuesta. Es decir los factores de elevación de la encuesta deben sumar esta población.[8]

---

[6] Por la misma razón, y dado que existe una tabla de poblaciones municipales por sexo y edades simples, es posible saber, a este nivel de desagregación, qué poblaciones tienen registros dudosos y cuáles no.

[7] En realidad 46,815,916.44.

[8] Estas cifras proceden de INE (2014). Sin embargo, la inspección de la suma de los factores de elevación de la muestra, procedente de los microdatos, mostraba una discrepancia de 4 personas.

A partir de la población en viviendas familiares principales y de la población empadronada en establecimientos colectivos se observó que **el origen de la discrepancia se debe a un descuadre en Barcelona (+2), Sevilla**

El proceso de calibración utiliza el método estándar del INE: CALMAR (Deville y Särndal 1992; Deville, Särndal y Sautory 1993), se efectúa a nivel de municipio, y es función del tamaño del mismo. La figura 1 muestra la información que se calibra para los diferentes tamaños de municipios (INE 2014).

La existencia de dos colectivos poblacionales de referencia, población residente y población residente en viviendas principales, complica notablemente el proceso de desagregación de los microdatos, puesto que las variables desagregadas deben ajustarse a marginales poblacionales que no pueden tomarse directamente del FPC. El FPC ofrece información para la población residente total, mientras que la base de datos municipal, construida a partir de los microdatos, debe ajustarse a la población residente en viviendas principales.

Por esta razón, como paso previo se estimó a nivel municipal, por los métodos expuestos en el apartado 3, la población residente en viviendas principales por sexos y en dos grupos de edad: los menores de 16 años, y las personas de 16 años y más.

Para efectuar esta estimación se tuvo en cuenta el hecho de que muchos municipios que no tienen establecimientos colectivos, y por tanto en estos la población total coincide con la residente en viviendas familiares principales. En concreto, solo en 2,621 municipios hay establecimientos colectivos; el resto, 5,495 municipios, no disponen de establecimientos colectivos, y por tanto en estos la población total coincide con la residente en viviendas familiares principales.[9]

(+1) y Zaragoza (+1). La razón de porque el INE pierde 4 personas en el proceso de calibrado es desconocida. La solución adoptada para que todo ajuste, sin alterar la cifra población total, ni tampoco la residente en viviendas familiares principales, que implicaría alterar los factores de elevación, fue simplemente incrementar la población empadronada en establecimientos colectivos en 2 personas para Barcelona, y en 1 persona para Sevilla y Zaragoza. La base de datos asociada a este trabajo incluye este ajuste, y por tanto discrepa de lo publicado por el INE para esta variable en las 4 personas indicadas.

Por tanto, tras el ajuste, los números anteriores quedan de la siguiente forma:

- **Población total**: 46,815,916.44
- **Población empadronada en establecimientos colectivos**: 241,190.87
- **Población residente en viviendas familiares principales**:

$$46{,}815{,}916.44 - 241{,}190.87 = 46{,}574{,}725.57$$

Que coincide con la suma de los factores de elevación de la muestra.

[9] Sin embargo, la inspección de las cifras de población indica que en 5,608 municipios no existe población empadronada en colectivos. Solo en los 2,508 municipios restantes existe población empadronada en colectivos. En consecuencia, hay 113 municipios que tienen establecimientos colectivos, pero no población empadronada en ellos.

Figura 1. Información sobre el calibrado de la muestra del Censo 2011

Población del municipio	Información que se calibra
Menos de 51 habitantes	Total de población
Entre 51 y 200 habitantes	Total de población Desglose por sexo
Entre 201 y 2.000 habitantes	Total de población Desglose por sexo y edades en grandes grupos
Entre 2.001 y 10.000 habitantes	Total de población Desglose por sexo y por edades en grupos decenales
Entre 10.001 y 50.000 habitantes	Total de población Desglose por sexo y por edades en grupos decenales Desglose por sexo y por nacionalidad (española/extranjera)
Entre 50.001 y 100.000 habitantes	Total de población Desglose por sexo y por edades en grupos quinquenales Desglose por sexo y por nacionalidad (española/extranjera)
Más de 100.00 habitantes	Total de población Desglose por sexo y por edades en grupos quinquenales Desglose por sexo y por nacionalidades más frecuentes

Fuente: INE (2014).

2.3. Estructura territorial en los microdatos del censo de 2011

A pesar de que la muestra debe tener representatividad municipal y de que la fracción de muestreo aumenta conforme disminuye el tamaño del municipio, lo que no debería plantear problemas de representatividad en los municipios más pequeños –el muestreo es exhaustivo para los municipios inferiores a los 200 habitantes– los microdatos del censo solo ofrecen información a nivel municipal para aquellos municipios de más de 20,000 habitantes, donde la fracción teórica de muestreo es del 9%. El resto de municipios se encuentran agrupados en 4 estratos de tamaño dentro de cada provincia. Los estratos son los siguientes:

- Hasta los 2,000 habitantes (Código 991).
- Entre 2,001 y los 5,000 habitantes (Código 992).
- Entre los 5,001 y los 10,000 habitantes (Código 993).
- Entre los 10,001 y los 20,000 habitantes (Código 994).

La distribución de los municipios por provincia y tramo se ofrece en la tabla 1.[10]

Tabla 1. Geografía por tramos de municipios en los microdatos del censo 2011.

Municipios por Provincia y Tramo						
cp	Provincia	Hasta 2.000 hab.	De 2.001 a 5.000 hab.	De 5.001 a 10.000 hab.	De 10.001 a 20.000 hab.	Más de 20.000 hab.	Total de Municipios
01	Álava	42	6		2	1	51
02	Albacete	62	17	2	2	4	87
03	Alacant/Alicante	66	18	20	13	24	141
04	Almería	62	19	9	6	6	102
05	Ávila	233	10	4		1	248
06	Badajoz	97	41	17	4	5	164
07	Illes Balears	14	13	17	11	12	67
08	Barcelona	121	58	51	37	44	311
09	Burgos	360	6	2		3	371
10	Cáceres	188	21	7	3	2	221
11	Cádiz	6	6	10	7	15	44
12	Castellón/Castelló	104	11	9	3	8	135
13	Ciudad Real	62	16	11	8	5	102
14	Córdoba	23	24	14	6	8	75
15	A Coruña	12	29	31	11	11	94
16	Cuenca	222	9	5	1	1	238
17	Girona	159	29	14	11	8	221
18	Granada	95	34	18	14	7	168
19	Guadalajara	267	13	4	2	2	288
20	Guipúzcoa	45	10	13	14	6	88
21	Huelva	35	24	7	7	6	79
22	Huesca	189	6	1	5	1	202
23	Jaén	33	36	13	9	6	97
24	León	178	21	5	4	3	211
25	Lleida	193	23	10	4	1	231
26	La Rioja	153	12	5	2	2	174
27	Lugo	24	30	8	4	1	67
28	Madrid	69	31	31	15	33	179
29	Málaga	44	29	9	3	16	101
30	Murcia	5	4	6	13	17	45

[10] Puesto que la población del Censo 2011 es un número real, la clasificación de los municipios por tamaño no es una cuestión trivial. Obsérvese que los criterios de clasificación por tamaños, tomados directamente del INE, están pensando en una población entera, lo que no es el caso para el censo de 2011, y no queda claro que es lo que sucede con los municipios que están en el límite de los intervalos.

