Pavía Miralles, JM and Veres Ferrer, EJ (2016) "Desagregando Estadísticas de Población", in *Investigaciones en Métodos Cuantitativos para la Economía y la Empresa*, Herrerías, JM and Callejón J (eds.), Editorial Universidad de Granada, pp. 543-555.

DESAGREGANDO ESTADÍSTICAS DE POBLACIÓN

JOSE M. PAVÍA
ERNESTO J. VERES FERRER
Grupo de Investigación en Procesos Electorales y Opinión Pública,
Facultat d'Economía, Departamento de Economía Aplicada
Universitat de Valencia

1. INTRODUCCIÓN

Debido a cuestiones de confidencialidad y a fin de evitar que cualquier persona concreta pueda ser identificada (de forma directa, o indirectamente en conexión con cualquier otra información publicada), las leyes de protección de datos de carácter personal, tanto nacionales como internacionales, obligan a que las estadísticas referidas a variables sociales, económicas y demográficas deban ser publicadas agregadas por unidades espaciales (BOE, 1999; OJEU 2013; USC 2002).

La sección censal es la unidad espacial más pequeña utilizada en España para diseminar variables sociodemográficas. A partir de la explotación estadística del padrón municipal, el Instituto Nacional de Estadística (INE) ofrece para cada una de las más de 35.000 secciones censales en que se divide el territorio español, y de forma abierta y gratuita, cifras relativas a: (i) el número de residentes por sexo y edad (agrupados en grupos quinquenales), (ii) el número de residentes por sexo y nacionalidad (agrupados por continentes y para las principales nacionalidades), el número de residentes por sexo clasificados en función de la relación de su lugar de nacimiento y de residencia, y el número de residentes por sexo y país de nacimiento (INE, 2016). Todas las cifras de población que ofrece el INE en el ámbito de la sección censal en su página web vienen además referidas al uno de enero de cada año.

Aunque la información disponible en la web del INE es muy valiosa y bastante detallada, en ocasiones el analista precisa para sus estudios de datos más desagregados y/o referidos a otros instantes temporales. Tal es el caso de las elecciones, en las que el número de nuevos electores (por haber alcanzado la mayoría de edad desde unas elecciones anteriores) por sección censal (o mesa) es una variable no disponible y de interés para la implementación de técnicas de inferencia ecológica (e.g., King, 1997) o para su utilización en

modelos de predicción electoral basados en pequeñas áreas (e.g., Pavía-Miralles, 2005; Pavía y Larraz, 2012).

Dado que conocer el número de nuevos electores en cada (mesa o) sección censal no permite la identificación individual de ninguna persona, previa petición y bajo pago, el analista interesado puede comprar tal información al INE. Lamentablemente, el coste de los datos suele ser muy elevado para el analista medio y la producción de la información puede sufrir retrasos que la conviertan en poco útil una vez recibida. Por ejemplo, si nuestro objetivo es emplear la variable como un predictor en un modelo especificado a nivel de sección censal, donde la variable respuesta son los porcentajes de intención de voto recolectados en una encuesta pre-electoral mediante un muestreo por conglomerados (con las secciones censales como conglomerados), no es inverosímil que los datos se reciban sin la antelación suficiente como para poder producir/publicar nuestras estimaciones antes del período de embargo de publicación de resultados de encuestas que impone la ley electoral en España (BOE, 1985).

Ante esta circunstancia el analista debe buscar estrategias que le permitan generar estimaciones de la variable de interés con un coste asumible y que puedan estar disponibles cuando son de utilidad. En este trabajo, se muestra una estrategia para estimar, a partir de la información pública disponible, el número de nuevos electores por sección censal.

La calidad de la aproximación propuesta es además evaluada comparando, tanto en cifras absolutas como en relativas, las estimaciones que se obtienen con las estadísticas reales para varios procesos electorales. En concreto, se han considerado como casos de estudio las elecciones a Corts Valencianes de 2015, las elecciones a Cortes de Castilla-La Mancha de 2015 y las elecciones a la Asamblea de Madrid de 2015; calculándose en todos los casos los nuevos electores respecto a las mismas elecciones celebradas en 2011.

El resto del documento está estructurado como sigue. En la sección segunda se ofrecen los detalles metodológicos de la aproximación. En la sección tercera se muestran las estimaciones obtenidas y se comparan con los valores reales. La sección cuarta está dedicada a un análisis de los errores de estimación con el objetivo de buscar patrones que nos ayuden a conocer las fortalezas y debilidades del estimador propuesto y, en su caso, a proponer mejoras. La sección quinta concluye.

2. METODOLOGÍA

El objetivo de este trabajo es proponer una estrategia para, a partir de la información pública más desagrada disponible a nivel de sección censal, estimar en todas las secciones censales en que se divide una determinada

provincia (incluidas las ciudades autónomas de Ceuta y Melilla) el número de nuevos electores que tienen derecho a voto por haber alcanzado la mayoría de edad respecto a un proceso electoral previo. En concreto, nos centramos en estimar el número de nuevos electores correspondientes al censo CER (Censo de Españoles Residentes).

