

La distribución espacial de la población en Andalucía

Instituto de Estadística y Cartografía de Andalucía

Iria Enrique Regueira
José E. Molina Trapero
Serafín Ojeda Casares
María Escudero Tena
Germán Pérez Morales

Índice

La distribución espacial de la población en Andalucía

Instituto de Estadística y Cartografía de Andalucía

1. Introducción
2. Ámbito de aplicación y análisis
3. Sistema de codificación
4. Fuentes de información
5. Georreferenciación de la población
 - 5.1 Enfoque bottom-up
 - 5.2 Enfoque top-down
 - 5.2.1 Información del seccionado censal
 - 5.2.2 Información catastral
 - 5.2.3 Modelos de desagregación espacial
6. Validación
7. Salvaguarda del secreto estadístico y la confidencialidad
8. Difusión de los datos
 - 8.1 Visualizador cartográfico

1. Introducción

La distribución espacial de la población a nivel de detalle, pese a ser una información trascendental para la gestión de múltiples políticas públicas y para actividades privadas, constituye, hasta el momento, una de las grandes deficiencias en materia de disponibilidad de datos. En la actualidad, y con carácter general para el conjunto de España, todas las actividades de planificación y gestión que demanden este tipo de información sólo cuentan con mapas del comportamiento espacial de la población a lo sumo a nivel municipal o, en el mejor de los casos, a nivel de sección censal, lo cual limita enormemente la capacidad de realizar análisis a escala de detalle (¿Cuántas personas viven a más de 500 metros de una parada de transporte público? ¿Cuántas personas se ven afectadas por el tráfico de una calle? ¿Cuántos niños en edad escolar viven en las proximidades de un centro educativo?).

Con esta limitación, la distribución espacial de la población se ha representado tradicionalmente mediante mapas de coropletas en los que la población de una determinada área administrativa (municipio o sección censal) se distribuye uniformemente sobre todo el espacio de dicha área. La hipótesis de base de la técnica coroplética es que la variable representada toma un valor constante en cada unidad espacial del sistema zonal que divide el territorio.

En el caso de que el sistema zonal sea de base administrativa, como ocurre con los municipios o secciones censales, el método coroplético es insatisfactorio ya que el comportamiento real de las variables a representar (el asentamiento de la población, en este caso) raramente se rige por divisiones administrativas. Así pues, los límites administrativos imponen restricciones que producen resultados engañosos al distribuir uniformemente la población en el territorio de las áreas administrativas.

Ello da lugar a que, por una parte, el aspecto visual de un mapa de coropletas sea diferente en función de la escala de análisis y, por otro lado, a que la representación cartográfica de los datos se vea muy influida por la forma en que los datos de base hayan sido agregados en las unidades espaciales del sistema zonal. La combinación de esos dos efectos, el de escala y el zonal, da lugar a lo que en el análisis espacial de datos se conoce como problema de la unidad espacial modificable. Este problema indica que la autocorrelación espacial local puede variar considerablemente en función del tamaño y la forma de las unidades espaciales sobre las que el fenómeno es medido.

Por consiguiente, el estudio de la distribución de la población sobre el territorio exige descender más allá del nivel de agregación municipal o de sección censal y superar el marco de los lindes administrativos. En este sentido se mueve la

iniciativa European Forum for Geography and Statistics (EFGS), auspiciada por un gran número de institutos nacionales de estadística de la UE, con el apoyo de Eurostat, que tiene por objeto la armonización de las estadísticas poblacionales europeas sobre la base de un *grid* o malla cartográfica de 1 x 1 km bajo un sistema geodésico de referencia común.

En Andalucía, como respuesta a este déficit de información sobre el asentamiento de la población en el territorio, y en línea con las recomendaciones del EFGS, el Instituto de Estadística y Cartografía de Andalucía (IECA) ha desarrollado un proyecto de investigación orientado a la geolocalización de la población, cuyos resultados han dado lugar a la elaboración del producto estadístico-cartográfico denominado "Distribución espacial de la población en Andalucía".

En este proyecto, el territorio de la Comunidad Autónoma de Andalucía ha sido dividido en un *grid* o malla de celdas de 250 m de lado. Como fuente principal de información para la localización de la población se han usado datos procedentes del Registro de Población de Andalucía (RPA), actualizados a fecha de 1 de enero de 2013. En el RPA figuran todas las personas empadronadas en algún municipio de Andalucía, añadiéndose información relativa a la dirección postal, así como ciertas características demográficas relativas al sexo, la edad y la nacionalidad de cada una de las personas registradas.

