serie ec
WP-EC 2015-02

Census *grid* 2011
Una evaluación metodológica

Francisco J. Goerlich e Isidro Cantarino

Ivie

Versión: abril 2015 / Version: April 2015

Edita / Published by:
Instituto Valenciano de Investigaciones Económicas, S.A.
C/ Guardia Civil, 22 esc. 2 1º - 46020 Valencia (Spain)

DOI: http://dx.medra.org/10.12842/WPASEC-2015-02

WP-EC 2015-02

Census *grid* 2011
Una evaluación metodológica*

Francisco J. Goerlich e Isidro Cantarino **

Resumen

Este trabajo presenta una evaluación, desde el punto de vista del usuario, de la malla regular (*grid*) de población, con resolución de 1 km², que el Instituto Nacional de Estadística (INE) ha hecho pública a partir de los resultados del Censo de Población y Viviendas 2011. Esta forma de difusión de resultados resulta muy novedosa y ofrece un gran valor analítico. Por primera vez esta información sobre la distribución espacial de la población se ha generado desde abajo (*Bottom-up*), es decir, a partir del conocimiento de las coordenadas de cada hogar, considerando como tales las del edificio donde reside. La disponibilidad de otra *grid* con idéntica resolución, elaborada por métodos de desagregación espacial a partir de la población censal por unidades administrativas e información auxiliar sobre coberturas del suelo (*Top-down*), nos permite examinar las mejoras asociadas a la geo-referenciación de la población acometida en el contexto de los cambios metodológicos del Censo de 2011. De forma simultánea ello nos permite analizar las bondades de la *grid* censal.

Palabras clave: Rejillas de población, Densidad de población, Geo-referenciación de la población, Censos, *Top-down* versus *Bottom-up*, Demografía.
Clasificación JEL: R12, R14, R52.

Abstract

This paper presents an evaluation, from the user point of view, of the regular population grid, with 1 km² resolution, that the Spanish National Statistical Institute (INE), has released as a product from the last Population and Dwellings Census 2011. This way of disseminating population data is novel, and has a lot of analytical potential uses, since population is no longer linked to administrative divisions. For the first time this information about the population distribution has been generated using a bottom-up approach, this is, by geo-referencing the population at its place of residence. The availability of another grid at the same spatial resolution, but generated using a top-down approach, this is by spatial disaggregation methods from administrative population data and other auxiliary land cover information, allow us to explore the benefits associated to geo-referencing the population in the context of the methodological changes introduced by the Population and Dwellings Census 2011. In parallel, we are able to evaluate the goodness of the census grid.

Keywords: Population grids, Population density, Population Geo-coding, Census, Top-Down versus Bottom-up, Demography.
JEL classification numbers: R12, R14, R52.

---

* Agradecimientos: Este trabajo ha evolucionado a partir de la presentación *Comparing bottom-up and top-down population density grids: The Spanish Census 2011, en el European Forum for Geography and Statistics* (EFGS) Conference 2014, 22-24 October, Krakow, Poland. Los autores agradecen los comentarios de los participantes en dicho foro, y en particular los de Ignacio Duque (INE), Jorge Luis Vega (INE), Carmen Teijeiro (INE) y Matina Halkia (JRC). Agradecemos la disponibilidad a facilitarnos información para la Comunidad de Madrid de Ángel Sanchez (IEM), Dolores Nuñez (IEM) y Rosario Arenas (IEM). Una parte importante de este trabajo no hubiera sido posible sin los datos de la Comunidad de Madrid que amablemente pusieron a nuestra disposición. Francisco J. Goerlich agradece la ayuda del proyecto del Ministerio de Ciencia y Tecnología ECO2011-23248. Resultados mencionados en el texto pero no ofrecidos están disponibles si se solicitan a los autores. Las figuras de este trabajo se aprecian mucho mejor en la versión electrónica en color de este trabajo que en su versión en papel en blanco y negro.

** F.J. Goerlich: Universidad de Valencia e Instituto Valenciano de Investigaciones Económicas (Ivie). I. Cantarino: Universidad Politécnica de Valencia. Autor de correspondencia: F.J. Goerlich, Universidad de Valencia, Departamento de Análisis Económico, e-mail: Francisco.J.Goerlich@uv.es.

1.- Introducción.

El Censo de Población y Viviendas 2011 realizado por Instituto Nacional de Estadística (INE 2011) ha incorporado importantes novedades, tanto desde el punto de vista metodológico, se abandona un censo clásico para emplear una metodología basada en una combinación de registros administrativos y una gran encuesta por muestreo, como desde el punto de vista de la disponibilidad de información (Goerlich, Ruiz, Chorén y Albert 2015). La resolución geográfica para la que el censo ofrece información ha disminuido notablemente, apenas se dispone de información para todos los municipios y la información por secciones censales es prácticamente inexistente con generalidad, ya que la proceder de una muestra está condicionada por el tamaño muestra de la misma.[1] Sin embargo, las nuevas técnicas de los Sistemas de Información Geográfica (GIS) han irrumpido con fuerza, y una de las principales novedades del censo es la georreferenciación exhaustiva de todos los edificios con al menos una vivienda familiar. Esto ha permitido diseñar un sistema de difusión de la información censal al margen de los límites administrativos, ya que se dispone de las coordenadas de cada hogar aproximadas por las del edificio donde reside.[2]

Lamentablemente, las coordenadas de los edificios no han formado parte de la información difundida por el censo. Sin embargo, si ha formado parte del sistema de difusión censal una aproximación al territorio no basada en la división administrativa del estado: Comunidades Autónomas, Provincias, Municipios y Secciones Censales.[3] La directiva comunitaria INSPIRE (Directive 2007/2/EC), diseñada para establecer una Infraestructura para la Información Espacial en el seno de la Comunidad Europea, ha establecido una malla (*grid*) geográfica armonizada a nivel europeo (Annoni 2005, INSPIRE 2014) susceptible de ser utilizada como soporte para difundir información estadística tradicional. La *grid* de referencia estándar tiene una resolución de 1 km^2. A partir de esta *grid* el INE ha incluido, entre sus sistemas de difusión de información censal, determinadas variables en este formato. Merece la pena mencionar que, aunque la *grid* de referencia Europea es de 1 km^2, el sistema de difusión censal en esta forma

[1] Ello hace que la información sea muy heterogénea en función de la muestra existente en cada sección, aunque ciertamente existen casos en los que se ofrece información comparable con el censo de 2001.

[2] La consulta de información para áreas definidas por el usuario es accesible en http://www.ine.es/censos2011/visor/.

[3] El Censo de 2011 no lleva asociado nomenclátor, siendo el primer censo moderno de la historia que presenta esta característica.

alcanza un resolución mucho mayor, pudiendo descender hasta una rejilla de 50 m de lado si la información muestral lo permite.[4] No obstante la información a nivel de usuario sólo se ofrece en la rejilla europea estándar de 1 km^2.

Lamentablemente la información existente en dicho formato es casi tan limitada como la disponible para las Secciones Censales. Si descargamos la información de la página *web* del INE[5] nos encontraremos con que, de los 145 indicadores potencialmente disponibles en dicho formato, solo está completa la *Población total* por celda, para el resto de indicadores siempre hay más o menos celdas en blanco por cuestiones de secreto y/o fiabilidad estadística. Adicionalmente, la población total sólo incluye la residente en viviendas principales, excluyéndose de esta forma la población en viviendas colectivas, y además se ofrece redondeada al 5 a nivel de celda. A pesar de estas limitaciones esta forma de difusión supone un importante paso adelante, más que por la utilidad intrínseca de la información ofrecida, por lo que representa en términos de la utilización de formatos que con seguridad se generalizarán en el futuro.[6]

Este trabajo evalúa, desde el punto de vista metodológico, la *grid* de población derivada del censo 2011. Esta no es la primera *grid* de población disponible para toda España, pero si la primera construida a partir de una geo-referenciación a nivel de coordenada de la población, es decir mediante métodos conocidos como *bottom-up* (EFGS 2014). Sin embargo, la *grid* procedente del INE objeto de evaluación en este trabajo no es la que procede de la *web* del Instituto, sino la que distribuye Eurostat.[7] Dicha *grid* tiene algunas ventajas respecto a la ofrecida directamente por el INE en su *web*. En primer lugar no está redondeada al 5, sino que ofrece cifras de población entera por celda sin ningún tipo de restricción de confidencialidad. En segundo lugar, incluye el total de la población, tanto la residente en viviendas principales como en

[4] El acceso al visor geográfico dispone de un documento metodológico sobre cómo se ha dividido el territorio, http://www.ine.es/censos2011_datos/rejilla_web.pdf.

[5] http://www.ine.es/censos2011_datos/cen11_datos_resultados_rejillas.htm.

[6] Durante el proceso de elaboración de este trabajo el INE (2015) difundió una interesante publicación sobre las características básicas de la sociedad española a partir de este sistema zonal y la información del Censo de 2011. Lamentablemente la información disponible al público es bastante inferior a la necesaria para la realización de dicho trabajo, pero muestra las potencialidades descriptivas de la información en dicho formato. El trabajo muestra, también, los avances cartográficos realizados desde el Censo de 1991, en el que, por primera vez, se puso sobre un mapa la densidad de población de los municipios españoles (INE 1994).

[7] http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography.

establecimientos colectivos, y en consecuencia es consistente con el total de población censal.[8] En lo sucesivo esta *grid* será referenciada como *GEOSTAT2011*.

La estructura del trabajo es la siguiente. A continuación examinamos, brevemente, las *grids* de población disponibles para el conjunto nacional desde una perspectiva histórica, así como sus métodos de producción. Adicionalmente, se genera una *grid* a partir de la población del censo y métodos dasimétricos de desagregación espacial, conocidos generalmente como *top-down*. Dada la obvia superioridad de los métodos *bottom-up*, la sección siguiente examina el origen de los errores cometidos normalmente por los métodos de desagregación que utilizan coberturas del suelo como información auxiliar. Esto permite examinar hasta qué punto dichos métodos pueden mejorar la calidad de la información, y que tipos de errores serán inevitables independientemente de los algoritmos de desagregación empleados. A continuación evaluamos directamente la *grid* del INE frente a un fichero de coordenadas de la población a nivel de punto, ya que toda la información que disponemos del INE es a nivel de celda de 1km^2 de resolución. Una sección final resume las conclusiones del trabajo.

2.- *Census grid* 2011: Un poco de historia.