Aunque en determinados procesos electorales (elecciones al Parlamento Europeo y elecciones locales) podría haber (en un número no despreciable de secciones censales) también nuevos electores pertenecientes al colectivo CERE (Censo electoral de Extranjeros Residentes en España), en aras a simplificar la exposición centraremos nuestro escrutinio en los nuevos electores CER. Con ello se evitan las mayores complejidades de notación y logísticas que conlleva manejar las variables por nacionalidades con derecho a voto (que son distintas dependiendo de la elección) y, sobre todo, las dificultades que en el proceso de estimación introducen los aspectos legales de inscripción previa en el censo electoral (BOE, 2011). Las ideas que se exponen, no obstante, serían trasladables a la estimación del número (potencial) de nuevos electores CERE.

Asimismo supondremos, como es habitual, que la elección previa respecto de la cual se pretenden calcular el número de nuevos electores corresponde a la última elección del mismo tipo o de diferente tipo celebrada. Es decir, admitimos que si, por ejemplo, pretendiésemos estimar el número de nuevos electores CER en las secciones censales de Granada correspondientes a las elecciones generales de 2015, los cálculos se realizarían respecto a algunas de las siguientes elecciones: las generales de 2011, las elecciones europeas de 2014, las elecciones autonómicas de 2015 o las elecciones locales de 2015.

Además de las consideraciones anteriores, suponemos que el momento en que deseamos realizar la estimación se sitúa en un instante anterior, no excesivamente alejado, de la celebración de la elección.[1]

Previo a cualquier proceso electoral en España las últimas cifras de población disponibles suelen estar referenciadas al 1 de Enero del año anterior a la elección. De hecho, las cifras referidas a un año t cualquiera se suelen publicar a lo largo del primer trimestre del año $t+1$. Esto, unido a las hipótesis de los párrafos anteriores, implica que el número de nuevos electores en cada sección se encuentren necesariamente contabilizados entre los residentes registrados en las últimas cifras publicadas previas a la elección dentro de los grupos de 15 a 19 años y de 20 a 24 años, asumiendo que no ha habido (entre la fecha de referencia de las cifras de población y la fecha de

1. Obviamente la estimación se podría hacer después de celebradas las elecciones (incluso años después), no obstante, la asunción realizada se corresponde con el escenario más verosímil correspondiente a una situación de utilidad inmediata, más allá de la utilidad académica.

celebración de la elección) fallecimientos ni traslados de domicilios dentro de esos colectivos[2].

El objetivo es, por tanto, estimar en cada sección censal cuántas de las personas contabilizadas dentro de los grupos de 15 a 19 años y de 20 a 24 años son nuevos electores. Para ello se hará uso de toda la información pública relevante disponible a nivel de sección censal.

En concreto, y a fin de poder expresar analíticamente el estimador denotamos por:

- t_1 la fecha en la que se celebró la elección previa respecto a la cual se desea estimar el número de nuevos electores;
- t_2 la fecha de referencia de las cifras de población disponible;
- t_3 la fecha de celebración de las elecciones actuales[3];
- P_j^{15-19} al número de residentes contabilizados con entre 15 y 19 años cumplidos en t_2 en la sección censal j;
- P_j^{20-24} al número de residentes contabilizados con entre 20 y 24 años cumplidos en t_2 en la sección censal j;
- E_j al número total de residentes contabilizados en t_2 en la sección censal j con nacionalidad española;
- X_j al número total de residentes contabilizados en t_2 en la sección censal j sin nacionalidad española;
- E^d al número total de residentes contabilizados en t_2 con d años cumplidos (para $d = 15, 16, \dots, 24$) y nacionalidad española en la provincia de estudio;
- NE al número total de nuevos electores CER que han alcanzado la mayoría de edad entre t_1 y t_3 en la provincia objeto de estudio;[4] y,
- $\widehat{ne}_j$ a la estimación del número de nuevos electores CER que han alcanzado la mayoría de edad entre t_1 y t_3 en la sección j-ésima de la provincia objeto de estudio.

La estimación $\widehat{ne}_j$ se obtiene después de aplicar secuencialmente una serie de hipótesis y transformaciones a los datos. En particular, se supone: (i) que dentro de cada sección censal la proporción de españoles en cada grupo quinquenal de edad es igual a la proporción de españoles en el conjunto de la sección; (ii) que la suma de españoles (y extranjeros) en el conjunto de las

2. Las bajas tasas de mortalidad que se registran a esas edades invitan a pensar que las desviaciones que pueda introducir esta hipótesis no debieran ser significativas; mientras que el supuesto de no movilidad geográfica dentro de estos grupos es una hipótesis necesaria cuyo impacto, aunque puntual y localmente significativo, no debería ser, en general, muy alto.