Esta información ha sido georreferenciada asignando las coordenadas geográficas correspondientes a las personas residentes en cualquier vivienda situada en Andalucía. La información georreferenciada proveniente del RPA se ha agregado a la malla regular de celdas de 250 m con el fin de proteger la confidencialidad de la información estadística. Esta aproximación *bottom-up* ha permitido la georreferenciación del 87,4% de la población registrada en el RPA.

El 12,6% de la población restante ha sido asignado a las celdas correspondientes usando técnicas estadísticas. Para ello se ha utilizado información auxiliar proveniente del RPA y del Catastro de Urbana (aproximación *top-down*). En general, un enfoque *top-down* genera un mapa dasimétrico de distribución de la población por desagregación espacial, partiendo de datos demográficos por unidades administrativas y aplicando métodos estadísticos de desagregación con información auxiliar.

Un mapa dasimétrico es un mapa de coropletas en el que las unidades espaciales del sistema zonal (habitualmente basado en límites administrativos) se subdividen en unidades espaciales de homogeneidad relativa definidas a partir de información auxiliar complementaria.

Los métodos de desagregación espacial forman parte de un problema más general en el análisis de datos espaciales: la transferencia de datos geográficos entre distintos sistemas zonales. Las técnicas cartográficas empleadas en la elaboración de un mapa dasimétrico, entendidas como métodos de desagregación espacial, son del tipo de interpolación zonal, ya que implican la transformación de los datos de un sistema zonal o conjunto de unidades espaciales (unidades de origen, por ejemplo, secciones censales) en otro sistema zonal (unidades de destino, por lo general de forma y tamaño regular para propiciar la homogeneidad, como es el caso de una malla cartográfica o grid).

Al proceso de desagregación espacial suele aplicársele la restricción picnofiláctica: al realizar la transferencia de los valores de la variable de estudio desde un sistema zonal administrativo (secciones censales) a un sistema zonal homogéneo (grid) ha de respetarse la restricción de volumen, es decir, ha de poderse reconstruir la población de la sección censal por agregación de la población de las celdas en ella contenidas.

La difusión de los resultados del proyecto ha tenido como referente uno de los objetivos generales del Plan Estadístico y Cartográfico de Andalucía 2013-2017: aprovechar el potencial que genera la integración de la información estadística y cartográfica para contribuir al desarrollo de la sociedad del conocimiento y difundir los datos estadísticos y cartográficos como información útil y reutilizable para la toma de decisiones por la sociedad andaluza. Para la consecución de sus objetivos, el Plan fija una serie de estrategias de desarrollo, entre las que resulta pertinente destacar la orientada a la elaboración de productos de difusión que realicen un tratamiento conjunto e integrado de todos los tipos de información utilizada, interrelacionando la de carácter estadístico con la de naturaleza cartográfica, y documentando mediante metadatos sus características técnicas. Asimismo, para facilitar la estandarización, el intercambio, la integración y la accesibilidad a la información, el Plan da prioridad a los medios de difusión que favorezcan la interoperabilidad de los datos, e identifica la Infraestructura de Datos Espaciales de Andalucía (IDEAndalucía) como una de las infraestructuras destinadas a facilitar el acceso a los datos.

Con estas orientaciones, para la difusión de los resultados en internet¹ se ha desarrollado una capa de información geográfica descargable, un servicio WMS y un visualizador cartográfico. En éste se puede consultar la información en tres mapas: "Población total", "Población por nacionalidad" y "Población por grupos de edad".

La disponibilidad de una fuente de información a mayor nivel de resolución (en grid de 250 metros, en el caso del trabajo que aquí se presenta) abre la posibilidad

¹ <http://www.juntadeandalucia.es/institutodeestadisticaycartografia/distribucionpob>

de realizar aplicaciones y estudios en ámbitos hasta ahora inexistentes, dada la insuficiencia de representatividad territorial de los datos: emplazamientos de servicios e infraestructuras públicas y determinación de las áreas de influencia (colegios, centros de salud, transportes, oficinas de empleo, etc.), estudios vinculados a salud ambiental (personas afectadas por la pluma de contaminación o por ruido), evaluación de riesgos (personas mayores de 65 años que viven en una zona con riesgo de inundación), geomarketing (lugares de concentración de sectores de población objetivo para un negocio, en función de edades y/o nacionalidades), etc.

2. Ámbito de aplicación y análisis

- Población objetivo: Población residente en Andalucía registrada en el RPA (Registro de Población de Andalucía).
- Unidad de análisis: Hogares.
- Unidad de difusión: Habitantes.
- Máxima desagregación territorial: celdas de 250 x 250 m de lado.