Por ejemplo, un municipio para el que la población sea 2,000.8 habitantes, ¿deber ser clasificado en el estrato 1, código 991, o en el estrato 2, código 992?

La regla exacta aplicada por el INE para la clasificación por tamaños no se menciona en ningún sitio, pero algo de experimentación mostró que la clasificación por tamaños se efectúa partir de las poblaciones municipales en enteros, obtenidas mediante el redondeo estándar. De acuerdo con esta regla, el municipio anterior sería clasificado en el estrato 2, puesto que su población 'entera' serían 2,001 habitantes. Lo contrario sucedería si la población del municipio fuera de 2,000.3 habitantes.

Municipios por Provincia y Tramo							
cp	**Provincia**	**Hasta 2.000 hab.**	**De 2.001 a 5.000 hab.**	**De 5.001 a 10.000 hab.**	**De 10.001 a 20.000 hab.**	**Más de 20.000 hab.**	**Total de Municipios**
31	Navarra	213	37	12	7	3	272
32	Ourense	61	21	4	5	1	92
33	Asturias	36	11	10	14	7	78
34	Palencia	180	6	4		1	191
35	Palmas de Gran Canaria (Las)	2	2	8	9	13	34
36	Pontevedra	4	21	12	16	9	62
37	Salamanca	349	3	6	3	1	362
38	Santa Cruz de Tenerife	6	16	12	8	12	54
39	Cantabria	55	27	9	6	5	102
40	Segovia	198	7	3		1	209
41	Sevilla	14	25	30	19	17	105
42	Soria	175	5	2		1	183
43	Tarragona	122	32	14	6	10	184
44	Teruel	225	8	1	1	1	236
45	Toledo	112	63	15	11	3	204
46	Valencia/València	132	55	28	20	31	266
47	Valladolid	201	13	7	1	3	225
48	Vizcaya	60	19	13	9	11	112
49	Zamora	244	1	1	1	1	248
50	Zaragoza	256	22	9	4	2	293
51	Ceuta					1	1
52	Melilla					1	1
	España	**5,808**	**1,000**	**553**	**361**	**394**	**8,116**

Fuente: INE (2013a).

Los 394 municipios de más de 20,000 habitantes son perfectamente identificables en los microdatos. Además de estos, cuando el municipio no supera los 20,000 habitantes, también es posible identificar los casos en los que hay un solo municipio por estrato: 8 casos. Todos ellos, 402, aparecen con fondo amarillo en la tabla 1. Finalmente también es posible identificar en los microdatos el municipio más pequeño de España, desde el punto de vista demográfico: *Illán de Vacas*, en la provincia de Toledo, 45080, 1 habitante. Basta con identificar el registro de la provincia 45, Toledo, del estrato 1, 991, con factor de elevación igual a 1.[11] Esta identificación exacta ya no es posible para el municipio siguiente en tamaño de población: *Jaramillo Quemado*, en la provincia de Burgos, 09184, con 5 habitantes.[12]

[11] Y comprobar que la información demográfica de dicho registro coincide con la derivada del FPC: un varón de 59 años, nacido en España y de nacionalidad española.

[12] En los microdatos del censo existen 47 registros con factor de elevación igual a 1, pero ninguno de ellos en la provincia de Burgos.

En consecuencia podemos identificar directamente en los microdatos 403 municipios; para el resto, 7,713 municipios, solo podemos conocer los valores agregados del estrato al que pertenecen. El objetivo de este trabajo es obtener información, sobre determinadas variables disponibles en los microdatos, para dichos municipios.

Además del fichero de *Personas y Hogares*, con 4,107,465 registros de personas y 1,621,643 registros de hogares, al que hace referencia este trabajo, el INE también ha hecho público un fichero de microdatos de *Viviendas y Edificios*. Dicho fichero contiene 2,326,247 registros correspondientes a viviendas, tanto principales como no principales, ya sean secundarias o vacías. Este fichero no ha sido objeto de atención en este trabajo, aunque sus variables podrían ser desagregadas por métodos similares a los expuestos en el apartado siguiente.

Debe observarse que a efectos censales los **hogares** están constituidos por las **viviendas familiares principales**. [13]

2.5. El sistema de Tablas a Medida en la difusión censal

Además del fichero de microdatos, los productos de difusión del Censo 2011 incluyen un sistema de consultas de *Tablas a Medida* en la que el usuario tiene la posibilidad de seleccionar, dentro de un ámbito geográfico y un dominio, las variables que le interesen.

Es necesario remarcar que el sistema de *Tablas a Medida* está construido sobre la muestra, y en consecuencia la **población de referencia** es la **residente en viviendas principales**, en este sentido es **consistente con los microdatos**. El sistema es, sin embargo, bastante limitado[14] para obtener información completa y con generalidad para la totalidad de los municipios por las siguientes razones:

1. El sistema aplica una serie de **reglas de confidencialidad** (INE 2014), de forma que para mostrar la información de un ámbito geográfico dado se debe verificar que: (*i*) o bien el cociente entre el número de unidades muestrales existentes (sin elevar) para ese ámbito geográfico y el número de celdas[15] implicadas en la consulta sea mayor o igual a 5, o (*ii*) bien que el número de unidades muestrales de una celda individual en la consulta sea igual o superior a 5 unidades muestrales.

[13] El censo define como "**Hogar**: *Grupo de personas residentes en la misma vivienda*." (INE 2011, p.-15).

[14] Ello a pesar de que en abril de 2015 se modificó el sistema de consultas y la regla de confidencialidad para mostrar más información, haciendo el sistema menos restrictivo, de forma que la unidad territorial más pequeña ya no condiciona que se muestren los datos.

[15] Corregidas por un factor de sensibilidad en función de las variables implicadas en la consulta.