3. Habitualmente $t_1 < t_2 < t_3$, pero no es necesario.

4. Esta variable suele ser publicada alrededor de un mes antes de la fecha de celebración de la elección.

secciones censales para cada grupo quinquenal debe coincidir con el total de españoles (extranjeros) en la provincia en ese grupo quinquenal; (iii) que la distribución de españoles para cada edad dentro de cada grupo quinquenal en cada sección censal coincide con la distribución por edades de españoles en el grupo quinquenal en el conjunto de la provincia; (iv) que dentro de cada edad las fechas de cumpleaños se distribuyen uniformemente; y, (v) que la suma del número de nuevos electores en el conjunto de las secciones censales debe coincidir con el número de nuevos electores en la provincia.

La aplicación secuencial de las hipótesis y condiciones anteriores permite obtener en cada sección censal un número decimal (bruto) de nuevos electores, $\widehat{bne}$. Estos números son también depurados utilizando un mecanismo de redondeo y ajuste para que se cumpla (v) en números enteros. En concreto, cada una de las estimaciones brutas es redondeada al entero más próximo y calculada la diferencia, D, entre la suma de los valores enteros y NE, de forma que (a) si D es positivo se resta uno a los D valores cuya diferencia entre la estimación bruta y la estimación entera sea más pequeña y (b) si D es negativo se suma uno a los D valores cuya diferencia entre la estimación bruta menos la estimación entera sea más grande.

En particular, de forma analítica el proceso secuencial en cinco etapas anterior se puede expresar como sigue:

(i) Se estima en cada sección censal ($j = 1, \dots, N$) la proporción de españoles en cada grupo quinquenal de edad asumiendo que la distribución de españoles/extranjeros en cada grupo es igual al del conjunto de la sección:

$$E_j^{15-19} = \frac{P_j^{15-19}}{E_j + X_j} E_j \qquad\qquad E_j^{20-24} = \frac{P_j^{20-24}}{E_j + X_j} E_j$$

(ii) Se ajustan (utilizando estimadores ratio) las estimaciones obtenidas en (i) para que la suma para las N secciones censales de los españoles por grupo quinquenal coincida con el total de españoles en la provincia por grupo quinquenal:

$$AE_j^{15-19} = \frac{\sum_{d=15}^{19} E^d}{\sum_{k=1}^{N} E_k^{15-19}} E_j^{15-19} \qquad\qquad AE_j^{20-24} = \frac{\sum_{d=20}^{24} E^d}{\sum_{k=1}^{N} E_k^{20-24}} E_j^{20-24}$$

(iii)-(iv) Se obtienen estimaciones iniciales del número bruto de nuevos electores por sección censal, $\widehat{ibne}$, bajo las hipótesis de que la distribución de españoles por edad dentro de cada grupo quinquenal en cada sección censal coincide con la distribución por edades de españoles en el grupo quinquenal del conjunto de la provincia y de distribución uniforme de fechas de cumpleaños dentro de cada edad.

$$\widehat{ibne}_j = AE_j^{15-19}(f_1(e_{min}) - f_1(e_{max})) + AE_j^{20-24}(f_2(e_{max}) - f_2(e_{min}))$$

donde:

$e_{min} = 18 - a(t_3, t_2)$ es la edad mínima que podría tener un elector en el instante t_2 para no habiendo tenido la edad mínima para votar en t_1 tenga derecho a voto en t_3; con $a(t_3, t_2)$ representando la distancia en años entre t_3 y t_2.
$e_{max} = e_{min} + a(t_2, t_1)$ es la edad máxima que podría tener un elector en el instante t_2 para no habiendo tenido la edad mínima para votar en t_1 tenga derecho a voto en t_3; con $a(t_2, t_1)$ representando la distancia en años entre t_2 y t_1.
Y las funciones f_1 y f_2 definidas mediante:

$$f_1(d) = \frac{1}{\sum_{d=15}^{19} E^d} \begin{cases} 0 & d \notin [15, 20] \\ E^{19}(20-d) & 19 \leq d \leq 20 \\ E^{19}+E^{18}(19-d) & 18 \leq d \leq 19 \\ \sum_{d=18}^{19} E^d+E^{17}(18-d) & 17 \leq d \leq 18 \\ \sum_{d=17}^{19} E^d+E^{16}(17-d) & 16 \leq d \leq 17 \\ \sum_{d=16}^{19} E^d+E^{15}(16-d) & 15 \leq d \leq 16 \end{cases}$$

$$f_2(d) = \frac{1}{\sum_{d=20}^{24} E^d} \begin{cases} 0 & d \notin [20, 25] \\ E^{20}(d-20) & 20 \leq d \leq 21 \\ E^{20}+E^{21}(d-21) & 21 \leq d \leq 22 \\ \sum_{d=20}^{21} E^d+E^{22}(d-22) & 22 \leq d \leq 23 \\ \sum_{d=20}^{22} E^d+E^{23}(d-23) & 23 \leq d \leq 24 \\ \sum_{d=20}^{23} E^d+E^{24}(d-24) & 24 \leq d \leq 25 \end{cases}$$