La primera distribución de la población para España en formato de malla geográfica regular procede del censo de 2001 y fue realizada por el *Joint Research Center* (JRC) (Gallego 2010, Gallego, Batista, Rocha y Mubareka 2011).[9] El objetivo último era disponer de estadísticas de población, a nivel europeo, que no dependieran de límites administrativos, y que fueran almacenadas en un sistema zonal que permitiera la integración con otro tipo de información, fundamentalmente medioambiental. Esta *grid* todavía es actualmente distribuida por la Agencia Europea del Medio Ambiente (EEA).[10]

Dicha *grid* fue elaborada mediante métodos dasimétricos de desagregación espacial (Eicher y Brewer 2001) a partir de las poblaciones municipales del censo de

[8] En realidad dicha *grid* ofrece 127 habitantes más que la cifra de población censal, como consecuencia del redondeo a enteros de la población a nivel de celda, ya que las cifras de población del censo 2011 se ofrecen con decimales, lo que deriva de la metodología de recuento de la población (INE 2012).

[9] Esta pequeña reseña histórica obvia los proyectos de cobertura mundial realizados mediante teledetección y en los que también está incluida España (Bhaduri, Bright, Coleman y Dobson 2002, CIESIN 2005, Bhaduri, Bright, Coleman y Urban 2007).

[10] http://www.eea.europa.eu/data-and-maps/data/population-density-disaggregated-with-corine-land-cover-2000-2.

2001 y *Corine Land Cover* (*CLC*) como información auxiliar. En aquel momento no fue posible evaluar la calidad de los resultados para España, puesto que no existía población geo-referenciada respecto a la que comparar la bondad de los métodos utilizados. Una distribución de la población en este formato representaba una mejora sustancial respecto al cálculo de densidades por municipios tradicional, y que supone implícitamente que todo el territorio que sirve de soporte a la cifra de población está habitado. Sin embargo, y a pesar de dicha mejora, la distribución generada no era tremendamente realista. Para España suponía que, a escala de una cuadrícula de 1 km^2, el 85% del territorio tenía población residente. Claramente dicha dispersión está muy lejos de la realidad, dado el modelo de asentamiento del interior y sur peninsular. El origen del problema procedía fundamentalmente de la baja resolución de la información auxiliar utilizada en el proceso de desagregación: *Corine Land Cover*, la base de datos sobre coberturas del suelo de referencia a escala Europea. Tal y como muestran Goerlich y Cantarino (2012) el modelo de asentamiento de una gran parte de España, junto con la resolución de *CLC* produce que en más de la mitad de los municipios españoles *CLC* no reporte zona urbana. El resultado es que el algoritmo de desagregación dispersa en exceso la población sobre coberturas agrícolas, y más generalmente sobre amplias zonas del territorio que no están habitadas. En consecuencia, la densidad de población en zonas urbanas es infra-estimada y la densidad de población en zonas rurales es sobre-estimada.

Existe amplia evidencia en la literatura sobre desagregación espacial mediante métodos dasimétricos que muestra como la resolución de la información auxiliar utilizada en el proceso de desagregación es mucho más importante que la fineza de los algoritmos (Martin, Tate y Langford 2000), y claramente la resolución de *CLC* aplicada sobre datos municipales era insuficiente para producir resultados satisfactorios.

Eurostat se marcó como objetivo disponer de una *grid* de población a escala Europea con referencia 2006. Dicha malla debería mejorar la de 2001, por lo que, para aquellos países que no disponían de población geo-referenciada a nivel de coordenada los esfuerzos se dirigieron a aumentar la resolución de la información auxiliar. A nivel Europeo se hicieron ensayos con la capa de sellado del suelo,[11] 1ha de resolución,

[11] Distribuida por *Copernicus*, *The European Earth Observation Programme*, dependiente de la Comisión Europea y de la EEA: http://land.copernicus.eu/pan-european/high-resolution-layers/imperviousness/view

(Steinocher 2011a, 2011b), así como intentos de mejorar *CLC* con información adicional de mayor resolución (Batista e Silva 2011, Batista e Silva, Lavalle, y Koomen 2013). Afortunadamente, para España se hizo público el *Sistema de Información de Ocupación del Suelo de España* (*SIOSE*, IGN2011) con fecha de referencia 2005, que además de presentar una resolución mucho mayor que *CLC*, desarrollaba un modelo de datos orientado a objetos tremendamente versátil y con mucha mayor información que el simple listado de coberturas jerárquico como el implantado en *CLC*.

A partir de la población a nivel de sección censal procedente del Padrón, lo que supone una importante mejora de resolución especialmente en el ámbito urbano, y aprovechando todo lo posible el modelo de datos de *SIOSE2005*, Goerlich y Cantarino (2011, 2012, 2013) elaboraron una *grid* de población para España que fue incorporada a la *grid* Europea distribuida por Eurostat, conocida como *GEOSTAT2006*. Esta *grid* fue validada frente al Padrón geo-referenciado de la Comunidad de Madrid, con un error relativo del 4%, muy reducido frente a los errores reportados por Gallego (2010) y Gallego, Batista, Rocha y Mubareka (2011) para los países en los que podía efectuarse esta validación en 2001, y que se situaba entre el 17% y el 35%.

Los algoritmos de desagregación aplicados resultaron ser muy similares a los utilizados anteriormente para la elaboración de la *grid* de 2001, por lo que las mejoras logradas pueden atribuirse a la mayor resolución de la información poblacional de partida, secciones censales *versus* municipios, así como a la de la información auxiliar sobre coberturas del suelo, que en el caso de *SIOSE* incluso contiene información sobre el tipo de edificación. El resultado fue un 19% de celdas de $1km^2$ con población residente, una cifra notablemente inferior a la obtenida en 2001. El cuadro 1 compara la distribución relativa de frecuencias de ambas *grids* y permite observar a simple vista sus diferencias, en la de 2001 un 63% de las celdas tienen menos de 20 habitantes, en la de 2006 ese porcentaje no llega a la mitad, 29%. Lo contrario sucede en el otro extremo de la distribución. Las celdas que tienen al menos 500 habitantes en 2001 apenas llegan al 2%, pero superan el 11% en 2006. Estas diferencias son todavía más acusadas en los extremos. En definitiva, la distribución de la población que se deriva de la *grid* de 2006 es notablemente más concentrada que la que mostraba su predecesora, y este efecto se debe, en exclusividad, a la mayor resolución de la información empleada en su elaboración.

La *grid* de 2006 representaba un objetivo intermedio hasta la realización del censo de 2011, el primero elaborado bajo reglamentación Europea. De dicho censo se obtendría la tercera *grid* Europea, y a ser posible debía generarse mediante métodos de geo-referenciación de la población a nivel de coordenada puntual, por lo que una vez disponible la base de datos de coordenadas y población asignada a las mismas, la obtención de la *grid* debería ser inmediata. La georreferenciación de todos los edificios con alguna vivienda proporciona el marco ideal para que el INE generara una *grid* de población *bottom-up* a partir del censo de 2011, lo que significa un salto cualitativo metodológico importante frente a sus dos predecesoras anteriores, 2001 y 2006, generadas ambas por métodos *top-down*.

El cuadro 1 muestra las características principales de la *grid* producida por el INE a partir del censo y distribuida por *Eurostat*, *GEOSTAT2011*.[12] La superficie habitada, a 1 km^2 de resolución, apenas supera el 12% del territorio nacional, lo que supone 1/3 menos de celdas habitadas respecto a la de 2006.[13] La distribución de frecuencias muestra de nuevo las importantes diferencias; si bien estas no son notables en el centro de la distribución, se acentúan en los extremos: el porcentaje de celdas con menos de 20 habitantes apenas supera el 21%, mientras que las celdas con más de 500 habitantes casi alcanzan el 19%. *GEOSTAT2011* muestra, pues, una concentración de la población realmente elevada, no solo el porcentaje de celdas habitadas es notablemente menor que en *GEOSTAT2006*, sino que dichas celdas presentan una elevada concentración con un número realmente escaso de celdas con menos de 5 habitantes (4.4%), algo menos de la mitad que en 2006. Las diferencias son, en cualquier caso, mucho menores que entre las *grids* de 2001 y 2006.

Si las diferencias entre las distribuciones de 2001 y 2006 pueden atribuirse con claridad a cuestiones relacionadas con la resolución de la información de base, ya que el método estadístico de distribución de la población es muy similar en ambos casos. Las

[12] El cuadro 1 también muestra, en su última fila, los estadísticos básicos para la *grid* distribuida por el INE en su *web*, lo que deja patente porque dicha *grid* no puede utilizarse para un ejercicio de esta naturaleza. No sólo esta *grid* no recoge el total de la población, excluyendo la residente en viviendas colectivas, sino que además el proceso de redondeo al 5 sesga los resultados hacia una mayor concentración de la población: ¡No se muestran celdas con población inferior a los 5 habitantes, aunque si existan, ya que los valores inferiores a 2.5 se redondean a 0!

[13] Una comparación con el resto de países Europeos muestra que España es el segundo país en términos de concentración de la población, sólo por debajo de Islandia, con un 2% de celdas habitadas, pero por delante de los nórdicos como Noruega (17%), Suecia (25%) o Finlandia (30%), países en los que su *grid* también ha sido generada mediante procedimientos *bottom-up*.

diferencias entre *GEOSTAT2006* y *GEOSTAT2011* pueden deberse tanto el método de producción, *top-down versus bottom-up*, como a cambios en la concentración real de la población. El periodo inter-censal 2001-2011 resulta ser el de mayor crecimiento poblacional de la historia, casi 6 millones de personas, en gran parte debido a la inmigración; y este stock de nuevos efectivos no se ha distribuido en modo alguno de forma uniforme a lo largo del territorio (Goerlich, Ruiz, Chorén y Albert 2015). Para aislar el efecto de los diferentes métodos de producción del lapso temporal entre ambas rejillas de población Goerlich y Cantarino (2014) aplicaron la misma metodología empleada en la elaboración de *GEOSTAT2006* a la población por secciones censales del censo 2011 y *SIOSE2009* como información auxiliar en el proceso de desagregación. El resultado resumido de dicho ejercicio se muestra en la fila 4 del cuadro 1. Observamos cómo dicha actualización no afecta prácticamente en nada a la distribución de la población mostrada en *GEOSTAT2006*. El mensaje principal de dicho ejercicio resulta pues bastante claro: las diferencias entre la distribución de la población mostradas en *GEOSTAT2006* y *GEOSTAT2011* se deben, en su práctica totalidad, a los métodos de producción, y no a cambios apreciables en la distribución de la población a la escala de referencia, celdas de 1 km^2.

Cuadro 1. Resumen de las principales características de las *grids* de población para España: 2001, 2006 y 2011.