Estas dos condiciones dan lugar a que, ante una consulta determinada, sea posible obtener dos tablas separadas con niveles de información diferente. Por ejemplo, supongamos que se consultan una serie de municipios. Pueden darse tres casos: (*a*) si para un municipio dado se cumple (*i*) entonces el municipio es incluido en el listado para los que se muestra información completa, es decir para todas las celdas implicadas en la consulta; (*b*) si dicho municipio no cumple (*i*), pero algunas de las celdas implicadas cumplen (*ii*), entonces se genera un listado, diferente del anterior, en el que se ofrece información sólo para las celdas que cumplen dicho criterio, el resto se omite; y (*c*) si el municipio no cumple ni (*i*) ni (*ii*) simplemente es omitido y no se ofrece ninguna información.

2. Para aquellas celdas incluidas en el caso (*i*) anterior, es decir cuando se muestra la información completa del municipio en cuestión, si dichas celdas contienen menos de 5 unidades muestrales (sin elevar) entonces aparecen marcadas con un asterisco, '*', por considerarse que pueden tener errores elevados de muestreo.

   Por tanto, la diferencia entre los valores de dos celdas: 25 y 25*, es que en el segundo caso la estimación resulta de un número de unidades muestrales inferior a 5.

3. Para garantizar el secreto estadístico todos los datos se redondean al entero múltiplo de 5 más cercano.

   En consecuencia un valor de 0 no se sabe con seguridad si se debe al redondeo, una estimación inferior a 2.5, o simplemente a que el valor para esa celda en cuestión es realmente 0. Es de suponer que cuando no hay unidades muestrales para una celda esta se indica como 0,[16] mientras que cuando si las hay, pero estas son inferiores a 5 y la estimación resultante es inferior a 2.5, entonces se indica como 0*.[17] Bajo esta interpretación la distinción entre 0 y 0*, es simplemente que, en el primer caso esa característica no se da en el ámbito territorial considerado, mientras que en el segundo es un efecto del redondeo.

   Desde el punto de vista práctico, de la incorporación de la información del sistema de *Tablas a Medida* en las estimaciones, acabó siendo necesario distinguir entre 0 y 0*, ya que de otra forma, es decir asimilar 0* a 0 en todos los casos, produjo situaciones de inconsistencia entre las celdas de un municipio para determinadas variables y el marginal al que debía ajustarse. La razón es

[16] De otra forma, si la ausencia de unidades muestrales se asimila a menos de 5 unidades muestrales, todos los 0´s deberían aparecer como 0*´s.

[17] Podría darse el extraño caso en el que con 5 o más unidades muestrales la estimación resultante fuera inferior a 2.5, ya que existen en los microdatos factores de elevación muy próximos a 0, pero este debe ser un caso extremadamente poco probable.

que el ajuste RAS, que se utiliza en el proceso de estimación tal y como se explica más adelante, preserva los 0´s; y había situaciones en las que asimilar 0* a 0 generaba 0´s en todos los casos posibles, mientras que el marginal daba una cifra positiva.[18]

Por esta razón, en la información procedente de *Tablas a Medida* se asimiló, de forma un tanto arbitraria, 0* a 2, lo que permitió que el proceso de ajuste RAS no asignara un 0 a esas celdas.

4. No se dispone de una *API* (*Application Program Interface*) con el que efectuar descargas, siendo el procedimiento de selección y descarga de los diversos ámbitos y variables totalmente manual.

   A pesar de todas estas limitaciones, algo de experimentación mostró que la incorporación de la información de las *Tablas a Medida* mejoraba notablemente las estimaciones municipales por el procedimiento descrito a continuación. Por esta razón el sistema de consultas de *Tablas a Medida* se incorporó extensivamente en la medida de lo posible. Ello requirió un notable esfuerzo de descarga y organización de la información.

En función de la variable en cuestión, y su nivel de detalle, la información que es posible conseguir para los municipios es muy variable, aunque en ningún caso cubre el conjunto completo de municipios. Adicionalmente, para el caso de variables de personas, los ámbitos demográficos del sistema de *Tablas a Medida* son 3: (*i*) *residentes en viviendas principales*, (*ii*) *ocupados de 16 o más años*, y (*iii*) *cursan algún tipo de estudios y no trabajan*. En ocasiones ello genera algún problema de encaje con los microdatos, donde, por ejemplo, no están definidos los estudiantes como tal, excepto para los menores de 16 años.

[18] Un ejemplo puede ser ilustrativo de los problemas encontrados al asimilar 0* a 0. En el caso de la variable ESCOLAR, definida para los menores de 16 años, el FPC nos dice que el municipio Castro de Filabres (04033) no tiene población empadronada en colectivos y la población menor de 16 años son 6 personas. Sin embargo, las *Tablas a Medida* nos indican que la población que no acude a un centro escolar, ESCOLAR = 2, es 0, mientras que la que si acude a un centros escolar, ESCOLAR = 1, es 0*. Si asimilamos 0* a 0, opción con la que se experimentó inicialmente para no favorecer estimaciones muy pequeñas, entonces no hay forma de reconciliar estas dos cifras. El algoritmo de estimación hace que esas 6 personas, que no pueden ser asignadas a Castro de Filabres (04033), acaben siendo asignadas a otros municipios del mismo estrato.

3. De los micro-datos a los municipios

Este apartado describe los diversos métodos utilizados en la desagregación de los microdatos para el conjunto completo de los municipios del censo 2011. Cualquiera que sea el método seguido, este debe cumplir una **propiedad básica**:

1. Las estimaciones deben ser **consistentes con los microdatos**. En consecuencia la **población de referencia** es la **residente en viviendas principales**.

Esta propiedad básica tiene una serie de implicaciones:

2. Para los 403 municipios que pueden ser identificados en los microdatos, las estimaciones se toman directamente de los mismos.

3. La consistencia con los microdatos implica que: (*i*) para cada municipio, los valores desagregados deben coincidir con la variable de referencia a nivel municipal: población en residente en viviendas principales, o cualquier otra a la que deba agregar según los microdatos; y (*ii*) para cada estrato, los valores desagregados a nivel de municipio deben coincidir con los que se obtiene de los microdatos para ese estrato. La información para (*i*) debe ser conocida de forma externa a los microdatos, y la (*ii*) de los propios microdatos.

4. Si el método de desagregación inicial no verifica por construcción la consistencia mencionada en el punto anterior, lo que suele ser la regla y no la excepción, esta se alcanza mediante **ajuste proporcional iterativo** (*ipf* – Deming y Stephan 1940; Stephan 1942), conocido más popularmente en economía como el método RAS (Bacharach 1965).

5. **Validación**: Los métodos expuestos también pueden aplicarse a los municipios que se identifican claramente en los microdatos. En consecuencia, estos municipios, cuando forman parte de un estrato con más de un municipio, 379 casos, constituyen el conjunto de validación. De esta forma podemos tener una idea del error agregado de estimación. Si bien solo para el caso de municipios de más de 20.000 habitantes, lo que, sin duda, constituye un conjunto de validación sesgado.