3. Sistema de codificación

Se ha generado una capa de celdas regulares, utilizando para ello la herramienta de generación de mallas desarrollada por Eurostat² incorporada a ArcGIS 10.

Esta malla regular se ha generado siguiendo las recomendaciones del proyecto ESSnet GEOSTAT 1A³. Su construcción inicial tuvo lugar bajo el sistema de referencia ETRS89-LAEA, adaptándola posteriormente al ETRS89-UTM Zona 30N, por ser el sistema de referencia estándar para la cartografía en España y Andalucía. Como resultado, se obtuvo una capa en formato shape con un total de 1.416.093 celdas (250 x 250 m cada una de ellas) que cubre el territorio de la Comunidad Autónoma de Andalucía.

4. Fuentes de información

- *Registro de Población de Andalucía (RPA)*: recoge todos los hogares y todas las personas con residencia en Andalucía. Además proporciona información acerca de su dirección postal y características demográficas básicas, tales como el sexo, la edad y la nacionalidad de cada uno de los miembros del hogar. Los datos de este registro, creado y mantenido por el IECA, proceden

² <http://www.efgs.info/data/eurogrid/eurostat-grid-generation-tool-for-arcgis/view>

³ <http://www.efgs.info/geostat/1A>

de fuentes administrativas. Los datos utilizados en este trabajo corresponden a la población andaluza con fecha de 1 de enero de 2013.

- *Callejero Digital de Andalucía (CDA)*: directorio creado y mantenido por el IECA en colaboración con ayuntamientos y diputaciones provinciales. Este directorio contiene información sobre vías y portales proveniente de archivos elaborados por el Instituto Nacional de Estadística (INE), Catastro y otros.
- *Catastro de Urbana*: archivo administrativo gestionado por la Dirección General de Catastro del Ministerio de Hacienda y Administración Pública, que contiene información catastral relativa a áreas residenciales, características de los edificios, características de las vías de comunicación, etc.

En particular se ha utilizado para este trabajo la siguiente información:

- Bienes inmuebles según tipo (industrial, comercial, residencial, etc.).
- Bienes inmuebles según tipo de vía donde se localizan (avenida, calle, plaza, carril, etc.).
- Bienes inmuebles según año de construcción.
- Área urbana y área urbana con bienes inmuebles residenciales (datos auxiliares usados para delimitar las áreas donde potencialmente puede residir la población).

5. Georreferenciación de la población

5.1 Enfoque bottom-up

El RPA proporciona una dirección postal para cada individuo y grupo de individuos (hogares). En el CDA cada uno de los portales tiene su dirección postal y sus coordenadas geográficas correspondientes. Mediante técnicas de enlace de registros, las direcciones postales de los hogares contenidos en el RPA se vinculan con el portal correspondiente del CDA, asignándole así una coordenada geográfica a cada uno de ellos.

La georreferenciación de los portales y de la población que vive en ellos es posible a través de tecnología de sistemas de información geográfica (SIG) una vez que se han asignado las coordenadas. El referido proceso de enlace de registros ofrece también información sobre el grado de coincidencia entre la dirección postal de RPA y la de CDA (lo cual informa de la fiabilidad del enlace y, por lo tanto, de la coordenada asignada).

En esta fase se genera una capa de puntos, como muestra la figura 1. Cada punto representa un portal al que se le asocia el número de habitantes que residen en él y las características demográficas de éstos.