(a) Información sobre celdas habitadas

Nº	Fecha de Referencia	Origen de la Población	Promotor de la grid	Productor de la grid	Celdas habitadas		Distribución de frecuencias por tamaño de población en las celdas					
					Absolutas	%	< 5	5 - 19	20 - 199	200 - 499	500 - 4999	>= 5000
1	2001	Censo	EEA	JRC	434,738	85.0%	124,295	147,886	146,847	5,632	8,577	1,501
							28.6%	34.0%	33.8%	1.3%	2.0%	0.3%
2	2006	Padrón	Eurostat	UVEG/UPV	94,916	18.6%	8,823	19,124	46,047	9,793	9,258	1,871
							9.3%	20.1%	48.5%	10.3%	9.8%	2.0%
3	2011	Censo	Eurostat	INE	63,522	12.4%	2,800	10,881	30,036	7,893	9,740	2,172
							4.4%	17.1%	47.3%	12.4%	15.3%	3.4%
4	2011	Censo	Goerlich y Cantarino (2014)		95,169	18.6%	9,411	20,407	44,131	9,414	9,750	2,056
							9.9%	21.4%	46.4%	9.9%	10.2%	2.2%
5	2011	Censo	INE	INE	62,440	12.2%	0	11,484	31,107	7,945	9,750	2,154
							0.0%	18.4%	49.8%	12.7%	15.6%	3.4%

(b) Información sobre población en la grid

Nº	Fecha de Referencia	Origen de la Población	Promotor de la grid	Productor de la grid	Población en la *grid*		Distribución de frecuencias: Población según la distribución de las celdas					
					Absoluta	%	< 5	5 - 19	20 - 199	200 - 499	500 - 4999	>= 5000
1	2001	Censo	EEA	JRC	40,874,775	100.1%	270,194	1,602,814	7,292,006	1,783,650	13,344,865	16,581,246
							0.7%	3.9%	17.8%	4.4%	32.6%	40.6%
2	2006	Padrón	Eurostat	UVEG/UPV	44,708,964	100.0%	20,229	215,917	3,323,510	3,068,789	13,954,561	24,125,958
							0.0%	0.5%	7.4%	6.9%	31.2%	54.0%
3	2011	Censo	Eurostat	INE	46,816,043	100.0%	7,912	123,455	2,250,465	2,500,596	15,410,678	26,522,937
							0.0%	0.3%	4.8%	5.3%	32.9%	56.7%
4	2011	Censo	Goerlich y Cantarino (2014)		46,689,142	99.7%	21,674	229,071	3,147,629	2,961,097	14,953,164	25,376,507
							0.0%	0.5%	6.7%	6.3%	32.0%	54.4%
5	2011	Censo	INE	INE	46,574,735	99.5%	0	108,715	2,252,000	2,500,505	15,416,830	26,296,685
							0.0%	0.2%	4.8%	5.4%	33.1%	56.5%

Nota: Los estadísticos de la *grid* del JRC de 2001 se han obtenido de la extracción de la capa Europea, sin ajuste en la población de las celdas frontera.

EEA: Agencia Europea del Medio Ambiente

JRC: Joint Research Center

UVEG/UPV: Universidat de Valencia Estudi General / Universidad Politécnica de Valencia. Goerlich y Cantarino (2012, 2013)

INE: Instituto Nacional de Estadística.

Goerlich y Cantarino (2014): *Comparing bottom-up and top-down population density grids: The Spanish Census 2011* , European Forum for Geography and Statistics Conference (EFGS), Krakow (Polonia). 22 – 24 October, 2014.

3.- *Top-down* versus *bottom-up*: ¿Dónde fracasan los métodos de desagregación espacial?

Esta sección compara *GEOSTAT2011* con la *grid top-down* generada por Goerlich y Cantarino (2014) a partir de la población del censo 2011 y *SIOSE2009*. El ejercicio toma como 'verdadera distribución de la población', a la escala de análisis, *GEOSTAT2011*, puesto que los métodos *bottom-up* son claramente superiores a cualquier método estadístico de desagregación espacial, que por su propia naturaleza sólo puede aspirar a aproximar la realidad. En consecuencia nos preguntamos dónde fracasan los métodos *top-down*. El análisis es relevante porque no siempre es posible disponer de una *grid* construida a partir de población geo-referenciada a nivel de coordenada puntual, y en consecuencia ejercicios de desagregación continuarán siendo inevitables en el futuro, quizá para características de la población u otras variables, pero el tipo de errores cometidos por los algoritmos dasimétricos de re-escalado espacial son de naturaleza similar en la mayoría de los casos. El supuesto de que *GEOSTAT2011* representa la 'verdadera distribución de la población' será cuestionado en el apartado siguiente.

El estadístico estándar para evaluar las discrepancias respecto a una *grid* de referencia es ½ del error relativo total, que se encuentra comprendido entre 0 y 1:[14]

$$\delta = \frac{\Sigma_{i=1}^{n} \left| P_i - P_i^{ref} \right|}{2 \times P^{ref}} \tag{1}$$

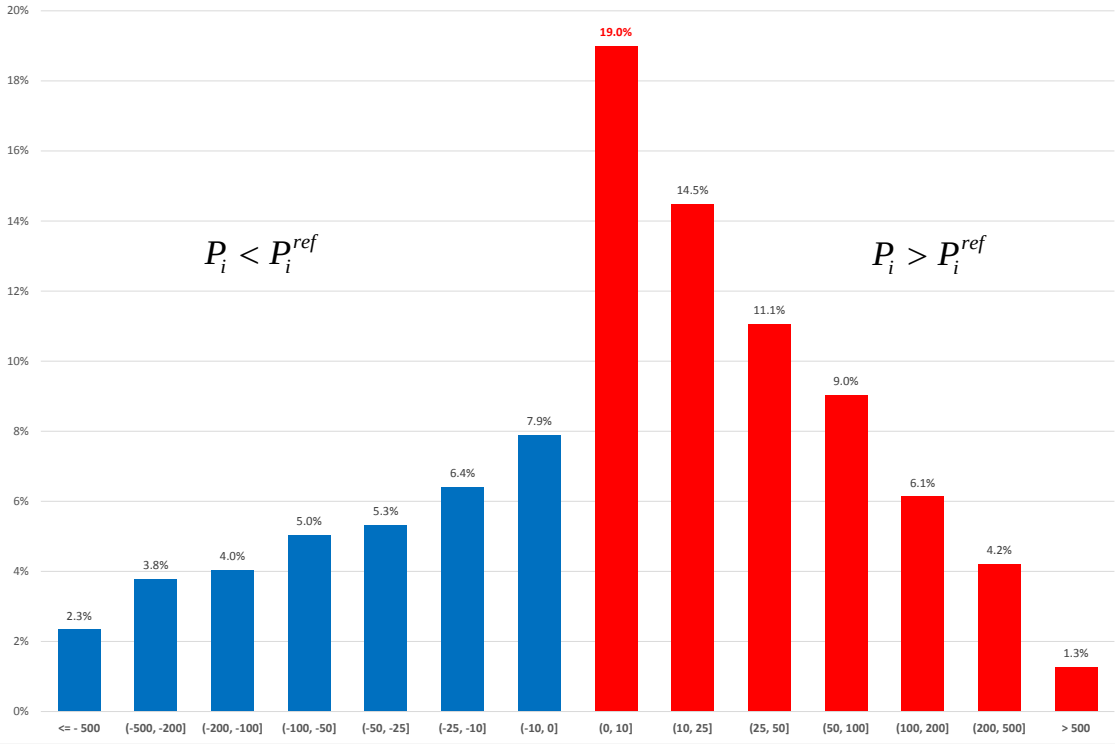
donde P_i^{ref} es la población de la *grid* de referencia, *GEOSTAT2011*, P_i la de la *grid* objeto de comparación generada mediante desagregación, *n* son el número de celdas de la malla, y $P^{ref} = \Sigma_{i=1}^{n} P_i^{ref}$. Dicho estadístico alcanza un valor del 10.1% en nuestro caso, en el entorno del doble de lo obtenido para la Comunidad de Madrid en la elaboración de *GEOSTAT2006*.

[14] Dicho estadístico es extremadamente simple, ya que sólo compara la población en cada celda sin tener en cuenta la localización espacial de los errores. En un mundo donde la geografía importa, lo lógico sería incluir un esquema de ponderación espacial en los errores, ya que no es lo mismo errores de asignación de población en celdas contiguas que en el otro lado de la rejilla, en consecuencia errores cercanos deberían pesar menos que errores lejanos. Por cuestiones de comparabilidad con la literatura estándar no introduciremos este esquema y el análisis se mantendrá a nivel de (1).

El coeficiente de correlación, calculado entre las poblaciones a nivel de celda, es de 0.99 entre ambas *grids*. Esta elevada correlación, unido a un error relativo moderado, y la importante discrepancia entre el número de celdas habitadas generada por ambos métodos de producción (cuadro 1), tiende a indicar que las discrepancias se concentran en un número relativamente grande de celdas, pero que afectan a poca población.

La figura 1 ofrece una primera comprobación de esta intuición. En ella mostramos el histograma de discrepancias por estratos de población distinguiendo los casos en los que $P_i > P_i^{ref}$, a la parte derecha, de los casos en los que $P_i < P_i^{ref}$, a la parte izquierda de la figura. El 20% de los errores se concentran en casos en los que la discrepancia no supera los 10 habitantes y $P_i > P_i^{ref}$. Casi la mitad de los errores (45%) se encuentran dentro de esta tipología, $P_i > P_i^{ref}$, y la población afectada no supera los 50 habitantes por celda.

Figura 1 - Dicrepancias a nivel de celda: estratos de población - *Top-Down versus* GEOSTAT 2011



Fuente: Eurostat para GEOSTAT2011 y elaboración propia (Goerlich y Cantarino 2014).

Para un análisis detallado de errores conviene clasificar las celdas de la *grid top-down* en tres grupos: (*i*) 'falsos positivos', cuando dicha *grid* asigna población pero la de referencia no, (*ii*) 'falsos negativos', cuando dicha *grid* no asigna población a la celda en cuestión, pero la de referencia si, y (*iii*) 'celdas correctas', en el sentido de que

presentan población en ambas *grids*, aunque su magnitud no tiene necesariamente porque coincidir. El error relativo (1) se descompone de forma nítida en estos tres componentes:

$$\delta = \underbrace{\frac{\sum_{P_i>0 \,\&\, P_i^{ref}=0} |P_i - 0|}{2 \times P^{ref}}}_{Falso\ \ Positivo} + \underbrace{\frac{\sum_{P_i>0 \,\&\, P_i^{ref}>0} |P_i - P_i^{ref}|}{2 \times P^{ref}}}_{Celdas\ \ Correctas} + \underbrace{\frac{\sum_{P_i=0 \,\&\, P_i^{ref}>0} |0 - P_i^{ref}|}{2 \times P^{ref}}}_{Falso\ \ Negativo} \qquad (2)$$

Lo que permite examinar no solo las celdas discrepantes, sino también la magnitud de las discrepancias en términos de población. El cuadro 2 ofrece esta descomposición y confirma las intuiciones anteriores: aunque en términos de celdas afectadas los errores cometidos por nuestra desagregación son elevados, un 42% de las celdas de la *grid top-down* son 'falsos positivos', y un 12% de las celdas de *GEOSTAT2011* son 'falsos negativos' –la desagregación dasimétrica no asigna población cuando en realidad si la hay–, la población afectada es poco relevante. Tan solo un 3% de la población es asignada a celdas realmente deshabitadas, mientras que un porcentaje inferior al 1% no se asigna correctamente a celdas con población según *GEOSTAT2011*.