6. **Medición del error para el conjunto de validación**: Como se observará en los métodos descritos a continuación, el proceso de desagregación genera para cada municipio una estimación de cada uno de los valores, *J*, que puede tomar una determinada variable, *X*. Denotamos esta estimación como $\hat{X}_j^m$, donde *m*

indexa el municipio, $m = 1, 2, 3, ..., M$, y *j* los valores que puede tomar la variable *X*, $j = 1, 2, 3, ..., J$. Para los municipios de validación X_j^m es conocido, de forma que podemos calcular el error absoluto (AE): $\left| \hat{X}_j^m - X_j^m \right|$.

A partir de él utilizamos **dos medidas de bondad del ajuste** en nuestras estimaciones: (*i*) el **promedio de los errores absolutos relativos** (*MARE*), en tanto por cien,

$$MARE = \frac{100}{M \times J} \times \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{\left| \hat{X}_j^m - X_j^m \right|}{X_j^m} \tag{1}$$

y, (*ii*) observando que la suma de los *AE*, $\Sigma_{m=1}^{M} \Sigma_{j=1}^{J} \left| \hat{X}_j^m - X_j^m \right|$, está comprendida entre 0, cuando no se comete ningún error, $\hat{X}_j^m = X_j^m, \ \forall m, j$, y dos veces la población de referencia, $N = \Sigma_{m=1}^{M} \Sigma_{j=1}^{J} X_j^m$, cuando el error es el máximo posible en cada caso, podemos definir una **medida global de error** como el **error absoluto relativo total** (*TARE*), en tanto por cien,

$$TARE = 100 \times \frac{\Sigma_{m=1}^{M} \Sigma_{j=1}^{J} \left| \hat{X}_j^m - X_j^m \right|}{2 \times N} \tag{2}$$

que puede ser interpretado como el porcentaje de población que es mal distribuido en conjunto.[19] Este es un estadístico habitual en el contexto de la desagregación espacial de la población (Goerlich y Cantarino 2013).[20]

Los métodos utilizados tienen una serie de limitaciones, algunas de las cuales derivan de la propia estructura censal, que conviene tener en cuenta:

1. En principio solo se han desagregado variables de los microdatos, sin cruces entre variables.[21] Los métodos descritos más abajo podrían aplicarse a cruces entre varias variables: por ejemplo, relación con la actividad (*RELA*) por nivel de estudios (*ESREAL*). Sin embargo, debe tenerse en cuenta que la fiabilidad de la desagregación debe disminuir con el número de celdas a estimar.

[19] Esto es, asignado a una celda a la que no corresponde.

[20] Al margen de que *MARE* y *TARE* puedan ser calculados a nivel global, siempre es posible examinar la distribución de los *ARE* individuales, $\frac{\left| \hat{X}_j^m - X_j^m \right|}{X_j^m}$. Se observó que esta distribución era siempre muy asimétrica. Además, puesto que el conjunto de validación puede ser visto como una matriz de *M* filas, municipios, y *J* columnas, valores que puede tomar la variable *X*, siempre es posible examinar los errores a nivel de municipio, o a nivel de valores de la variable en cuestión. Igualmente es posible examinar los errores a nivel de estrato en cada provincia.

[21] Con una excepción, la distribución conjunta de estudiante (*ESTUDIANTE*) y relación con la actividad principal (*RELA*).

2. La desagregación tiene un límite derivado de la propia naturaleza de muestra de la información censal. Por ejemplo, a nivel municipal en la información del fichero censal final ponderado disponemos solo de 7 grandes grupos de nacionalidad, además de la española, sin embargo a nivel nacional disponemos de 198 nacionalidades, además de los apátridas. Esta es la relación de códigos que aparece en los microdatos, 198 países, para las variables nacionalidad (*NACI*) y país de nacimiento (*CPAISN*). En principio podríamos pensar en estimar todas las nacionalidades a nivel municipal, sin embargo esto no es posible, simplemente por la falta de representatividad de la muestra, incluso a nivel nacional, para algunas variables. Por ejemplo, a nivel nacional el fichero censal final ponderado indica que existen en nuestro país 2,843 personas de nacionalidad Turca (código 180), sin embargo no hay ningún registro en los microdatos con nacionalidad (*NACI*) Turca, lo mismo sucede con otras nacionalidades. Existe por tanto un límite a lo que es posible desagregar que viene impuesto por la representatividad y consistencia de la muestra con el fichero censal final ponderado.

3. Al margen de las limitaciones a la desagregación mencionadas en el punto anterior, y que derivan de la falta de consistencia entre el FPC y los microdatos, dichas limitaciones también existen a nivel de las variables proporcionadas por los microdatos. Por ejemplo, las variables CNO y CNAE se ofrecen a dos dígitos, lo que representa 61 códigos en el primer caso y 88 en el segundo. Aunque los métodos descritos pueden desagregar a este nivel las variables CNO y CNAE para todos los municipios, es muy probable que, para los municipios más pequeños, la desagregación a 2 dígitos contenga muchos errores, especialmente si no existe para ellos información externa procedente del sistema de *Tablas a Medida*. Por otra parte, el trabajar con población en reales, y valores muy pequeños, los métodos tienden a repartir la escasa población entre un número relativamente grande de categorías, pudiéndose obtener estimaciones ridículamente pequeñas, muy por debajo de la unidad. Por ello estas estimaciones deberán ser utilizadas con las debidas cautelas. En general, estos errores disminuyen al agregar por categorías.

Podría argumentarse, con razón, que los procedimientos que ignoran la información suministrada por el sistema de *Tablas a Medida* del INE son ineficientes, en el mejor de los casos, o manifiestamente erróneos en otras ocasiones. Esta afirmación es cierta, y por esta razón los métodos mecánicos expuestos a continuación tuvieron en cuenta dicha información hasta donde fue posible, ya que el INE es bastante restrictivo en la difusión de información municipal.

Los algoritmos implementados, y sus programas asociados, son lo suficientemente flexibles como para incorporar información externa de las *Tablas a Medida*, o de cualquier otra fuente, cuando esta esté disponible. Basta con poner dicha información en un fichero con el formato adecuado, y la desagregación respetará esta información externa.

3.1. Desagregación de la Población Residente en Viviendas Principales (PRVP)

Como se indica al principio de este apartado la **población de referencia** es la **residente en viviendas principales**; y la **consistencia** con los **microdatos** exige que, para cada municipio, los **valores desagregados coincidan**, al agregarlos, **con el colectivo correspondiente de esta población**.