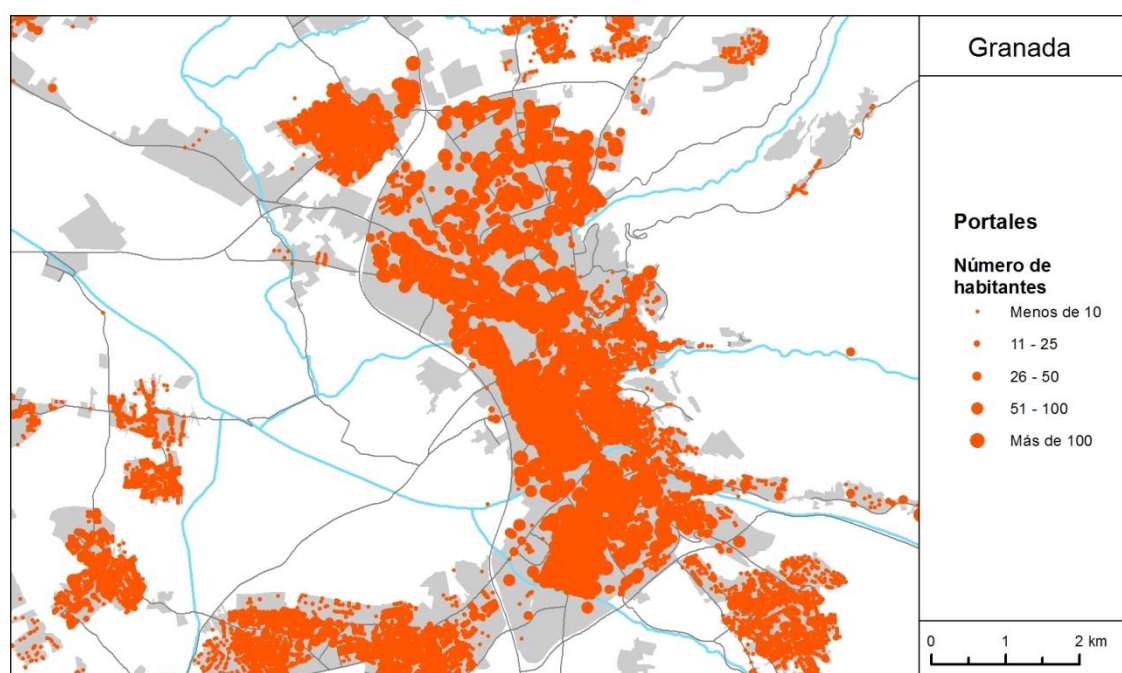


Figura 1. Portales y población

El paso siguiente consiste en la agregación de esa información a cada celda, añadiendo los datos de todos los puntos (portales y su población) localizados en cada celda. Esto se realiza en el SIG mediante un geoproceso de unión espacial. Cada portal se añade a su celda en función de su localización espacial. Como se muestra en la figura 2, cada celda contiene el número total de habitantes y de portales pertenecientes a la misma según la localización de los puntos (portales). Con este procedimiento se han localizado 7.453.107 habitantes (87,4% de la población registrada en el RPA), distribuidos en un total de 28.806 celdas que contienen información de población residente.

5.2 Enfoque top-down

Un 12,6% de la población residente en hogares no ha podido ser georreferenciada a través de la aproximación bottom-up. Para solventar esta circunstancia se ha diseñado un procedimiento top-down para la asignación de cada uno de estos hogares no georreferenciados a una celda. Una vez descartada la aproximación postal, la sección censal es el nivel máximo de desagregación territorial disponible en la información relativa a la geolocalización de estos hogares. La sección censal es una división administrativa del territorio inferior al municipio con un tamaño máximo de 2.000 votantes. Así pues, a partir de la información contenida en el RPA, para cada sección censal se dispone del número de hogares y de personas no georreferenciados que en ella residen. A partir de técnicas estadísticas y del uso de información auxiliar, el enfoque top-down permite

asignar cada uno de esos hogares a una de las celdas (o fragmentos de celda) que componen la correspondiente sección censal.