Cuadro 2. Descomposición del error relativo en 3 componentes: Falsos positivos *versus* falsos negativos. Comparación *GEOSTAT2011 versus grid top-down*

$$\underbrace{\frac{\sum_{P_i>0 \,\&\, P_i^{ref}=0} |P_i - 0|}{2 \times P^{ref}}}_{Falso\ \ Positivo} \quad \underbrace{\frac{\sum_{P_i>0 \,\&\, P_i^{ref}>0} |P_i - P_i^{ref}|}{2 \times P^{ref}}}_{Celdas\ \ Correctas} \quad \underbrace{\frac{\sum_{P_i=0 \,\&\, P_i^{ref}>0} |0 - P_i^{ref}|}{2 \times P^{ref}}}_{Falso\ \ Negativo}$$

10.1% = **1.6%** + **8.1%** + **0.4%**

Población afectada por los diferentes tipos de errores

1,517,388 — 3.2% de la población en la grid

371,044 — 0.8% de la población en la grid

Celdas afectadas por los diferentes tipos de errores

	39,523	55,646	7,876
	95,169		63,522
% de celdas habitadas	18.6%		12.4%

Fuente: Eurostat para GEOSTAT2011 y elaboración propia (Goerlich y Cantarino 2014).

El resultado es, como ya se mencionó al principio, que los métodos de desagregación espacial tienen a dispersar en exceso la población, esto se manifiesta en

un número relativamente elevado de celdas que no están realmente habitadas –'falsos positivos'–, aunque no afecta a volúmenes de población importantes. La existencia del error contrario, no asignar población a celdas que realmente la tienen –'falsos negativos'–, es mucho menos frecuente, pero no despreciable, y su origen debe ser analizado.

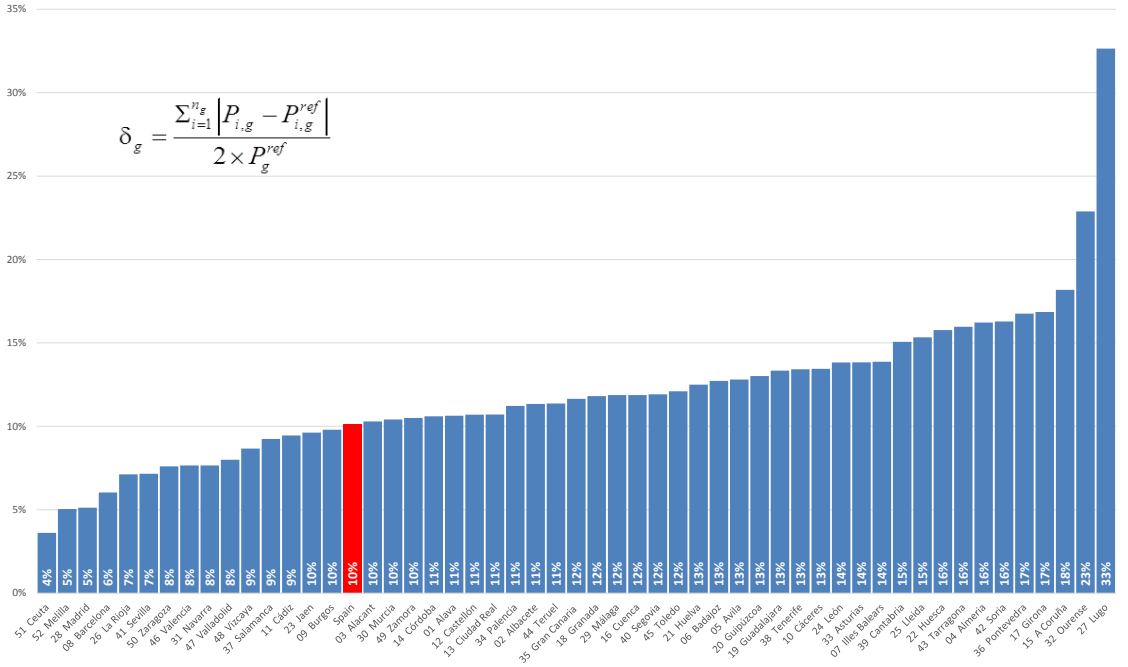
Resulta de interés examinar si existe un patrón de discrepancias a nivel provincial. Dada una partición del territorio en *G* regiones, exhaustivas y mutuamente excluyentes del territorio nacional, el error relativo se descompone como una suma ponderada de los errores de cometidos en las diferentes regiones:

$$\delta = \frac{\Sigma_{i=1}^{n} \left| P_i - P_i^{ref} \right|}{2 \times P^{ref}} = \sum_{g=1}^{G} \frac{P_g^{ref}}{P^{ref}} \cdot \frac{\Sigma_{i=1}^{n_g} \left| P_{i,g} - P_{i,g}^{ref} \right|}{2 \times P_g^{ref}} = \sum_{g=1}^{G} \frac{P_g^{ref}}{P^{ref}} . \delta_g \qquad (3)$$

El análisis de los errores a nivel provincial, $G = 52$, reveló la existencia de un patrón espacial muy marcado asociado a los diferentes tipos de asentamiento dentro del conjunto nacional. Como revela la figura 2, exceptuando Ceuta y Melilla, los errores oscilan entre un 5% para Madrid y un 33% para Lugo. Sin embargo lo más interesante de dicha figura es que los errores son claramente menores en provincias con una población muy concentrada y asentada sobre núcleos compactos, como Madrid, Barcelona, Sevilla, Zaragoza o Valencia. Mientras que los errores son muy elevados en provincias con población dispersa, entre las 5 provincias con los errores más elevados se encuentran las 4 gallegas.

Lo que la figura 2 muestra es que el error cometido en los procesos de desagregación depende no sólo de la resolución de la información de partida, sino también del patrón de asentamiento de la población, o alternativamente que dicho patrón afecta a la calidad de la información auxiliar dado un grado de resolución de la misma.

Figura 2 - Error relativo a nivel provincial: *Top-Down versus* GEOSTAT 2011



Fuente: Eurostat para GEOSTAT2011 y elaboración propia (Goerlich y Cantarino 2014).

Un análisis detallado, en muchos casos mediante inspección visual sobre ortofotos, de los falsos positivos y falsos negativos mostró que los métodos de desagregación fracasan fundamentalmente en 3 direcciones, algunas de ellas no son responsabilidad del algoritmo de desagregación, sino de la calidad –o la resolución– en la información de partida o de decisiones del propio investigador:[15]

1. **Errores de clasificación** en la información auxiliar sobre coberturas del suelo (*SIOSE*).

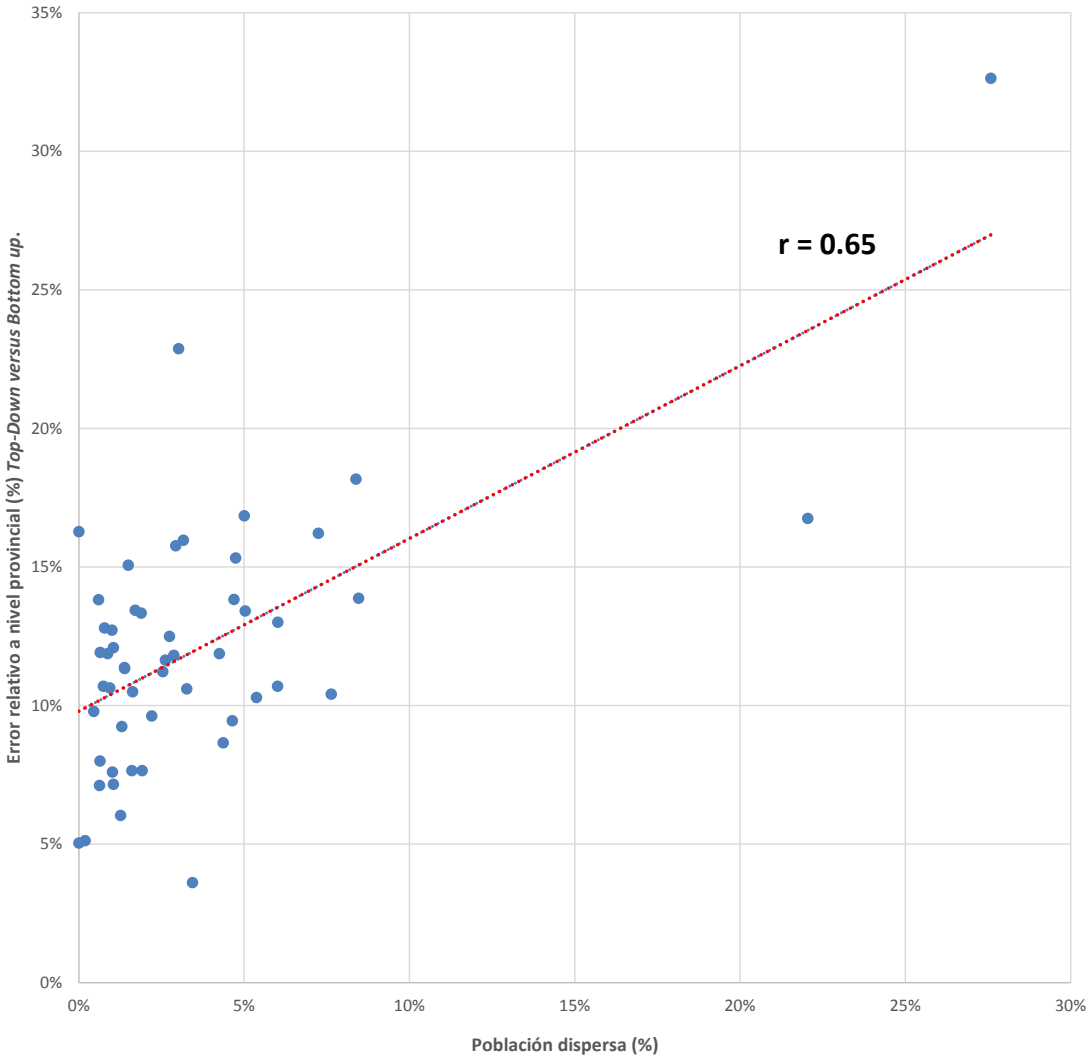
    1.1. Polígonos clasificados como residenciales, cuando en realidad no lo son, y a los que el algoritmo de desagregación acaba atribuyendo población: Falsos positivos.