En ocasiones, el colectivo correspondiente es el total de la población residente en viviendas principales (**PRVP**), pero en otros casos el colectivo se reduce a la clasificación por sexos o grupos de edad, menores de 16 años y de 16 o más años, e incluso en algún caso se necesita el cruce de estas variables o estimaciones previas de variables de clasificación de los microdatos. Son estos colectivos los que actúan de marginales a los que hay que ajustar las estimaciones.

Por esta razón, el primer paso antes de desagregar las variables de los microdatos consistió en la desagregación de la PRVP según los criterios mencionados. El procedimiento seguido fue muy sencillo. Para los 5,608 **municipios** en los que **no existe población empadronada en colectivos** esta **información** es **conocida** a partir del FPC, **y se toma de allí** –con una excepción que se menciona más adelante–. Estos municipios no se estiman y forman parte del conjunto de validación. Para el resto distinguimos entre dos casos: (*i*) municipios que tienen hasta 20,000 habitantes, y (*ii*) municipios de más de 20,000 habitantes.

1. **Municipios de hasta 20.000 habitantes**: Estos municipios **no** se identifican en los microdatos.

Para estos se obtuvo una estimación inicial distribuyendo el porcentaje correspondiente –por sexos, grupos de edad o ambos– para el total de la población sobre la población residente en viviendas principales a nivel municipal, cuyo total es conocido. Esta estimación inicial se ajustó por RAS a nivel de estrato de los microdatos, para que fueran consistentes con ellos a este nivel de agregación.

2. **Municipios de más de 20,000 habitantes**: Estos municipios se identifican en los microdatos.

Para estos municipios las estimaciones correspondientes de la PRVP según los criterios referidos se obtuvieron directamente de los microdatos.

Existen, sin embargo, dos excepciones a las reglas anteriores derivadas de la falta de consistencia entre el FPC y el calibrado de los microdatos.

En primer lugar, existen 14 municipios de más de 20,000 habitantes que no tienen población empadronada en colectivos, y por tanto para estos la población total coincide con la población residente en viviendas principales.[22] Sin embargo, en estos municipios lo que se obtiene del FPC no coincide con lo que se obtiene de los microdatos debido a que el calibrado no se realiza a este nivel de detalle (INE 2014). Es una **falta de consistencia entre el FPC y los microdatos que no es posible solucionar**, y hay que decidir dar prioridad a una u otra fuente de información. Aunque el criterio general es dar prioridad al FPC para la información demográfica, en este caso se dio prioridad a los microdatos, de forma que para estos 14 municipios, que pertenecen a 10 provincias diferentes, las estimaciones correspondientes de la PRVP según los criterios referidos se obtuvieron también directamente de los microdatos.

Hacerlo de otra forma hubiera generado una inconsistencia entre la distribución de otras variables, disponibles en los microdatos, y sus marginales en el estrato de los municipios de más de 20.000 habitantes para las provincias implicadas,[23] al tomar las estimaciones para estos municipios directamente de los microdatos, como parece natural.

En segundo lugar, para dos provincias hay estratos en los que todos los municipios carecen de población empadronada en colectivos –el estrato 2, código 992, de Las Palmas de Gran Canaria (35), y el estrato 1, código 991, de Pontevedra (36). En estos casos se produce la misma falta de consistencia señalada anteriormente, de forma que el marginal derivado de los microdatos simplemente se re-escala a la población correspondiente del FPC, para poder aplicar el proceso RAS.

La desagregación de la *Población Residente en Viviendas Principales* no incorpora información externa de las *Tablas a Medida*, ya que se comprobó que dicha incorporación empeoraba los resultados sobre los municipios de validación.

[22] Estos municipios son: Níjar (04066), Sant Antoni de Portmany (07046), Sant Josep de sa Talaia (07048), Conil de la Frontera (11014), Ames (15002), Maracena (18127), Algete (28009), Arroyomolinos (28015), Barañain (31901), Candelaria (38011), Los Realejos (38031), Mairena del Alcor (41058), Tomares (41093) y Alfafar (46022); y afectan a 10 provincias.

[23] A estos efectos, el conjunto de municipios de más de 20,000 habitantes de cada provincia constituyen un estrato propio, y el número de municipios de este estrato para las 10 provincias afectadas por la inconsistencia es de 147 municipios.

3.2. Desagregación de variables de personas en los microdatos

Considérese una variable categórica, *X*, para un municipio *m*, que toma *J* posibles valores. Por ejemplo, la variable '*Relación con la actividad*', ***RELA***, toma 6 posibles valores, y no se aplica cuando la edad del individuo es menor de 16 años. Por tanto, restringiendo la población a los de 16 o más años, en este ejemplo $J = 6$.

Puesto que cada persona del municipio objeto de estimación debe pertenecer a una de las *J* categorías posibles, la población de dicho municipio, N^m, puede escribirse como $N^m = \Sigma_{j=1}^{J} X_j^m$, donde del superíndice *m* indica el municipio correspondiente.[24] Los valores de X_j^m son desconocidos para cada *j* y *m*, y son las variables que pretendemos estimar. Conocemos la población del municipio, N^m, procedente de la información del fichero censal final ponderado, y también X_j para el estrato al que pertenece el municipio, $X_j = \Sigma_{m \in S} X_j^m$ donde *S* indica el estrato, procedente de los microdatos. En otras palabras, visto en formato de tabla, conocemos las marginales, pero no la distribución conjunta.

Una aplicación mecánica mediante un ajuste iterativo proporcional a partir de una distribución inicial uniforme generará, probablemente, pésimos resultados, por lo que consideramos la incorporación de información auxiliar a la estimación de esa distribución conjunta.

Supongamos que disponemos de otra partición de la población del municipio en *K* clases exhaustivas y mutuamente excluyentes. Igualmente podemos escribir ahora la población del municipio como $N^m = \Sigma_{k=1}^{K} N_k^m$, donde ahora N_k^m es conocido a partir de la información del fichero censal final ponderado.

Consideremos ahora el problema de estimar X_j^m. Por definición,

$$X_j^m = \sum_{k=1}^{K} X_{k,j}^m = \sum_{k=1}^{K} N_k^m \frac{X_{k,j}^m}{N_k^m} \qquad (3)$$

El estimador propuesto estima las tasas que aparecen en (3), $\frac{X_{k,j}^m}{N_k^m}$, a partir del estrato al que pertenece el municipio, *S*, con la información disponible en los microdatos, y aplica dichas tasas a la partición de la población considerada a nivel municipal. Es decir,

[24] Esta notación supone que X_j representan poblaciones de la categoría *j*, no los códigos de dichas categorías, que es como aparece en los microdatos.

$$\hat{X}_j^m = \sum_{k=1}^{K} N_k^m \frac{X_{k,j}^S}{N_k^S} \tag{4}$$

donde $N_k^S = \Sigma_{m \in S} N_k^m$ y $X_{k,j}^S = \Sigma_{m \in S} X_{k,j}^m$. En consecuencia, (4) sustituye en (3) las verdaderas tasas, $\frac{X_{k,j}^m}{N_k^m}, \ \forall k$, por las tasas estimadas para el estrato al que pertenece el municipio, $\frac{X_{k,j}^S}{N_k^S}, \ \forall k$, al no poder identificar los registros del municipio dentro del estrato, y aplica las mismas tasas a todos los municipios del estrato.