(v) Se ajustan (utilizando estimadores ratio) las estimaciones brutas obtenidas en el paso anterior para que la suma para las N secciones censales del número de nuevos electores coincida con el total de la provincia:

$$\widehat{bne}_j = = \frac{NE}{\sum_{k=1}^{N} \widehat{ibne}_k} \widehat{ibne}_j$$

Finalmente, y utilizando el proceso descrito previamente las estimaciones decimales brutas, $\widehat{bne}$, son aproximadas a soluciones enteras. Todo el proceso anterior ha sido programado, en el software estadístico R (R Core Team, 2016), en una función cuyo código y descripción de utilización puede encontrarse en Pavía y Veres (2016).

3. EVALUACIÓN DEL ESTIMADOR

Con el estimador propuesto en el apartado anterior y utilizando los datos de población disponibles en la web del INE a fecha 1 de enero de 2014, se han estimado para las elecciones a Corts Valencianes de 2015, las elecciones a

Cortes de Castilla-La Mancha de 2015 y las elecciones a la Asamblea de Madrid de 2015 el número de nuevos electores CER que (respecto a las mismas elecciones celebradas en 2011) corresponderían a cada sección censal de cada una de las provincias de Alicante, Valencia, Castellón, Albacete, Ciudad Real, Cuenca, Guadalajara, Toledo y Madrid. El estimador propuesto es evaluado comparando, en cifras absolutas y en relativas, las estimaciones obtenidas con las estadísticas reales facilitadas por el INE.

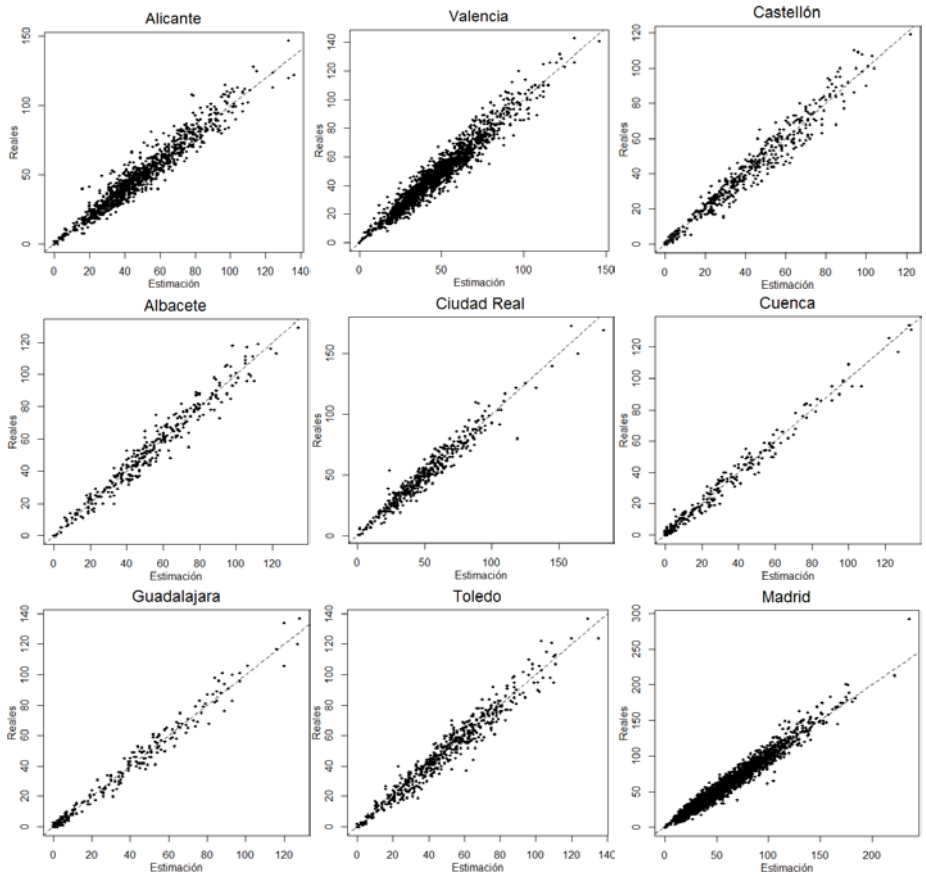

Figura 1. *Comparación a nivel de sección censal entre el número de nuevos electores reales y estimados en cada una de las provincias analizadas. Nuevos electores correspondientes a las elecciones autonómicas de 2015 respecto a las mismas elecciones de 2011. La distancia a la diagonal (línea discontinua) de cada punto informa sobre la magnitud del error asociada a cada estimación. Fuente: Elaboración propia.*

Ambas comparaciones, en términos absolutos y en términos relativos, son relevantes puesto que, por una parte, el número total de nuevos electores en cada sección censal es uno de los componentes de las distribuciones marginales de origen en la estimación de las matrices de transferencia de voto mediante técnicas de inferencia ecológica; mientras que, por otra parte, el número relativo de nuevos electores puede ser empleado como una de las variable explicativas en los modelos multivariantes predictivos de proporciones de votos.