[15] No incluimos en este listado desajustes geométricos entre las capas de la información auxiliar, *SIOSE* en nuestro ejercicio, y el fichero vectorial de contornos administrativos de la población objeto de desagregación, las secciones censales en nuestro caso. Esta última información se ha revelado extremadamente difícil de conseguir a nivel nacional con la precisión que sería deseable. En particular, dicho fichero existe como producto derivado del censo 2011 en la *web* del INE para su descarga, http://www.ine.es/censos2011_datos/cen11_datos_resultados_seccen.htm, sin embargo resulta curioso observar como las secciones censales de la tabla alpha-numérica del censo no coinciden con las del fichero de contornos de su representación cartográfica, donde existen 43 secciones censales más que para las que disponemos de población a partir del censo. Adicionalmente existen 17 secciones censales en la información censal con población nula, lo que sin duda se debe a errores de muestreo derivados del diseño muestral, o a problemas de falta de respuesta, ya que la población de las secciones censales no procede del Fichero Precensal, e incluye sólo la población residente en viviendas principales (Goerlich, Ruiz, Chorén y Albert 2015).

1.2. Polígonos clasificados como no residenciales, cuando en realidad si lo son, y a los que el algoritmo de desagregación no atribuye población: Falsos negativos.[16]

Ambos tipos de errores de asignación de la población son más frecuentes cuando la población es dispersa, que cuando está más concentrada sobre el territorio, tal y como muestra la figura 3 a nivel provincial.

Figura 3 - Error relativo *versus* dispersión en la población a nivel provincial



Fuente: Eurostat para GEOSTAT2011, INE-Nomenclátor 2012 y elaboración propia (Goerlich y Cantarino 2014).

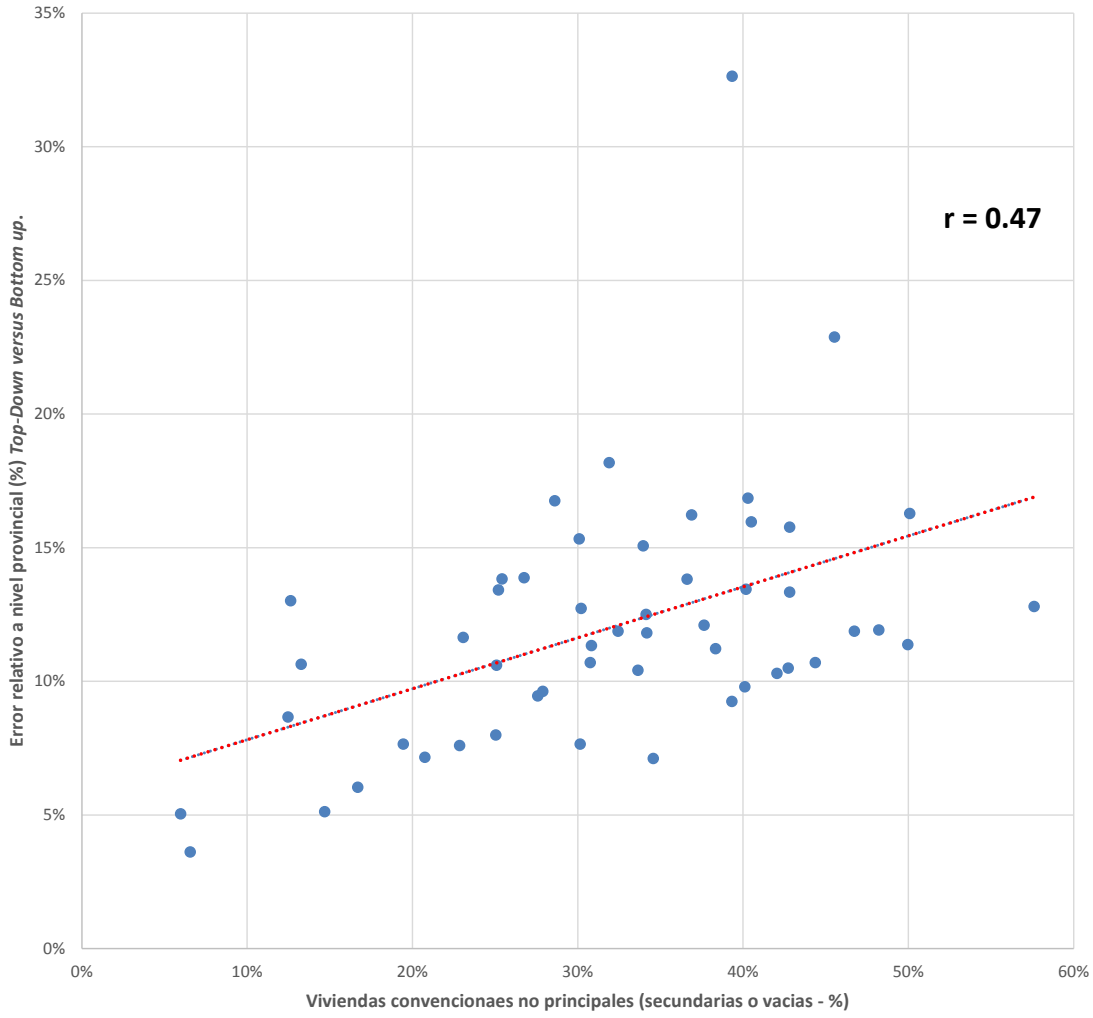
[16] Dentro de este grupo pueden encontrarse, tanto errores de clasificación del propio *SIOSE*, como casos en los que coberturas compuestas agrícolas o forestales, no recojan edificaciones de tipo residencial por falta de resolución, y que en la práctica si alberguen población residente. Este segundo caso no es, propiamente dicho, un error de clasificación, sino que es debido a la resolución de la información de partida. En ambos casos el algoritmo de desagregación no encuentra soporte donde asignar la población, lo que acaba generando un falso negativo. En algunas situaciones, si el polígono clasificado incorrectamente es suficientemente grande, la restricción de volumen a nivel de sección censal hace que la población de la misma se pierda en la *grid* resultante.

2. **Viviendas convencionales no principales (secundarias o vacías)**: Inevitable cuando la información auxiliar en el proceso de desagregación procede de las coberturas del suelo, que no incluye información sobre usos a nivel de vivienda, tal y como es nuestro caso. Por su propia naturaleza ello genera falsos positivos, ya que el algoritmo de desagregación asigna población a viviendas no ocupadas con carácter residencial.

   Este factor puede explicar, en gran parte, el exceso en el número de celdas habitadas que arroja la *grid top-down* frente a *GEOSTAT2011*, cuadro 2, ya que de acuerdo con la información del censo 2011 el 28% de las viviendas convencionales son viviendas secundarias o vacías, y por tanto no albergan población residente. La figura 4 muestra, a nivel provincial, una clara relación positiva entre el número de viviendas no principales y el error cometido en el proceso de desagregación.

   Ambos factores considerados conjuntamente, errores de clasificación y viviendas convencionales no principales (secundarias o vacías), muestran un coeficiente de correlación múltiple con el error relativo a nivel provincial de 0.80.

Figura 4 - Error relativo *versus* viviendas secundarias o vacías a nivel provincial



Fuente: Eurostat para GEOSTAT2011, INE-Censo 2011 y elaboración propia (Goerlich y Cantarino 2014).

3. **Decisiones sobre ‘donde vive la población’**. Cualquier ejercicio dasimétrico de desagregación espacial requiere de una decisión a priori sobre que coberturas van a soportar población y cuáles no. Esta selección condiciona, de partida, los resultados finales y puede generar tanto falsos positivos, al seleccionar coberturas que pueden albergar población, pero que finalmente no la tienen, como –más frecuentemente– falsos negativos, al impedir asignar población a coberturas que realmente si la tienen.

   Esta es una decisión difícil, en la que el investigador tiene que alcanzar un equilibro entre ambos tipos de errores, normalmente con información muy limitada al respecto.

La figura 5 muestra un caso concreto, y real, de esta situación, de trata del Monasterio de Santa María del Paular, en la provincia de Madrid. *SIOSE* clasifica el polígono –contorno azul en la figura 5– como complejo religioso –cobertura compuesta *ERG*–, nuestra selección de coberturas al generar la *grid top-down* no permite población en este tipo de coberturas, y en consecuencia no asignamos población a una celda que en realidad si la tiene. Generamos de esta forma un falso negativo en nuestra *grid*. Una comparación con la población a nivel de coordenada puntual muestra que en dicho lugar residen 9 personas –punto amarillo en la figura 5–, que deberían ser adecuadamente tenidas en cuenta por una *grid bottom-up*, que asignaría correctamente esa población a la celda correspondiente.

Figura 5 – Falso negativo debido a la selección de coberturas

Fuente: SIOSE 2009, IEM-Padrón 2012 y elaboración propia.

Sin embargo, si permitiéramos que las coberturas de tipo complejo religioso albergaran población con generalidad entonces la mayor parte de ellas resultarían habitadas, cuando en la práctica no lo están. El mismo problema lo encontramos con la población residente en complejos industriales; permitir con generalidad población en estas coberturas dispersaría en exceso la población, ya que acabaría poblando, en mayor o menor grado, la práctica totalidad de los polígonos industriales del territorio objeto de análisis. Por otra parte, los métodos de desagregación ya tienden a dispersar la población más de lo que debieran, como muestra el hecho de que los falsos positivos son mayores que los falsos negativos –cuadro 2–.

Así pues, el investigador debe realizar una elección difícil en un contexto de incertidumbre, y en el que reglas generales no son probablemente de aplicación. Incluso los denominados métodos dasimétricos 'inteligentes' (Mennis y Hultgren 2006), basados en el muestreo empírico para determinar pesos variables por tipo de cobertura para la redistribución de la población, no están exentos de este tipo de error. Los pesos determinados en una parte del territorio pueden no ser de aplicación en otra. El análisis anterior muestra claramente como el patrón de asentamiento poblacional influye de forma notable sobre la magnitud del error en la desagregación.

Una cuantificación de cada una de estas fuentes de error sobre el error total en la generación de la *grid top-down* frente a *GEOSTAT2011*, (1), es tremendamente difícil, entre otras cosas porque dichos errores están muy correlacionados, y a su vez son dependientes de la resolución de la información de partida.

4.- Una evaluación directa de *GEOSTAT2011*.

La sección anterior compara la desagregación espacial de la población del censo 2011 a un formato de malla regular de 1 km^2 de resolución, utilizando *SIOSE2009* como información auxiliar y las secciones censales como unidad estadística para la población, con la *grid GEOSTAT2011*, disponible a la misma escala y ofrecida directamente por *Eurostat* a partir de la georreferenciación de la población a nivel de coordenada puntual proveniente del censo 2011.

Dichas coordenadas, que son las que soportan la *grid*, o las de los edificios con alguna vivienda familiar fruto de la operación censal, no son información pública. En consecuencia, a nivel nacional la comparación que hemos efectuado en la sección

anterior es todo lo que podemos realizar. Y es suficientemente ilustrativa de que los métodos dasimétricos de desagregación espacial tienden a dispersar la población más de lo que debieran, incluso cuando la información de partida es de muy elevada resolución.