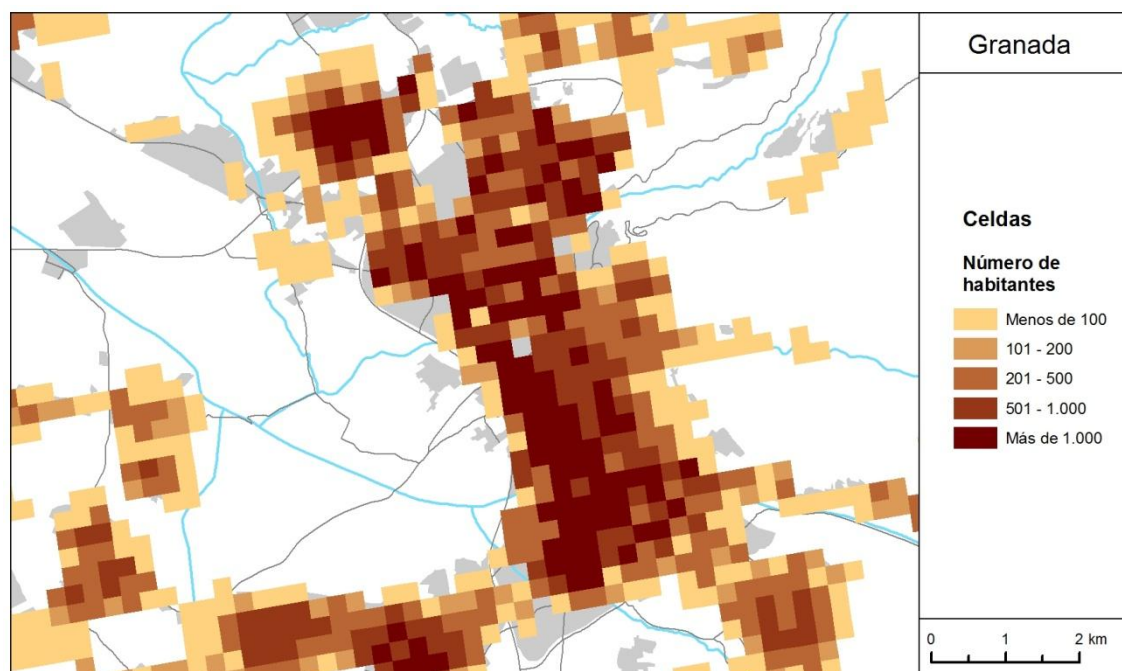


Figura 2. Número de habitantes por celda

Como paso previo a la elaboración del modelo de localización de hogares, se llevó a cabo un importante trabajo de procesamiento de la información geográfica. Éste se realizó fundamentalmente con la información proveniente de dos capas muy útiles para la construcción del modelo de localización: la capa de secciones censales elaborada por el INE y la capa del parcelario del Catastro de Urbana, elaborada por la Dirección General de Catastro.

5.2.1 Información del seccionado censal

La localización de los hogares no georreferenciados está disponible a nivel de sección censal. El objetivo principal de esta fase del proyecto es transferir esos datos (hogares no georreferenciados) desde la sección censal a la celda que le corresponde dentro de esa sección censal (transferencia de datos entre sistemas zonales).

Con este fin se genera una capa por la intersección de la capa de secciones censales con la de la malla de celdas regulares. Esta nueva capa establece la división en entidades de secciones y celdas, mostrando qué celdas quedan total o parcialmente dentro de cada una de las secciones censales en Andalucía. Las entidades espaciales generadas en esta capa de intersección no son ni secciones ni celdas, sino la intersección de ambas, la *sección/celda* (ver figura 3). Ésta es la unidad espacial de referencia para la realización del análisis, de tal modo que el modelo de localización asigna población desde la sección censal a la entidad

sección/celda y cada celda se reconstruye posteriormente por la suma de todas las sección/celdas que la componen.