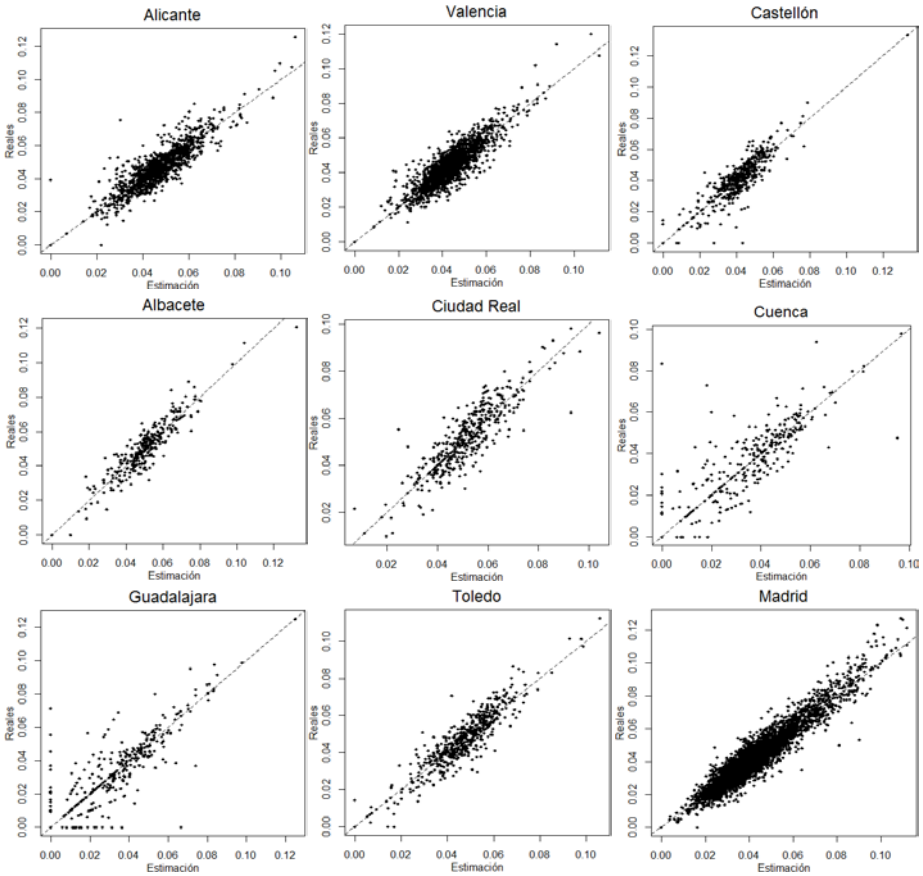

Figura 2*. Comparación a nivel de sección censal entre las proporciones, respecto a censo, de nuevos electores reales y estimados en cada una de las provincias analizadas. Nuevos electores correspondientes a las elecciones autonómicas de 2015 respecto a las mismas elecciones de 2011. La distancia a la diagonal (línea discontinua) de cada punto informa sobre la magnitud del error asociada a cada estimación. Fuente: Elaboración propia.*

La Figura 1 muestra gráficamente los diagramas de dispersión correspondientes a los números de nuevos electores reales y estimados para cada una de las secciones censales de cada una de las provincias analizadas. La Figura 2 presenta los mismas cantidades pero relativizadas, divididas por el censo de cada sección censal. En ambos conjuntos de gráficos, la distancia a la diagonal (marcada con una línea discontinua en cada panel) da idea del error cometido en las estimaciones.

Como se observa, las estimaciones logradas son, en general, de gran calidad. A la vista de las nubes de puntos dibujadas en las Figuras 1 y 2, las soluciones que se obtienen con el estimador propuesto están muy próximas a los valores reales. De hecho, las correlaciones están muy próximas a la unidad, especialmente en cifras absolutas (ver Cuadro 1). En términos relativos, las correlaciones son algo menores, aunque todas ellas en el entorno de 0.9.

Cuadro 1. *Correlaciones entre valores reales y estimados en valores absolutos y en términos relativos respecto a censo en números enteros y en decimales.*

	Alicante	Valencia	Castellón	Albacete	C. Real	Cuenca	Guadal.	Toledo	Madrid
Absolutos	0,9647	0,9659	0,9762	0,9776	0,9718	0,9925	0,9933	0,9753	0,9801
Rel. Enteros	0,8774	0,8742	0,8710	0,9225	0,8768	0,8439	0,8795	0,9143	0,9463
Rel. Brutos	0,8794	0,8735	0,8811	0,9234	0,8773	0,8417	0,8718	0,9146	0,9466

Fuente: Elaboración propia.