Sin embargo, al igual que para la *grid GEOSTAT2006* (Goerlich y Cantarino 2012) disponemos del Padrón 2012 para la Comunidad de Madrid georreferenciado a nivel de coordenada puntal. Dicho fichero contiene las coordenadas de las Aproximaciones Postales Principales (APP) del 99.4% de la población del Padrón, y a partir de él es directo obtener una *grid* a la resolución que se desee. Para celdas de 1 km^2 dicha *grid* contiene 2.449 celdas habitadas, un 30% del total. Podemos suponer que el error de dicha *grid* es prácticamente nulo, y representa la verdadera distribución de la población a la escala considerada.

Si comparamos la *grid top-down* para Madrid con la generada por agregación de coordenadas –*bottom-up*– el error relativo es del 6%, sólo ligeramente superior al que obteníamos para *GEOSTAT2006* (Goerlich y Cantarino 2012), y muy similar al que se obtiene para esa provincia cuando *GEOSTAT2011* se toma como *grid* de referencia (5%) –figura 2–. Por su estructura de asentamiento urbano, Madrid es la provincia que muestra menor error entre los métodos *top-down* y *bottom-up*.

La descomposición de dicho error, entre 'falsos positivos' y 'falsos negativos', así como las celdas implicadas en dichos errores y la población afectada por los mismos, se muestra en el cuadro 3. El patrón de errores es muy similar a lo descrito en la sección anterior. La *grid* generada por desagregación muestra un mayor número de celdas habitadas que la obtenida a partir de las coordenadas a nivel de punto, 32% frente a 29%, aunque las celdas con 'falso positivo' están muy poco habitadas, por lo que la población afectada es muy escasa. En términos de población, los 'falsos negativos' son prácticamente inexistentes, de forma que los errores de los métodos de desagregación se manifiestan en una mayor dispersión de la que observamos en la realidad.

Un análisis a nivel municipal permitió confirmar que las discrepancias están asociadas a la dispersión de la población y al volumen de viviendas no principales, secundarias o vacías. Aunque, más allá de estas características generales, un patrón definido de los errores es difícil de encontrar.

Cuadro 3. Descomposición del error relativo en 3 componentes: Falsos positivos *versus* falsos negativos. Comparación *grid bottom-up* Madrid *versus grid top-down*

$$\underbrace{\frac{\sum_{P_i>0 \,\&\, P_i^{M,ref}=0}\left|P_i-0\right|}{2\times P^{M,ref}}}_{Falso\ \ Positivo} \qquad \underbrace{\frac{\sum_{P_i>0 \,\&\, P_i^{M,ref}>0}\left|P_i-P_i^{M,ref}\right|}{2\times P^{M,ref}}}_{Celdas\ \ Correctas} \qquad \underbrace{\frac{\sum_{P_i=0 \,\&\, P_i^{M,ref}>0}\left|0-P_i^{M,ref}\right|}{2\times P^{M,ref}}}_{Falso\ \ Negativo}$$

6.0% = **0.3%** + **5.7%** + **0.0%**

Población afectada por los diferentes tipos de errores

36,952		5,026
0.6% de la población en la grid		0.1% de la población en la grid

Celdas afectadas por los diferentes tipos de errores

598	2,141	308
	2,739	2,449
% de celdas habitadas	32.3%	28.9%

Fuente: IEM-Padrón 2012 georeferenciado y elaboración propia (Goerlich y Cantarino 2014).

Sin embargo, la disponibilidad de la población a nivel de coordenada para la Comunidad de Madrid nos permite comparar dos *grids bottom-up*, *GEOSTAT2011* y la construida directamente por nosotros para Madrid. Ambas deberían ser prácticamente idénticas, ya que el desfase temporal entre la fecha de referencia es de sólo dos meses y las poblaciones en ambas *grids* difieren en sólo un 0.2%.

Sin embargo, las celdas habitadas para Madrid que reporta *GEOSTAT2011* son 1,948, un 23%, lo que en términos absolutos supone 501 celdas habitadas menos, y 7 puntos porcentuales de diferencia. *GEOSTAT2011* parece concentrar la población más de lo que encontramos en la realidad, si esta la juzgamos a partir de la *grid* generada para Madrid.

Ciertamente esta discrepancia no es despreciable, sobre todo si tenemos en cuenta que Madrid es la región donde los errores tienden a ser menores –figura 2–, por lo que estas cifras representan, probablemente, una cota inferior a la magnitud de las discrepancias entre ambas *grids*, aunque las dos hayan sido producidas mediante métodos similares.

El error relativo entre ambas resulta ser del 5%. La descomposición de dicho error se muestra en el cuadro 4, y aunque de nuevo la población afectada por los errores 'falso positivo' y 'falso negativo' es escasa, dicho cuadro muestra, en términos de

celdas, el patrón inverso al observado en el cuadro 3: las celdas en la que *GEOSTAT2011* no atribuye población, 'falsos negativos', pero las coordenadas del Padrón indican que si la hay, son relativamente numerosas, 653, lo que representan un 27% del total de celdas habitadas según Padrón. Ciertamente la población en dichas celdas es muy reducida, curiosamente bastante inferior a la población afectada por los 'falsos positivos", a pesar de que, en este caso, el número de celdas donde *GEOSTAT2011* asigna población, pero Padrón no, es muy bajo, del orden del 8%.[17]

Cuadro 4. Descomposición del error relativo en 3 componentes: Falsos positivos *versus* falsos negativos. Comparación *grid bottom-up* Madrid *versus GEOSTAT2011*

$$\underbrace{\frac{\sum_{P_i>0 \,\&\, P_i^{M,ref}=0} |P_i - 0|}{2 \times P^{M,ref}}}_{Falso\ \ Positivo} \qquad \underbrace{\frac{\sum_{P_i>0 \,\&\, P_i^{M,ref}>0} |P_i - P_i^{M,ref}|}{2 \times P^{M,ref}}}_{Celdas\ \ Correctas} \qquad \underbrace{\frac{\sum_{P_i=0 \,\&\, P_i^{M,ref}>0} |0 - P_i^{M,ref}|}{2 \times P^{M,ref}}}_{Falso\ \ Negativo}$$

5.0%	=	**0.2%**	+	**4.7%**	+	**0.1%**

Población afectada por los diferentes tipos de errores

31,433	12,810
0.5% de la población en la grid	0.2% de la población en la grid

Celdas afectadas por los diferentes tipos de errores

152	1,796	653
	1,948	2,449
% de celdas habitadas	23.0%	28.9%

Fuente: IEM-Padrón 2012 georeferenciado y Eurostat para GEOSTAT2011.

El mensaje general que se desprende de la comparación de los cuadros 3 y 4 es doble. Primero, tomando la *grid bottom-up* de Madrid como referencia, los métodos de desagregación espacial tienden a dispersar en exceso la población, aun cuando la resolución de la información utilizada en el proceso de desagregación sea elevada, pero esto afecta a volúmenes de población poco importantes. Esta conclusión ya la habíamos observado, y ha sido repetidamente señalada por la literatura (Gallego 2010).

Segundo, un mensaje algo más sorprendente, *GEOSTAT2011* tiende a concentrar la población más de lo que debiera en términos de celdas habitadas. En conjunto, si los resultados mostrados en los cuadros 3 y 4 fueran extrapolables a nivel nacional, cálculos aproximados indicarían que el porcentaje de celdas habitadas que debería mostrar una

[17] Es difícil pensar en una explicación para este hecho, que quizá se deba, parcialmente, a la población de Padrón no geo-referenciada, alrededor de unas 40 mil personas.

grid censal, generada a partir de coordenadas de la población, debería situarse en el entorno del 16% del total, es decir unas 80 mil celdas habitadas, aproximadamente a medio camino entre la *grid top-down* y *GEOSTAT2011*.[18] Es cierto, no obstante, que estamos hablando de diferencias en términos de población afectada en el margen, es decir, pequeñas en ambos casos, y cuyo error relativo probablemente se mantendría alrededor del 10%.

La cuestión de interés es buscarle una explicación a las diferencias entre la *grid* para la Comunidad de Madrid, generada a partir de coordenadas puntuales, y *GEOSTAT2011,* que en principio sigue el mismo método de producción –cuadro 4–. En ambos casos se trata de *grids bottom-up* construidas a partir de ficheros de población geo-referenciada. La única información disponible sobre *GEOSTAT2011* en la *web* de *Eurostat* indica que el método de producción es mediante agregación de la población residencial (METHD_CL = A = *Agregated*), y no se ha aplicado ninguna medida restrictiva por cuestiones de confidencialidad, al contrario de lo que sucede con la *grid* distribuida por el INE en su *web*, y mencionada en la introducción.[19] En la metodología censal lo único que se indica es que este tipo de producto, no dependiente de las unidades estadísticas tradicionales de recogida de la información, "*… puede realizarse dado que cada hogar tiene asignadas unas coordenadas GPS aproximadas (las del edificio donde habita)*" (INE 2011, p.-96). En consecuencia, la explicación que ofrecemos a continuación es tentativa y exploratoria, ya que no puede ser confirmada por una metodología de producción disponible públicamente.[20]

[18] Estos cálculos simplemente mantienen la estructura relativa entre las tres *grids* que aparecen en los cuadros 3 y 4, y la extrapola al total nacional.

[19] El fichero de descripción y evaluación metodológica para *GEOSTAT2011* correspondiente a España no está disponible para versión 1 de la *grid* en la *web* de *Eurostat*.

[20] Una comparación detallada entre *GEOSTAT2011* y las poblaciones municipales del Censo 2011 mostró algunas peculiaridades no recogidas en este trabajo, que se ocupa fundamentalmente de cuestiones metodológicas asociadas a la producción de *grids* de población. En concreto, una intersección de *GEOSTAT2011* con el fichero vectorial de contornos municipales del Instituto Geográfico Nacional mostró que un municipio (Haza, 09155, en la provincia de Burgos) no dispone de población en la *grid*, (a pesar de que su población, 26 habitantes, no está ajustada por factores de recuento –es un número natural– y su núcleo urbano está situado completamente dentro de una celda de la *grid*); y que de los algo más de 2.000 municipios en los que todas sus celdas habitadas son interiores a su contorno municipal, en casi las ¾ partes de los casos la población de la *grid* no es totalmente consistente con la población que se obtiene de las cifras censales. Algunas de estas discrepancias podrían justificarse por efecto de redondeos, pero otras son de difícil explicación. Abordamos estas cuestiones de detalle en otro trabajo (Goerlich, Cantarino y Reig 2015).