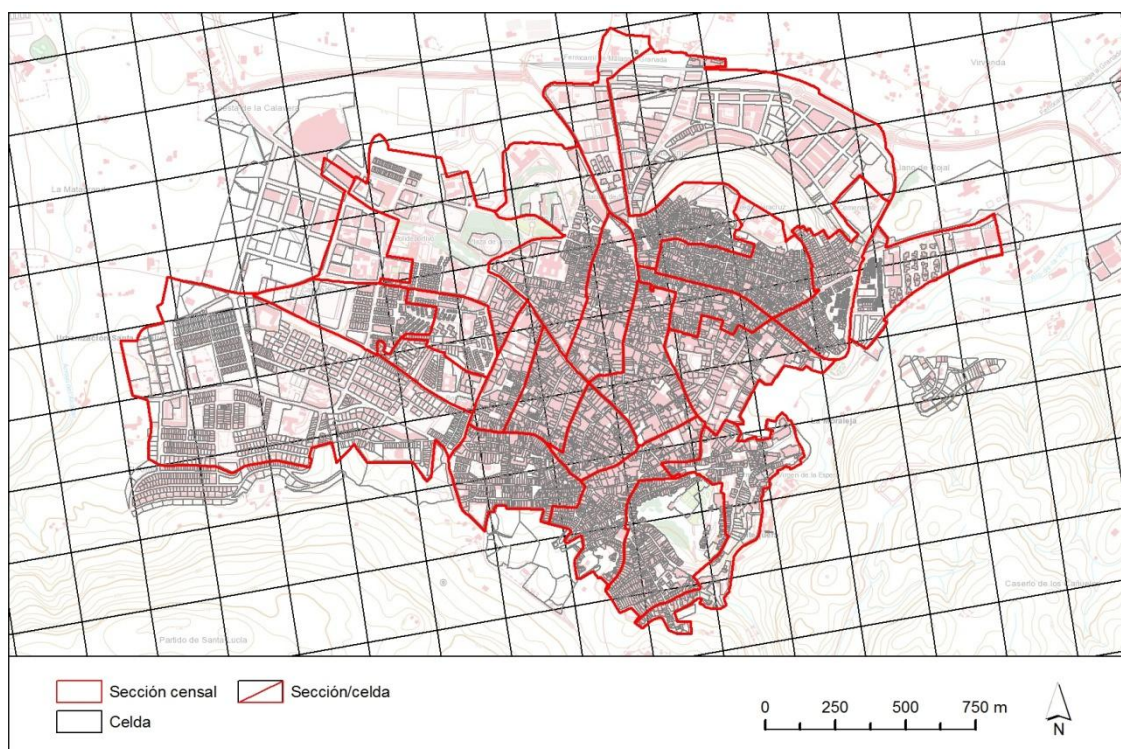


Figura 3. Unidades espaciales: secciones censales, celdas y sección/celdas

5.2.2 Información catastral

El Catastro de Urbana proporciona información correspondiente a la ocupación del territorio, identificando y localizando, entre otros usos, parcelas con uso residencial.

Las parcelas con uso residencial son la clave del proyecto, de ahí que la información de la que se dispone en Catastro relativa a sus atributos y sus características haya sido analizada en detalle.

El procesamiento de la información catastral comienza con las operaciones de fusión de todas las capas municipales del Catastro de Urbana. Se generaron ocho capas provinciales conteniendo la información del Catastro de Urbana. En algunos casos, antes de la fusión, las capas municipales han tenido que ser transformadas desde el sistema de coordenadas ED50-UTM Zona 30N al sistema ETRS89-UTM Zona 30N.

La unidad de delimitación territorial del Catastro es la parcela, mientras que la unidad de análisis en el proyecto es la sección/celda. Por ello la intersección entre la capa de parcelas con la de sección/celdas da lugar a la división de cada parcela por sección/celda. Así, se genera una nueva capa en la que figuran las parcelas que contiene cada una de las sección/celdas. Una parcela puede estar total o

parcialmente dentro de una sección/celda. La unidad espacial de información en esta capa de intersección es la *sección/celda/parcela* (ver figura 4).

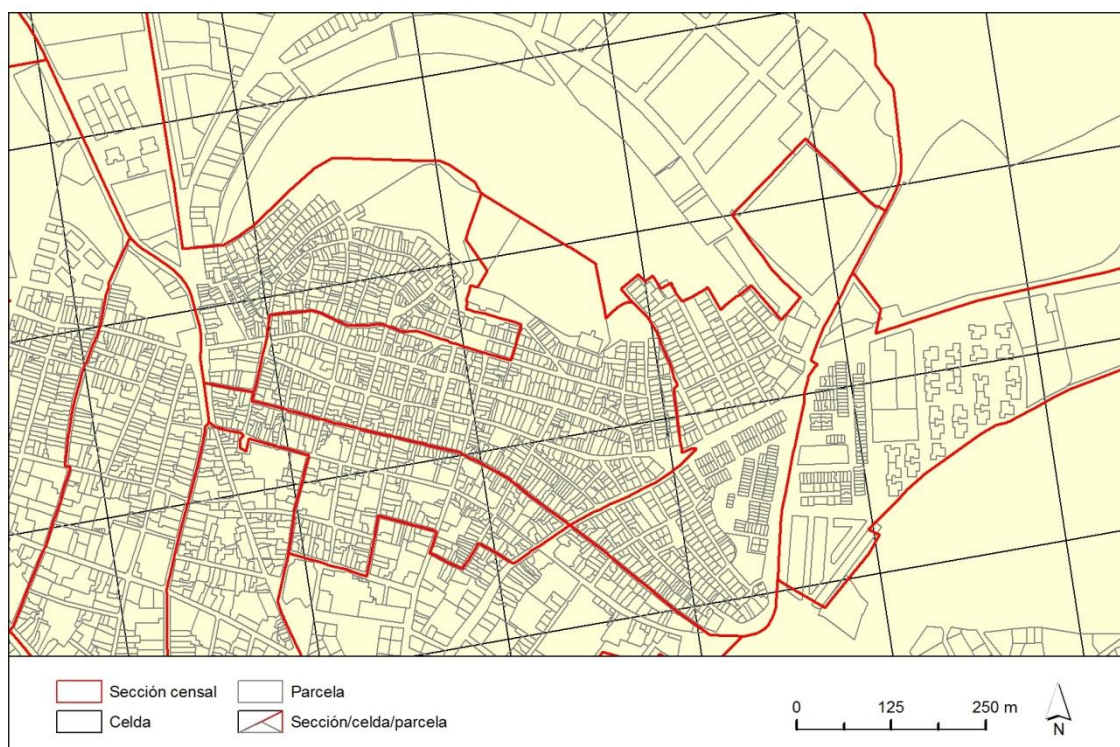


Figura 4. Unidades espaciales: secciones censales, celdas, parcelas y sección/celda/parcelas

Consecuentemente la porción de cada parcela contenida en una sección/celda se calcula a partir de la superficie de cada polígono (sección/celda/parcela). Esas porciones son utilizadas para transferir los atributos y características de una parcela a sus sección/celda/parcelas integrantes.