El método para obtener $\hat{X}_j^m$ a partir de (4) es sencillo, puede encuadrarse en los denominados métodos demográficos tradicionales en el contexto de las estimaciones en pequeñas áreas (Rao 2003, capítulo 3), o de los denominados estimadores sintéticos (Rao 2004, capítulo 4.2)[25] y puede implementarse de forma generalizada y automática para diversas variables de los microdatos del censo.

Un estimador se denomina sintético si un estimador directo fiable para un área mayor, que cubre varias pequeñas áreas, es utilizado en la obtención de un estimador indirecto para las áreas pequeñas, bajo el supuesto de que las áreas pequeñas tienen características comunes al área más grande. Claramente (4) cae dentro de esta definición, donde el supuesto implícito, es que todos los municipios del estrato *S* presentan las mismas tasas, $\frac{X_{k,j}^S}{N_k^S}, \ \forall k$, y los municipios del estrato solo se diferencian en su estructura demográfica. Más adelante mostraremos como relajar parcialmente esta información a partir de la información disponible en el propio censo.

Este método es conocido como el **método de las propensiones** (Bell et al 1995), y es aplicado por el INE (2013c) en diversos contextos.

Colom, Goerlich, Molés y Murgui (2015) ofrecen una justificación del método en el contexto de los modelos de superpoblación del muestreo tradicional, cuando no es posible identificar los registros de las unidades concretas dentro de un dominio más amplio. Este es el caso de la estructura de los microdatos del censo 2011. Una forma alternativa de ver (4) es

$$\hat{X}_j^m = \sum_{k=1}^{K} \frac{N_k^m}{N_k^S} X_{k,j}^S \tag{5}$$

[25] De hecho el estimador (4) es básicamente el estimador sintético con información auxiliar (4.2.3) en Rao (2004, capítulo 4.2.2, página 47).

que enfatiza como el valor de X_j a nivel de estrato para cada elemento en la partición, $X^S_{k,j}$, es re-escalado por la proporción que representa la población del municipio dentro del estrato, $\frac{N^m_k}{N^S_k}$.

Estos autores muestran como (4) es, en este contexto, un estimador insesgado, aunque ineficiente. No obstante los errores estándar estimados son muy pequeños, y de una magnitud similar a la que el INE ofrece en gran parte de las investigaciones por muestreo. Además, dicho procedimiento genera resultados prácticamente idénticos a los obtenidos mediante modelización de la variable objeto de desagregación utilizando modelos de elección discreta.

Una vez se dispone de $\hat{X}^m_j$ para las J categorías de la variable, y para todos los municipios del estrato, estas estimaciones iniciales se ajustan a los totales marginales conocidos, N^m y X_j, mediante ajuste bi-proporcional iterativo. La estimación se realiza pues a nivel de estrato.

Como partición utilizamos la población municipal por sexos y edades simples hasta la edad de 100 y más años, ya que esta partición está disponible a partir del fichero censal final ponderado, lo que genera un total de 202 celdas, 101 por cada sexo, $K = 202$ en (4).

Ya hemos indicado que la aplicación de (4) descansa sobre el supuesto de que el municipio para el que efectuamos la estimación tiene las mismas características que el estrato al que pertenece, y que las diferencias entre los municipios del mismo estrato radican en su estructura demográfica. Ello implica que cuanto más relacionada esté la variable en cuestión con la demografía, y cuanto más homogéneos sean los municipios dentro del estrato menores serán los errores de estimación.

Sin embargo, podemos mejorar la estimación de las tasas en (4), $\frac{X^S_{k,j}}{N^S_k}, \ \forall k$, utilizando información del propio censo.

1. Aunque no es posible identificar a que municipio pertenecen los registros de un estrato, si es posible eliminar algunos registros del estrato que con seguridad no pertenecen al municipio que estamos estimando. Puesto que disponemos de la población por sexos y edades simples –es la partición que estamos utilizando–, y además conocemos la población por nacionalidades y país de nacimiento en 8 grandes grupos siempre podemos filtrar estos registros para cada municipio

antes de construir $\frac{X_{k,j}^{S}}{N_{k}^{S}}$. Dicho filtrado se realiza además a nivel de hogar.[26]

Sería de esperar que esto mejorara fundamentalmente las estimaciones en municipios pequeños, al acercar las estimaciones de las tasas a la estructura de su población.

2. En ocasiones se observó que el supuesto de homogeneidad de los municipios dentro del estrato no era muy apropiado para algunos municipios y variables. Este era el caso, por ejemplo, de municipios de tamaño intermedio, entre 5,001 y 10,000 habitantes, estrato 3, o entre 10,001 y 20,000, estrato 4, en los que las características de estos municipios eran muy diferentes si pertenecían a un área urbana o se trataba de municipios relativamente grandes, pero que no estaban en el área de influencia de una gran ciudad. Este caso es difícil de solucionar con generalidad. Se ensayó con la posibilidad de añadir a los registros del estrato a partir del cual vamos a calcular $\frac{X_{k,j}^{S}}{N_{k}^{S}}$, los registros de los municipios con más de 20,000 habitantes que fueran físicamente contiguos –compartan lindes municipales– al municipio en cuestión. Así como la adición iterativa de todos los vecinos físicamente contiguos, hasta que ya no se añadiera ninguno más. Sería de esperar que este ajuste mejorara fundamentalmente las estimaciones de los municipios intermedios que estén cerca de un área urbana, y en consecuencia bajo su área de influencia.

   Sin embargo, este ajuste, que dio buenos resultados para algunos municipios, empeoró las estimaciones de otros, de forma que no quedaba claro que, en conjunto, los resultados mejoraran.

3. Finalmente, y puesto que es posible conseguir estimaciones del sistema de consultas de *Tablas a Medida* del INE se decidió mantener el método original, fórmula (4), en el que la información municipal se estima con información del estrato correspondiente y la pirámide demográfica del municipio en cuestión, pero explotar al máximo la información procedente del sistema de *Tablas a Medida* –lo que no es factible con generalidad–, esta se incorpora al proceso, sustituyéndose, para esos municipios, las estimaciones derivadas de (4) por las de las *Tablas a Medida*, antes del ajuste bi-proporcional a nivel de estrato.