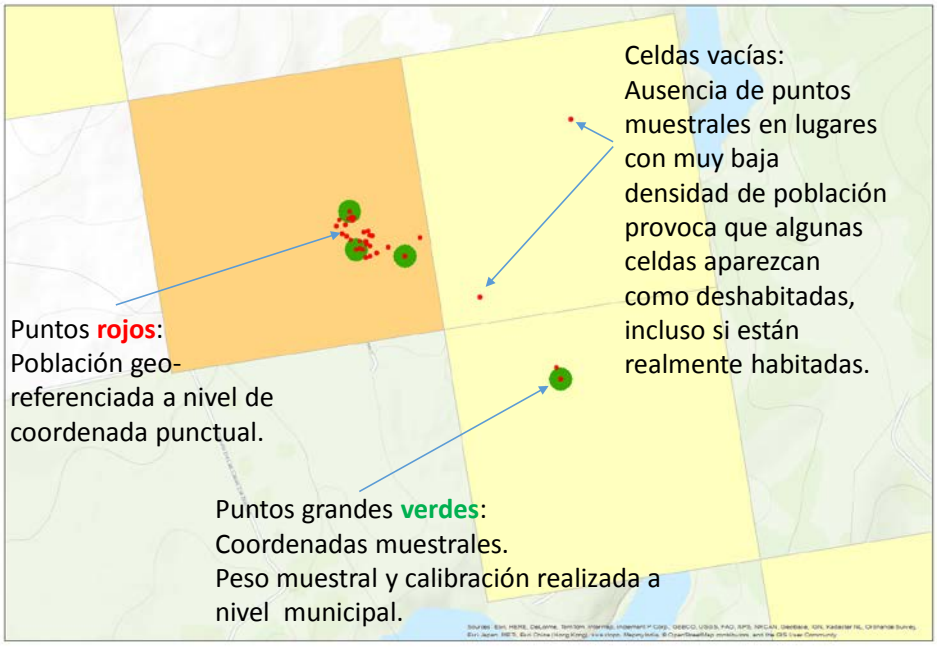
Estimamos que el problema reviste cierto interés, no solo por ofrecer una explicación a las discrepancias observadas, sino fundamentalmente porque puede ayudar a mejorar los métodos de desagregación espacial, o incluso los procedimientos *bottom-up* bajo determinadas circunstancias, llegando a proponer procedimientos mixtos que pueden ser superiores a los métodos *bottom-up* cuando la información disponible no es completa (Enrique, Molina, Ojeda, Escudero y Pérez 2013; Kraus, Moravec y Klauda 2013).

La explicación más razonable de las discrepancias observadas vuelve sobre la metodología censal con la que hemos iniciado el trabajo. El censo de 2011 ha adoptado una metodología mixta, en la que el recuento de la población y sus características más básicas proceden de un ajuste del Padrón, como fichero básico de referencia de la población residente, pero el resto de características procede de una gran encuesta por muestreo, cuya fracción de muestro teórica estaba diseñada para el 12% de la población y que en la práctica se va visto reducida al 9%. No obstante, se selecciona muestra en todas las secciones censales (INE 2011, Goerlich, Ruiz, Chorén y Albert 2015, Capítulo1).

Aunque el censo prevé una georreferenciación exhaustiva de los edificios con alguna vivienda familiar, principal o no, con lo que en principio cabría pensar que la georreferenciación podría haberse llevado a cabo a nivel del fichero precensal –al menos para el total de población y sus características más básicas–, en la práctica la georreferenciación de los hogares se ha producido a nivel de la muestra recogida. El resultado es que, aunque la literatura sobre generación de *grids* de población ha discutido ampliamente dos métodos de producción, *top-down versus bottom-up* (EFGS 2012), el INE ha implementado una versión de *bottom-up* que podríamos denominar *survey bottom-up*. En esta versión la *grid* se genera a partir de coordenadas a nivel de punto, pero dichas coordenadas llevan asociado un peso muestral como el registro de cualquier encuesta. Naturalmente, la bondad de la estimación finalmente resultante depende del diseño muestral en relación a la escala a la que deseemos ofrecer información, pero como todas las celdas habitadas no han sido objeto de muestro –ni sería factible, ni se conocen a priori– la *grid* resultante no puede cubrir todas las celdas realmente habitadas. Lo más frecuente serán errores de ‘falsos negativos’ en celdas con muy poca densidad de población, ya que en estas celdas la probabilidad de seleccionar algún hogar será muy baja. Adicionalmente los factores de elevación, calibrados a nivel

municipal, serán representativos a esta escala, pero no tienen por qué serlo a nivel de celda. Estos aspectos de diseño muestral espacial (Kumar 2007; Wang, Stein, Gao y Ge 2012) no han sido tenidos en cuenta, lo que tiene cierto impacto sobre los resultados finales, tanto en términos de celdas habitadas como de población residente en las mismas. La figura 6 trata de ilustrar visualmente esta idea, y proporciona cierta intuición del elevado número de 'falsos negativos', así como de la magnitud de error relativo de *GEOSTAT2011* en comparación con la *grid* de la Comunidad de Madrid, tal y como se muestra en el cuadro 4.

Figura 6 – Ilustración del efecto del muestreo de coordenadas sobre la generación de una *grid* de población *bottom-up*.

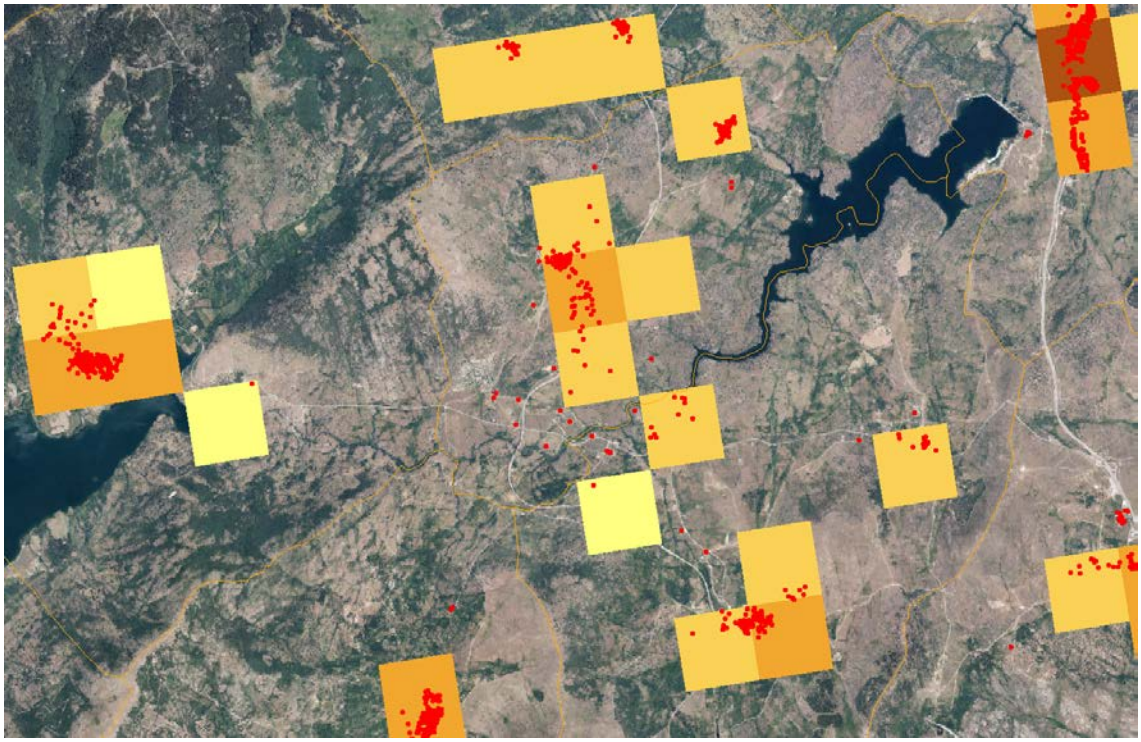


Es posible ofrecer alguna evidencia que trate de corroborar esta intuición. En primer lugar, una inspección directa de las coordenadas de Padrón frente a las celdas de *GEOSTAT2011* parece confirmar el argumento anterior. La figura 7 ofrece un ejemplo, y permite observar con claridad cómo *GEOSTAT2011* deja escapar con relativa facilidad varias coordenadas con población –puntos rojos en la figura 7–, mientras que en las zonas con una densidad elevada de puntos las celdas se identifican correctamente.

En segundo lugar, es posible examinar de forma aproximada el efecto que un muestreo similar al del censo –aunque obviamente no idéntico– tiene sobre la

generación de una *grid* de población mediante el método que hemos denominado *survey bottom-up*. Nuestro fichero de población geo-referenciada para la comunidad de Madrid dispone no solo del número de personas por coordenada, sino también del número de hogares, y del tamaño de cada uno de ellos en dicha coordenada. En consecuencia podemos simular mediante Monte Carlo el proceso seguido por el censo en la generación de *GEOSTAT2011*.[21]

Figura 7 – Población geo-referenciada de Padrón frente a celdas de *GEOSTAT2011*



Nota: Puntos rojos: coordenadas con población de Padrón 2012.

Fuente: IEM-Padrón 2012 georeferenciado y Eurostat para GEOSTAT2011.

Para ello se determinó el municipio y la sección censal de cada coordenada, y se asignaron fracciones de muestreo según el tamaño municipal (INE 2011, p.-79). Dicha muestra se distribuyó proporcionalmente al tamaño en cada sección censal, y se supuso

[21] Este ejercicio es sólo una aproximación por varias razones. En primer lugar el muestreo del ejercicio de Monte Carlo solo es una aproximación al seguido en el censo 2011. La unidad secundaria de muestreo en el censo es la vivienda familiar –ocupada o no–, mientras que nosotros muestreamos directamente hogares, lo que es asimilable a viviendas principales, de forma que no tenemos el problema de las viviendas secundarias o vacías. En segundo lugar, nuestro muestreo aplica la fracción de muestreo teórica del censo y supone una tasa de respuesta uniforme del 90%. Dentro de cada sección censal la muestra es proporcional a su tamaño. No disponemos de dos marcos de hogares con distinto proceso de selección de la muestra como en el caso del censo (INE 2011, Sección 8). En tercer lugar, los factores de elevación simplemente escalan la muestra al total de la población municipal, sin incorporar ningún otro tipo de información adicional.

una tasa de respuesta del 90%. Se introdujo la restricción de que debía haber muestra en todas las secciones censales, de al menos un hogar. A continuación, una vez seleccionada la muestra, los factores de elevación se calcularon como la inversa de la probabilidad de selección y se ajustaron –calibraron– al total de la población a nivel municipal. Finalmente, la muestra seleccionada, convenientemente ponderada por los factores de elevación estimados, se proyectó sobre las celdas de la *grid*, y se calcularon dos estadísticos frente a la *grid* generada directamente a partir de las coordenadas de la población geo-referenciada: (*i*) las celdas que dan error 'falso negativo', y (*ii*) el error relativo total.