El análisis del Catastro se centra en las capas de parcelas urbanas que contienen información de usos, tanto residenciales como cualesquiera otros. Este análisis permite identificar las áreas que potencialmente pueden acoger población residente (por ejemplo, las áreas residenciales, comerciales y recreativas, frente a las áreas industriales o en construcción).

El Catastro de Urbana aporta también información de los bienes inmuebles relativa a la edad de estos, tipo de vía donde se localiza el bien inmueble, etc. Esta información resulta relevante para el proyecto porque la no georreferenciación de los hogares se debe fundamentalmente a desajustes en el algoritmo de enlace de registros. De este modo la información referida al grado de consolidación urbana (áreas urbanas de construcción reciente, áreas con alto porcentaje de tipos de vías poco habituales, áreas con alto porcentaje de edificios de almacén, etc.) podría arrojar luz sobre los factores que dificultaron el proceso de enlace entre los registros de las dos bases de datos (CDA y RPA).

La información sobre bienes inmuebles se ha agregado por parcela; en otras palabras, todos los atributos de los bienes inmuebles que están dentro de cada parcela han sido sumados o se les ha calculado la media aritmética, dependiendo de los casos.

Como se ha mencionado anteriormente, la unidad de análisis espacial de este proyecto es la sección/celda. No obstante, los atributos de cada parcela han sido transferidos a la unidad sección/celda/parcela, ya que ésta es la unidad de transferencia entre sistemas zonales. Así pues, los atributos de cada parcela han sido distribuidos entre las sección/celda/parcelas que la componen y, entonces, dichos atributos son agregados a nivel de sección/celda para obtener el número total de bienes inmuebles, año medio de construcción, porcentaje de bienes inmuebles registrados en avenidas, etc., por sección/celda.

5.2.3 Modelos de desagregación espacial

Una vez que toda la información alfanumérica y espacial disponible del RPA y del Catastro ha sido procesada, se desarrolla un proceso de análisis descriptivo. La población no georreferenciada no se distribuye uniformemente a través del territorio por lo que hay que llevar a cabo un análisis de la información disponible de RPA y Catastro a nivel de sección censal. Las secciones censales que contienen población no georreferenciada presentaron determinados patrones de carácter sociodemográfico, aunque la no georreferenciación no pudo ser asociada exclusivamente a las características de la población, como se ha mencionado anteriormente.

Basándose en este análisis descriptivo y en las correlaciones observadas, se han definido cuatro modelos de regresión de los hogares no georreferenciados, dependiendo de la tasa de incidencia de no georreferenciación en cada sección censal:

Modelo 1. Secciones censales con *intensidad alta de no georreferenciación* ($X > 40,22\%$), que dependen de: año medio de construcción de los edificios, superficie de las parcelas que contienen bienes inmuebles residenciales, número de bienes inmuebles con uso de almacén o industrial, hogares georreferenciados, porcentaje de población georreferenciada de la UE15, porcentaje de la población georreferenciada con 65 años y más, y porcentaje de hogares georreferenciados residentes en áreas diseminadas.

Modelo 2. Secciones censales con *intensidad media-alta de no georreferenciación* ($20,11\% < X \leq 40,22\%$), que dependen de: número de bienes inmuebles en

vías relevantes⁴, porcentaje de bienes inmuebles en otro tipo de vías⁵, porcentaje de la superficie de las parcelas que contienen bienes inmuebles residenciales sobre el área total, número de bienes con uso de almacén o industrial, hogares georreferenciados, porcentaje de población de la UE15 georreferenciada, porcentaje de población georreferenciada con 65 años y más, y la variable proxy de intensidad de no georreferenciación:

$$\frac{\text{Bienes inmuebles residenciales} - \text{hogares georreferenciados}}{\text{Superficie de las parcelas que contienen bienes inmuebles residenciales}}$$

Modelo 3. Secciones censales con *intensidad media-baja de no georreferenciación* ($10,62\% < X \leq 20,11\%$), que dependen de: número de bienes inmuebles en vías relevantes, superficie de las parcelas que contienen bienes inmuebles residenciales, porcentaje de bienes inmuebles en otros tipos de vías, porcentaje de la superficie de las parcelas que contienen bienes inmuebles residenciales sobre el área total, número de bienes inmuebles con uso de almacén o industrial, porcentaje de población de la UE15 georreferenciada, porcentaje de población georreferenciada con 65 años y más, y la variable proxy de intensidad de no georreferenciación.