Para completar el análisis, en el Cuadro 1 también se ofrecen las correlaciones que se obtienen con las estimaciones relativas logradas sobre las estimaciones brutas, $\widehat{bne}_j$, en decimales. A la vista de los resultados, y a pesar de que cuando el interés se centra en las cifras relativas podría argumentarse que podrían obtenerse mejores estimaciones si se trabajase con las estimaciones no redondeadas (brutas), se constata que trabajando con estimaciones relativas no existen grandes diferencias entre las estimaciones en decimales y las estimaciones enteras; aunque ciertamente parecen ligeramente preferibles, por muy poco, las primeras.

4. ANÁLISIS DE ERRORES

Los análisis realizados en la sección anterior muestran que el estimador propuesto es capaz de generar estimaciones de gran calidad. En este apartado se realiza un análisis de los errores de estimación, $ne_j - \widehat{ne}_j$, con el objetivo de buscar algunos patrones que nos ayuden a conocer las fortalezas y

debilidades del estimador propuesto para, en su caso, proponer alguna mejora o variante del estimador.

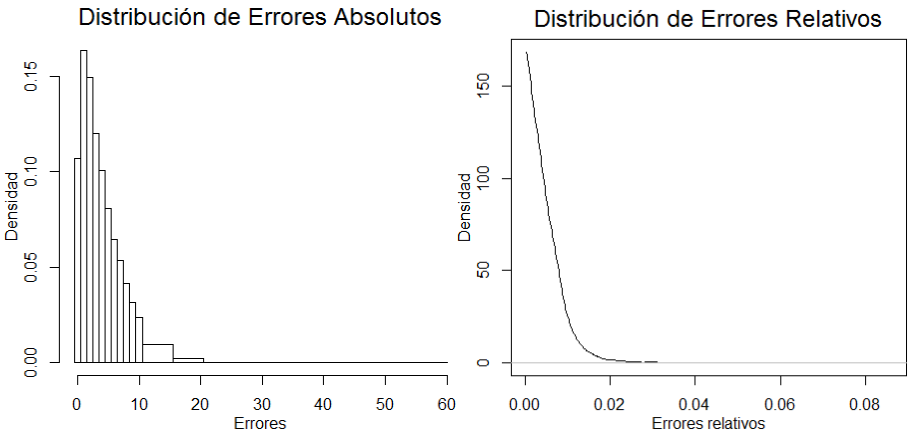

Figura 3. *Distribuciones de errores. Panel izquierdo histograma de las diferencias en valor absoluto a nivel de sección censal entre el número de nuevos electores reales y estimados en el conjunto de provincias analizadas. Panel derecho densidad estimada de las diferencias relativas en valor absoluto, respecto al censo, a nivel de sección censal entre el número de nuevos electores reales y estimados en el conjunto de provincias analizadas. Densidad estimada utilizando la función* density.reflected *de la librería* GoFKernel *(Pavía, 2015). Fuente: Elaboración propia.*

La Figura 3 presenta gráficamente la distribución de los errores de estimación en valores absolutos (módulo de valores reales menos estimaciones): el panel izquierdo mediante un histograma sobre los errores asociados a las estimaciones y el panel derecho empleando una estimación núcleo de la densidad de los errores relativos, calculados como fracción del censo de cada sección censal. Como era de esperar, la mayoría de los errores en términos absolutos no superan una cifra importante, estando casi todos ellos por debajo de 10 unidades, siendo 1 (que se corresponde con estimar un nuevo elector de más o uno de menos) el valor modal. Las diferencias no son además especialmente grandes en términos relativos, de hecho para cerca del 100% de los casos los errores representan, en términos relativos, menos de un 2% del censo correspondiente. Este comportamiento global de los errores, sin embargo, puede esconder ciertos patrones, que son visibles cuando se analizan los errores condicionando a algunas de las características de la sección, como son la proporción de extranjeros de la sección, el tamaño de la misma o el tamaño del hábitat al cual pertenece.