Este proceso se replicó 1000 veces. En promedio, las celdas con 'falso negativo' fueron 522, con un error estándar de 12, lo que representa un 21% del total de las celdas con población; no muy lejos de lo observado en el cuadro 4 para *GEOSTAT2011*, donde las celdas con 'falso negativo' representan un 27%. Si hemos de otorgar alguna fiabilidad a estas cifras, lo que nos indican, extrapolando los resultados al total nacional, es que las celdas habitadas derivadas de la *grid* censal deberían estar más cerca de las 80 mil, que de las 63 mil actuales que proporciona *GEOSTAT2011*. Estas celdas, sin embargo, estarían muy poco pobladas. En el Monte Carlo recogen sólo el 0.1% de la población, el mismo porcentaje que el observado en el cuadro 4. El error relativo promedio, derivado del ejercicio de simulación, es del 7.9%, con un error estándar de 0.9. Algo más elevado de que lo que observamos en el cuadro 4 para *GEOSTAT2011*.

Ciertamente el ejercicio de simulación no puede explicar todas las diferencias observadas, pero ilustra de forma clara lo que la figura 7 muestra de forma visual: las *grids survey bottom-up* concentran la población más de lo que debieran, justo al contrario que las *grids top-down*.

5.- Conclusiones.

Este trabajo ha presentado brevemente una comparación de la *grid* de población derivada del Censo de 2011, elaborada a partir de coordenadas puntuales de la población, frente a una actualización de su inmediata predecesora: *GEOSTAT2006*, generada por métodos dasimétricos de desagregación espacial a partir de los datos de población del propio censo y una versión actualizada de *SIOSE2009*.

Ello ha permitido corroborar un resultado ya conocido. Los métodos de desagregación espacial dispersan en exceso la población. Un análisis de las fuentes de error permiten descubrir un marcado patrón espacial de las discrepancias, que pueden explicarse por 3 factores fundamentales: (*i*) el tipo de asentamiento poblacional, disperso *versus* concentrado, (*ii*) el problema de las viviendas no principales, que no es tenido en cuenta en el proceso de desagregación, y (*iii*) las propias decisiones del analista sobre que coberturas soportarán población residente. El análisis también ha puesto de manifiesto que, aunque las diferencias en términos del número de celdas son notables, las que hacen referencia a la población son mucho menores.

La disponibilidad de un fichero geo-referenciado de población a nivel de coordenada puntual para la Comunidad de Madrid nos ha permitido evaluar ambas *grids* frente a una generada por nosotros mismos por agregación de dichas coordenadas. Para la *grid top-down* las conclusiones son básicamente las mismas que las que acabamos de señalar en el párrafo anterior. La comparación frente a *GEOSTAT2011* muestra algunos resultados de interés. Por una parte, dicha *grid* parece concentrar la población más de lo que debiera, es decir muestra menos celdas habitadas de las que existen en la realidad. Ello es debido, probablemente, a un efecto diseño, es decir que la geo-referenciación de la población se ha efectuado a nivel de la muestra del censo, y no a nivel de toda la población recogida en el mismo –el fichero precensal ponderado–. Por otra parte, este efecto de concentración, aun abarcando a un número de celdas no despreciable, no afecta a grandes volúmenes de población, sino más bien al contrario, se trata de celdas con muy baja densidad, y en muchos casos aisladas de los núcleos principales.

Hasta donde nosotros conocemos esté método de elaborar *grids* de población, al que hemos denominado *survey bottom-up*, es novedoso y no ha sido empleado con anterioridad. La literatura ha concentrado sus esfuerzos en describir y depurar, o bien métodos estadísticos de desagregación espacial con información auxiliar –*top-down*–, o bien en normalizar la producción de estadísticas de población geo-referenciada a nivel de coordenada puntual –*bottom-up*– (EFGS 2012, 2014). Es necesario, sin embargo, evaluar la bondad de los nuevos métodos en situaciones concretas.

Referencias

1. **Annoni, A. (2005, ed.)** *European Reference Grids*. EUR Report 21494 EN. European Commission – Joint Research Centre. Proceedings of the "European Reference Grids" workshop, Ispra (Italy), 27-29 October 2003. [http://www.ec-gis.org/sdi/publist/pdfs/annoni2005eurgrids.pdf].

2. **Batista e Silva, F. (2011)** "The effect of ancillary data in population dasymetric mapping: A test case using the original and a modified version of CORINE Land Cover" presentado en el *European Forum for Geography and Statistics Conference* (EFGS), Lisbon (Portugal). 12 – 14 Octubre, 2011.

3. **Batista e Silva, F.; Lavalle, C. y Koomen, E. (2013)** "A procedure to obtain a refined European land use/cover map". *Journal of Land Use Science*, 8, 3, 255-283.

4. **Bhaduri, B.; Bright, E.; Coleman, P. y Dobson, J. (2002)** "LandScan: Locating people is what matters". *Geoinformatics*, 5, 2, 34–37. [http://www.ornl.gov/sci/landscan/].

5. **Bhaduri, B.; Bright, E.; Coleman, P. y Urban, M. L. (2007)** "LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics". *GeoJournal*, 69, 103–117. [http://www.ornl.gov/sci/landscan/].

6. **Center for International Earth Science Information Network (CIESIN) (2005)** *Gridded Population of the World* (*GPW*), Version 3. CIESIN, Columbia University, Palisades, NY. [http://sedac.ciesin.columbia.edu/].

7. **EFGS (2012)** *GEOSTAT 1A Final Report – Representing Census data in a European population grid*. European Forum for Geography and Statistics. Disponible en http://ec.europa.eu/eurostat/documents/4311134/4350174/ESSnet-project-GEOSTAT1A-final-report_0.pdf/fc048569-bc1c-4d99-9597-0ea0716efac3. [consultado: 30/1/2015].

8. **EFGS (2014)** *GEOSTAT 1B Final Report*. European Forum for Geography and Statistics. Disponible en http://www.efgs.info/geostat/1B [consultado: 30/1/2015].

9. **Eicher, C. and Brewer, C. (2001).** Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science, 28*, 125-138.

10. **Enrique Regueira, I.; Molina Trapero, J. E.; Ojeda Casares, S.; Escudero Tena, M. y Pérez Morales, G. (2013)** "A population grid for Andalusia" Institute of Statistics and Cartography of Andalusia (IECA). September 7, 2013.

11. **Gallego, F. J. (2010)** "A population density grid of the European Union". *Population & Environment*, 31, 6, (July), 460-473. [http://www.springerlink.com/content/0199-0039/31/6/].

12. **Gallego, F. J.; Batista, F.; Rocha, C. and Mubareka, S. (2011)** "Disaggregating population density of the European Union with CORINE land cover". *International Journal of Geographical Information Science*, 25, 12, (December), 2051-2069. [http://dx.doi.org/10.1080/13658816.2011.583653].

13. **Goerlich, F. J. y Cantarino, I. (2011)** "Population Grid for Spain – SIOSE", presentado en el *European Forum for Geography and Statistics Conference* (EFGS), Lisbon (Portugal). 12 – 14 Octubre, 2011.

14. **Goerlich, F. J. y Cantarino, I. (2012)** *Una* grid *de densidad poblacional para España*. Informe Técnico. Fundación BBVA.

15. **Goerlich, F. J. y Cantarino, I. (2013)** "A population density grid for Spain". *International Journal of Geographical Information Science*, 27, 12, (December), 2051-2069. [http://dx.doi.org/10.1080/13658816.2013.799283].

16. **Goerlich, F. J. y Cantarino, I. (2014)** "Comparing bottom-up and top-down population density grids: The Spanish Census 2011**"**, presentado en el *European Forum for Geography and Statistics Conference* (EFGS), Krakow (Polonia). 22 – 24 Octubre, 2014.

17. **Goerlich, F. J.; Cantarino, I y Reig, E. (2015)** *¿Territorios rurales o urbanos? Medio ambiente, demografía y accesibilidad*. Trabajo en curso.

18. **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. Bilbao. En prensa.

19. **IGN (2011)** *Sistema de Información de Ocupación del Suelo en España* —SIOSE2005—. Documento Resumen. Madrid, 10 de mayo de 2011. [http://www.siose.es/siose/].

20. **INSPIRE (2014)** "D2.8.I.2 INSPIRE Specification on Geographical Grid Systems – Guidelines", INSPIRE Thematic Working Group Coordinate Reference Systems and Geographical Grid Systems. Version 3.1 (2014-04-17). [http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2].

21. **Instituto Nacional de Estadística (INE 1994)** *Densidad de Población de los Municipios Españoles*. *Mapas Provinciales. Censos de Población y Viviendas 1991*. Madrid.

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

23. **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].

24. **Instituto Nacional de Estadística (INE 2015)** *¿Cómo es España? 25 mapas para descubrirla km^2 a km^2*. Subdirección General de Estadísticas de la Población. (Enero). Madrid. Disponible en http://www.ine.es/ss/Satellite?L=0&c=INEPublicacion_C&cid=1259945920521&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout&param1=PYSDetalleGratuitas [consultado: 20/1/2015].

25. **Kraus, J.; Moravec, S.; y Klauda, P. (2013)** "Disaggregation Methods For Georeferencing Inhabitants With Unknown Place Of Residence: The Case Study Of Population Census 2011 In The Czech Republic". Appendix 16 WP1B de EFGS (2014). Disponible en http://www.efgs.info/geostat/1B [consultado: 30/1/2015]

26. **Kumar, N. (2012)** "Spatial sampling design for a demographic and health survey". *Population Research Policy Review*, 26, 581-599.

27. **Martin, D.; Tate, N. J. y Langford, M. (2000)** "Refining population surface models: Experiments with Northern Ireland Census data". *Transactions in GIS*, 3, 285-301.

28. **Mennis, J. y Hultgren T. (2006)** "Intelligent dasymetric mapping and its application to areal interpolation". *Cartography and Geographic Information Science*, 33, 3, 179-194.

29. **Steinocher, K. (2011a)** "The European Dataset: The disaggregation issue", presentado en el *European Forum for Geography and Statistics Conference* (EFGS), Lisbon (Portugal). 12 – 14 Octubre, 2011.

30. **Steinocher, K. (2011b)** "A new population grid for Europe – chances and challenges", presentado en el *European Forum for Geography and Statistics Conference* (EFGS), Lisbon (Portugal). 12 – 14 Octubre, 2011.

31. **Weng, J.-F.; Stein, A.; Gao, B.-B. y Ge, Y. (2012)** "A review of spatial sampling". *Spatial Statistics*, 2, (December), 1 -14.

Anexo: Información asociada a este trabajo

La información que se detalla a continuación está disponible en el siguiente **enlace**. En ambos casos ficheros Excel con el identificador de celda y la población asociada.

- La *grid top-down*, resolución de 1 km^2, obtenida a partir de la población de las secciones censales del censo 2011 y *SIOSE2009* como información auxiliar.
- La *grid bottom-up*, resolución de 1 km^2, de la Comunidad de Madrid del Padrón 2012, obtenida a partir de población por coordenadas puntuales de las Aproximaciones Postales Principales (APP).

Ivie

Guardia Civil, 22 - Esc. 2, 1º
46020 Valencia - Spain
Phone: +34 963 190 050
Fax: +34 963 190 055

Website: http://www.ivie.es
E-mail: publicaciones@ivie.es