[26] Es decir, si en un municipio no hay varones de 97 años de edad se eliminan de los cálculos del municipio los registros de todos los hogares con algún varón de 97 años.

Con este procedimiento, para cada municipio se aplican unas tasas por sexo y edad, $\frac{X_{k,j}^{S}}{N_{k}^{S}}$, diferentes, que partiendo de los registros del estrato al que pertenece el municipio, eliminan los registros que, con seguridad, no pertenecen a dicho municipio, en base a la información demográfica disponible en el fichero censal final ponderado. A continuación, se incorpora la información disponible de las *Tablas a Medida* para los municipios para los que está disponible. Esta incorporación de información externa a los microdatos complicó notablemente el proceso de desagregación, a costa de ganar precisión en las estimaciones de forma considerable. Cada variable requirió sus propios ajustes.

De esta forma, con este método, convenientemente adaptado, se desagregaron las 55 variables de personas contenidas en el anexo.[27]

Un análisis de errores mostró que estos son despreciables para los municipios de validación, los de más de 20.000 habitantes, ya que para estos se dispone siempre de información en las *Tablas a Medida*. Puesto que para muchos otros municipios también se dispone de esa información, de forma total o parcial, los errores de estimación para estos serán igualmente despreciables, y cuanta más información se disponga del sistema de *Tablas a Medida* mejores serán las estimaciones para los municipios faltantes, ya que se procede a un ajuste a nivel de estrato en los microdatos, hasta donde el FPC y los microdatos son compatibles.

4. Conclusiones

Este documento presenta los métodos seguidos para la obtención, con generalidad, de variables a nivel municipal derivadas del censo 2011. Los métodos de desagregación son sencillos, y los *scripts* desarrollados permiten la incorporación de información externa derivada del sistema de *Tablas a Medida* del INE, lo que resultó esencial para mejorar la precisión de las estimaciones.

Los procedimientos empleados deben superar numerosas pequeñas inconsistencias entre los dos pilares sobre los que se ha construido el Censo del 2011, el fichero censal final ponderado y la muestra, de donde se obtienen todas las características de la población más allá de las meramente demográficas. Al margen de estas pequeñas inconsistencias las estimaciones generadas son totalmente consistentes, a nivel municipal y a nivel de los estratos a los que pertenecen los diversos municipios en los microdatos. Aunque todas las variables desagregadas son a nivel de persona, idénticos

[27] Además de los marginales correspondientes de la población residente en viviendas principales. Los detalles concretos para cada variable se explican en un documento metodológico extenso complementario a este trabajo.

métodos pueden emplearse para variables de hogar. Métodos similares podrían emplearse para las variables de viviendas y edificios.

Finalmente unas palabras de precaución. Los resultados deben ser tomados como lo que son, estimaciones sobre una muestra censal, con el objeto de disponer de estimaciones para todos los municipios, y con esta precaución deben ser utilizados. La información derivada del sistema de *Tablas a Medida* ha sido explotada al máximo en la medida de lo posible, pero esta es en ocasiones limitada o parcial, y nunca está disponible para la totalidad de los municipios.

Referencias

Bell, M.; Cooper, J.; *et al* (1995) *Household and Family Forecasting Models. A review*. Canberra. Department of Housing and Regional Development, 68 p.

Centro de Ciencias Humanas y Sociales (CCHS 2015) *El futuro de las estadísticas demográficas del INE y el Censo de Población de 2021*. Centro Superior de Investigaciones Científicas (CSIC). 21 y 22 de octubre de 2015. Información en línea: http://cchs.csic.es/es/event/congreso-futuro-estadisticas-demograficas-ine-censo-poblacion-2021. [Consultado 25/04/2016].

Colom Andrés, Mª C.; Goerlich Gisbert, F. J.; Molés Machí, Mª C. y Murgui Izquierdo, S. (2015) *Estimación de proporciones a partir de diseños no aleatorios: Aplicación al Censo de Población de 2011*, trabajo presentado en XXIX Congreso Internacional de Economía Aplicada. Métodos Cuantitativos para la Economía y la Empresa. ASEPELT 2015. Cuenca, 24-27 de junio de 2015.

Bacharach, M. (1965) "Estimating Nonnegative Matrices from Marginal Data". *International Economic Review*, 6, 3, 294-310.

Deming, W. E. y Stephan, F. F. (1940) "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known". *Annals of Mathematical Statistics*, 11, 4, 427-444.

Deville, J.-C. y Särndal, C.-E. (1992) "Calibration estimators in survey sampling". *Journal of the American Statistical Association*, 87, 376–382.

Deville, J.-C., Särndal, C.-E. y Sautory, O. (1993) "Generalized raking procedure in survey sampling". *Journal of the American Statistical Association*, 88, 1013–1020.

Elbers, C.; Lanjouw, J. O. y Lanjouw, P. (2003) "Micro-level estimation of poverty and inequality". *Econometrica*, 71, 1, 355–364.

Goerlich Gisbert, F. J. (2007) "¿Cuantos somos? Una excursión por las estadísticas demográficas del Instituto Nacional de Estadística (INE)". *Boletín de la Asociación de Geógrafos Españoles*, 45, 123-156.

Goerlich Gisbert, F. J. (2012) "Estimaciones de la población actual (*ePOBa*) a nivel municipal. Discrepancias Censo-Padrón a pequeña escala". *Boletín de la Asociación de Geógrafos Españoles*, 58, 83-104.

Goerlich Gisbert, F. J. y Cantarino Martí, I. (2013) "A population density grid for Spain". *International Journal of Geographical Information Science*, 27, 12, 2247–2263. doi:10.1080/13658816.2013.799283.

Goerlich Gisbert, F. J. y Cantarino Marti, I. (2016) "*Grid* poblacional 2011 para España. Evaluación metodológica de diversas posibilidades de elaboración", *Estudios Geográficos*. Pendiente de publicación.

Goerlich, F. J.; Ruiz, F.; Chorén, P. y Albert, C. (2015) "Cambios en la estructura y localización de la población. Una visión de largo plazo (1842-2011)", Fundación BBVA. 2015. Bilbao. pp.- 354.

Instituto Nacional de Estadística (INE *on line*) *¿Qué tipos de cifras de población publica el INE?* Documentación en línea: http://www.ine.es/daco/daco43/epoba/cifras.pdf [Consultado 20/05/2016].

Instituto Nacional de Estadística (INE 2011) *Proyecto de los Censos Demográficos 2011*. Subdirección General de Estadísticas de la Población. (Febrero).

Instituto Nacional de Estadística (INE 2012) *Metodología de cálculo de las cifras de población censal*. Documentación en línea: http://www.ine.es/censos2011/censos2011_meto_calculo.pdf. [Consultado 20/09/2013].