Aunque en general no existe una relación clara entre el tamaño de los errores relativos cometidos en cada sección censal y la proporción de extranjeros en la misma (ver Cuadro 2), en la provincia de Alicante si se observa una relación significativa y positiva: a mayor proporción de extranjeros mayor error de estimación, tanto cuando se considera el signo del error como cuando este se calcula en módulo (i.e., en valor absoluto). Lo que sin duda indica que tiende a haber una subestimación del número de nuevos electores CER en las secciones censales con mayor proporción de extranjeros. La explicación a este resultado apunta a que se debe al hecho de que, por una parte, la provincia de Alicante es la provincia de España con mayor proporción de extranjeros (20,62%) y a la circunstancia de que, por otra parte, gran parte de los extranjeros afincados en muchos de los municipios costeros de la provincia de Alicante son jubilados, lo que estaría provocando que la hipótesis (i), centrada en los grupos de 15 a 24 años, no sea especialmente afortunada en este caso y estaría artificialmente 'vaciando' más de lo debido de españoles los grupos de edad de 15 a 19 y de 20 a 24 años. En el caso de Cuenca, el resultado es el contrario, los errores son menores (y además se producen por sobreestimación) a mayor proporción de extranjeros. La distribución interna de extranjeros-españoles en cada sección censal unida al hecho de que el porcentaje de extranjeros entre 15 y 24 años en Cuenca es más de 4 puntos porcentuales superior al del total de extranjeros de la provincia podría explicar esta circunstancia.

***Cuadro 2**. Correlaciones entre proporción de extranjeros en la sección censal y errores relativos, respecto a censo, directos y en valor absoluto.*

	Alicante	Valencia	Castellón	Albacete	C. Real	Cuenca	Guadal.	Toledo	Madrid
Errores	0,270**	-0,071**	-0,019	-0,044	0,012	-0,169**	-0,081	-0,106*	-0,041**
|Errores|	0,242**	0,039	-0,034	0,007	0,062	-0,152**	0,080	0,052	0,075**

*Fuente: Elaboración propia. *Significativo a nivel 0.01; *Significativo a nivel 0.05.*

Ahondando más en los resultados destacados en el párrafo anterior, se ha realizado, tras clasificar las secciones censales según su porcentaje de extranjeros en cuatro grupos (0-5%, 5-10%, 10-20% y >20%), un ANOVA de los errores relativos (en valor absoluto), el cual permite concluir que en Cuenca no existen diferencias significativas entre grupos, mientras que en Alicante los errores del grupo de secciones censales con más proporción de extranjeros son significativamente mayores que los del resto de grupos. Unos resultados que vienen a confirmar las impresiones expuestas en el párrafo anterior y que pueden servir de guía para determinar cómo incorporar en el

proceso de estimación otra información disponible todavía no explotada. Cuestión esta que abordamos en la próxima sección.

Para completar el análisis, el Cuadro 3 ofrece información sobre las correlaciones entre los errores relativos en valor absoluto y los tamaños (medidos mediante el total de habitantes) de las secciones censales y de los municipios a los que pertenecen. Como era de esperar, por la ley de los grandes números y por las propias métricas utilizadas (donde los errores se miden en términos relativos), las correlaciones son negativas, indicando que el error relativo decrece a medida que se incrementa el tamaño de la unidad considerada. La excepción a esta regla se observa en Alicante, donde como ya hemos visto otras variables, como la proporción de extranjeros y su distribución en grupos de edad tiene un efecto en los errores significativamente mayor.

***Cuadro 3**. Correlaciones entre los errores relativos, respecto a censo, en valor absoluto y los tamaños de la sección censal y del municipio.*

Tamaño	Alicante	Valencia	Castellón	Albacete	C. Real	Cuenca	Guadal.	Toledo	Madrid
Sección	0,096**	-0,233**	-0,126**	-0,230**	-0,133**	-0,252**	-0,162**	-0,160**	-0,075**
Municipio	-0,022	-0,099*	-0,028	-0,077	-0,111*	-0,158*	-0,127*	-0,058	-0,042**

*Fuente: Elaboración propia. *Significativo a nivel 0.01; *Significativo a nivel 0.05.*

5. CONCLUSIONES

A partir de la explotación estadística del padrón municipal y gracias al INE, la información disponible en el ámbito de la sección censal en España es muy rica y valiosa. En muchas ocasiones, sin embargo, el analista precisa de una información más detallada y/o referida a un instante temporal distinto al empleado por el INE. Un ejemplo paradigmático es el de los procesos electorales, en los que puede resultar de interés disponer del número de nuevos electores que, por edad, se incorporan al censo electoral en cada sección censal en una nueva elección. Se trata de una variable no pública y necesaria para, por ejemplo, la estimación de las matrices de transferencia de voto utilizando técnicas de inferencia ecológica.

El presente trabajo presenta, y valida, una estrategia estadística que permite generar estimaciones de tal variable empleando exclusivamente información publicada. El complejo estimador introducido en esta investigación es evaluado comparando las estimaciones logradas con las cifras reales de nueve conjuntos de datos diferentes. A la luz de los resultados obtenidos, el estimador propuesto supera el escrutinio pues genera estimaciones de gran calidad y que podrían ser empleadas con confianza en

problemas de inferencia ecológica o en modelos de predicción basados en pequeñas áreas.