Instituto Nacional de Estadística (INE 2013a) *Censos de Población y Viviendas 2011*. Resultados detallados. Información en línea:

http://www.ine.es/censos2011_datos/cen11_datos_inicio.htm. [Consultado 17/12/2013].

Instituto Nacional de Estadística (INE 2013b) *Población residente en establecimientos colectivos (Encuesta de colectivos del Censo de Población y Viviendas 2011*. Metodología. Documentación en línea: http://www.ine.es/censos2011/censos2011_meto_pobla_colectivos.pdf [Consultado 20/05/2016].

Instituto Nacional de Estadística (INE 2013c) *La producción de información demográfica en el INE a partir del Censo de 2011.* Curso de la Escuela de Estadística de las Administraciones Públicas (EEAP). INE. Madrid, 14-15 de marzo de 2013.

Instituto Nacional de Estadística (INE 2014) *Censo 2011. Productos para consultar esta información.* Curso de la Escuela de Estadística de las Administraciones Públicas (EEAP). INE. Madrid, 3 de marzo de 2014.

Rao, J. N. K. (2003) *Small Area Estimation*. Wiley Series in Survey Methodology. John Wiley & Sons, Inc. Hoboken, New Jersey.

Reig, E.; Goerlich, F. J. y Cantarino, I. (2016) *Delimitación de Áreas Rurales y Urbanas a Nivel Local. Demografía, Coberturas del Suelo y Accesibilidad.* Fundación BBVA.

Stephan, F. F. (1942) "Iterative method of adjusting frequency tables when expected margins are known". *Annals of Mathematical Statistics*, 13, 2, 166-178.

Anexo: Microdatos – Desagregación de variables de personas

Variables de personas desagregadas mediante la fórmula (4) utilizando como partición la población municipal por sexos y edades simples. Los registros del estrato al que pertenece el municipio se filtran con los registros que no pueden pertenecer al municipio correspondiente en base a las características de sexo, edad, nacionalidad y país de nacimiento. La información del sistema de *Tablas a Medida* se incorpora en la medida en que está disponible.

Tabla A.1: Variables de Personas desagregadas por los métodos expuestos en el texto.

Variables que actúan de marginales en el proceso de desagregación		
1	Población residente en viviendas principales por grupos de edad	
	Menor de 16 años	PRPVM16
	De 16 y más años	PRVP16M
2	Población residente en viviendas principales por sexo	
	Hombres	PRVPVAR
	Mujeres	PRVPMUJ
3	Población residente en viviendas principales por sexo y grupos de edad	
	Hombres menores de 16 años	PRVPVARM16
	Hombres de 16 y más años	PRVPVAR16M
	Mujeres menores de 16 años	PRVPMUJM16
	Mujeres de 16 y más años	PRVPMUJ16M
Variables de clasificación de los microdatos		
4	Municipio de residencia actual y Municipio de residencia anterior	RES_ANTERIOR
5	Municipio de residencia actual y Municipio de residencia hace 1 año	RES_UNANO
6	Municipio de residencia actual y Municipio de residencia hace 10 años	RES_DANO
7	Pasa más de 14 noches en segundo municipio	SEG_VIV
8	Disponibilidad vivienda en segundo municipio	SEG_DISP
9	Estado civil	ECIVIL
10	Acude a un centro escolar	ESCOLAR
11	Nivel de estudios completados (grados)	GRADOS
12	Nivel de estudios completados (detalle)	ESREAL
13	Tipo de estudios realizados	TESTUD
14	Cuidar a un menor de 15 años	TAREA1
15	Cuidar a una persona con problemas de salud	TAREA2
16	Tareas benéficas o voluntariado social	TAREA3
17	Encargárse de la mayor parte de las tareas domésticas de su hogar	TAREA4
18	Indicador de si la mujer ha tenido hijos	HIJOS
19	Relación preferente con la actividad (activo / inactivo)	ACTIVO
20	Relación preferente con la actividad (detalle)	RELA
21	Tipo de jornada de trabajo	JORNADA
	Código de ocupación	
22	a 1 dígito	OCUPACION
23	a 2 dígitos	CNO
	Código de actividad económica a 2 dígitos	
24	Rama	RAMA
25	Letra	LETRA
26	a 2 dígitos	CNAE
27	Situación profesional	SITU
28	Condición socioeconómica	CSE
29	Estudiantes (ESCUR1): Si/No	ESTUDIANTE
	Estudios en curso: Tipo de Estudios	
30	01 - ESO, Educación secundaria para adultos	ESCUR01
31	02 - Programas de Cualificación Profesional Inicial	ESCUR02
32	03 - Bachillerato	ESCUR03
33	04 - Grado Medio de FP, de Artes Plásticas y Diseño y de EE. Deportivas o equivalentes	ESCUR04
34	05 - Enseñanzas de Escuelas oficiales de Idiomas	ESCUR05
35	06 - Enseñanzas Profesionales de Música y Danza	ESCUR06
36	07 - Grado Superior de FP, de Artes Plásticas y Diseño y de EE. Deportivas o equivalentes	ESCUR07
37	08 - Diplomatura universitaria, Arquitectura Técnica, Ingeniería Técnica o equivalente	ESCUR08
38	09 - Estudios de Grado Universitario y de Enseñanzas Artísticas y equivalentes	ESCUR09
39	10 - Licenciatura, Arquitectura, Ingeniería o equivalente	ESCUR10
40	11 - Máster oficial universitario, Especialidades Médicas o análogos	ESCUR11
41	12 - Doctorado	ESCUR12
42	13 - Otros cursos de educación reglada (Enseñanzas iniciales para adultos,...)	ESCUR13
43	14 - Cursos de Formación de los Servicios Públicos de Empleo	ESCUR14
44	15 - Otros cursos de formación no reglados.	ESCUR15
45	Estudiantes (Si/No) según relación con la actividad (3 categorías): 6 categorías	ESTURELA
46	Población que trabaja o estudia: Si/No	TRABAEST
47	Lugar de trabajo o estudio	LTRABA
48	Número de viajes diarios	NVIAJE
	Medio de desplazamiento	
49	01 - En coche o furgoneta como conductor	MDESP01
50	02 - En coche o furgoneta como pasajero	MDESP02
51	03 - En autobús, autocar, minibus	MDESP03
52	04 - En metro	MDESP04
53	05 - En moto	MDESP05
54	06 - Andando	MDESP06
55	07 - En tren	MDESP07
56	08 - En bicicleta	MDESP08
57	09 - Otros medios	MDESP09
58	Tiempo de desplazamiento	TDESP

Fuente: INE (2013) - Censo 2011