Un análisis pormenorizado de los errores de estimación, no obstante, muestra que quizás todavía podría existir cierto margen de mejora utilizando una información publicada todavía no introducida en el proceso; aunque eso sí a costa de complicar significativamente el procedimiento. Además de las variables consideradas, hay un par de variables disponibles en el ámbito municipal, no empleadas en la construcción de nuestro estimador, que también podrían aportar cierta información. Se trata de la población para cada municipio clasificada por nacionalidad (español/extranjero) y grandes grupos de edad (menores de 16 años, de 16 a 64 años, y 65 años y más) y de la población de cada municipio por edad (año a año).

Las variables anteriores podrían servir para modificar el estimador en algún sentido con el objetivo de buscar mejores predicciones. La incorporación de cualquiera de estas variables en el procedimiento de estimación conllevaría, no obstante, importantes complejidades desde el punto de vista de computación (y de comunicación del estimador), ya que para poder combinar la información municipal con la de secciones censales sería preciso disponer de valores parciales por municipio obtenidos mediante agregación de grupos de secciones censales.

Entre las variantes que cabría la pena explorar sería modificar el punto (iii) de construcción del estimador. De acuerdo con este punto se admite que “a distribución de españoles por edad dentro de cada grupo quinquenal en cada sección censal coincide con la distribución por edades de españoles en el grupo quinquenal en el conjunto de la provincia”, por lo que una alternativa de modificación podría consistir en incluir por un punto previo (a implementar antes o después del punto (i)) en el que se asumiese que la distribución por edad dentro de cada grupo quinquenal en cada sección censal coincide con la distribución por edades de la población por edad (año a año) del municipio al que pertenece.

Asimismo, otra modificación del proceso que merecería la pena ser explorada consistiría en trasladar las distribuciones de extranjeros/españoles por grandes grupos de edad de cada municipio sobre las distribuciones de extranjeros/españoles de cada una de las secciones censales del municipio, para en una segunda etapa dividir la población con entre 15 y 24 años en cada sección censal entre españoles y extranjeros (paso (ii) de nuestro proceso) a partir de suponer alguna distribución (por ejemplo uniforme) dentro de los grandes grupos y el peso relativo de cada una de ellas en los grupos objetivo. Pensamos que esta última modificación podría ser especialmente valiosa en municipios de no excesivo tamaño, es decir, en municipios que están

constituidos por un número pequeño de secciones censales donde las distribuciones municipales y por secciones censales estarán más próximas.

AGRADECIMIENTOS

Los autores desean agradecer la ayuda del Ministerio de Economía y Competitividad a través del proyecto "Estructura Social, Encuestas y Elecciones", referencia CSO2013-43054-R, correspondiente a la convocatoria 2013 del programa de proyectos de I+D+i del programa estatal de investigación, desarrollo e innovación orientado a los retos de la sociedad.

REFERENCIAS BIBLIOGRÁFICAS

BOE (1985): «Ley Orgánica 5/1985, de 19 de junio, del Régimen Electoral General», *Boletín Oficial del Estado*, 147, de 20 de junio de 1985, 19110-19134.

BOE (1999): «Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal», *Boletín Oficial del Estado*, 298, de 14 de diciembre de 1999, 43088-43099.

BOE (2011): «Ley Orgánica 2/2011, de 28 de enero, por la que se modifica la Ley Orgánica 5/1985, de 19 de junio, del Régimen Electoral General», *Boletín Oficial del Estado*, 25, de 29 de enero de 2011, 9504-9523.

INE (2016): *Estadística del Padrón Continuo*. Instituto Nacional de Estadística. Madrid. URL http://www.ine.es

KING, G. (1997): *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press. Princeton.

OJEU (2013): «Commission Regulation (EU) No 557/2013 of 17 June 2013 implementing Regulation (EC) No 223/2009 of the European Parliament and of the Council on European Statistics as regards access to confidential data for scientific purposes and repealing Commission Regulation (EC) No 831/2002 (1)», *Official Journal of the European Union*, 56, L164/16–L164/19.

PAVÍA, J.M. (2015): «Testing Goodness-of-fit with the Kernel Density Estimator: GoFKernel», *Journal of Statistical Software*, 66 (Code Snippet 1), 1-27.

PAVÍA, J.M.; LARRAZ, B. (2012): «Nonresponse Bias and Superpopulation Models in Electoral Polls», *Revista Española de Investigaciones Sociológicas*, 137, 237-264.

PAVÍA, J.M.; VERES, E. (2016): «Un nuevo estimador para disgregar totales poblacionales. El caso de los nuevos electores», *Anales de Economía Aplicada 2016*, XXX, 817-826.

PAVÍA-MIRALLES, J.M. (2005): «Forecasts from Non-Random Samples: The Election Night Case», *Journal of the American Statistical Association*, 100, 1113-1122.

R CORE TEAM (2016): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. URL http://www.R-project.org/

USC (2002): «Confidential Information Protection and Statistical Efficiency Act of 2002», *Public Law 107-347 "E-Government Act"*. 116 STAT.2962–116 STAT.2970.