Modelo 4. Secciones censales con *intensidad baja de no georreferenciación* ($X \leq 10,62\%$), que dependen de: número de bienes inmuebles en vías relevantes, superficie de las parcelas que contienen bienes inmuebles residenciales, porcentaje de bienes inmuebles en otros tipos de vías, número de bienes inmuebles con uso de almacén o industrial, hogares georreferenciados, porcentaje de población de la UE15 georreferenciada y porcentaje de población georreferenciada con 65 años y más.

Todos los modelos dieron lugar a coeficientes de ajuste satisfactorios para la regresión de los hogares no georreferenciados, tal y como se muestra en la tabla 1.

Tabla 1. Coeficientes de regresión de los hogares no georreferenciados

	<i>R² ajustada</i>
Intensidad alta de no georreferenciación	0,927
Intensidad media-alta de no georreferenciación	0,952
Intensidad media-baja de no georreferenciación	0,942
Intensidad baja de no georreferenciación	0,721

A partir de los modelos estimados a nivel de sección censal se ha obtenido una primera proyección de hogares no georreferenciados a nivel de sección/celda,

⁴ Calle, urbanización, carretera, plaza, barrio, área, polígono, otras vías.

⁵ Urbanización, avenida, diseminado, otras vías.

iniciando así un proceso iterativo de desagregación espacial, tal y como se muestra en la figura 5.

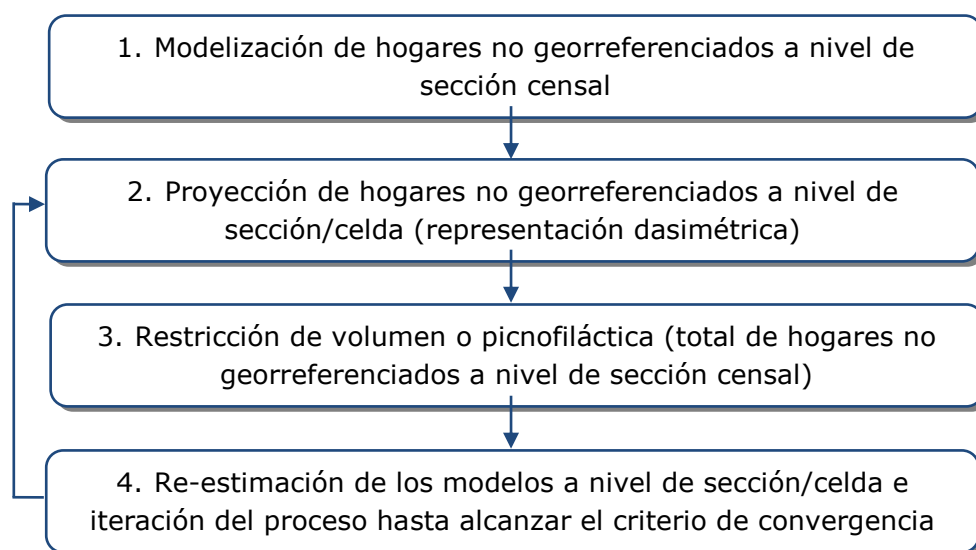


Figura 5. Proceso iterativo de desagregación espacial

El proceso se detiene una vez que se alcanza el criterio de convergencia, esto es, cuando la discrepancia entre los parámetros estimados en las últimas dos iteraciones es menor que 10^{-3} .

Los modelos estimados distribuyen los hogares no georreferenciados de cada sección censal entre las sección/celdas que la componen. Asignan un número determinado de hogares a una sección/celda, no a unas coordenadas X, Y. La selección de qué hogares del conjunto de hogares no georreferenciados de una sección censal son asignados a cada sección/celda se realiza de forma aleatoria.

Finalmente todos los hogares asignados (y sus miembros) en cada sección/celda son agregados por celda. Los hogares e individuos localizados en este proceso top-down se suman a los datos de hogares e individuos georreferenciados (enfoque bottom-up) para obtener el número total de hogares e individuos que hay en cada celda, completando así un *método híbrido* para desarrollar la malla regular de población. Los resultados para el conjunto de Andalucía se pueden ver en la figura 6. Al finalizar este proceso se obtiene un total de 40.419 celdas que albergan población.

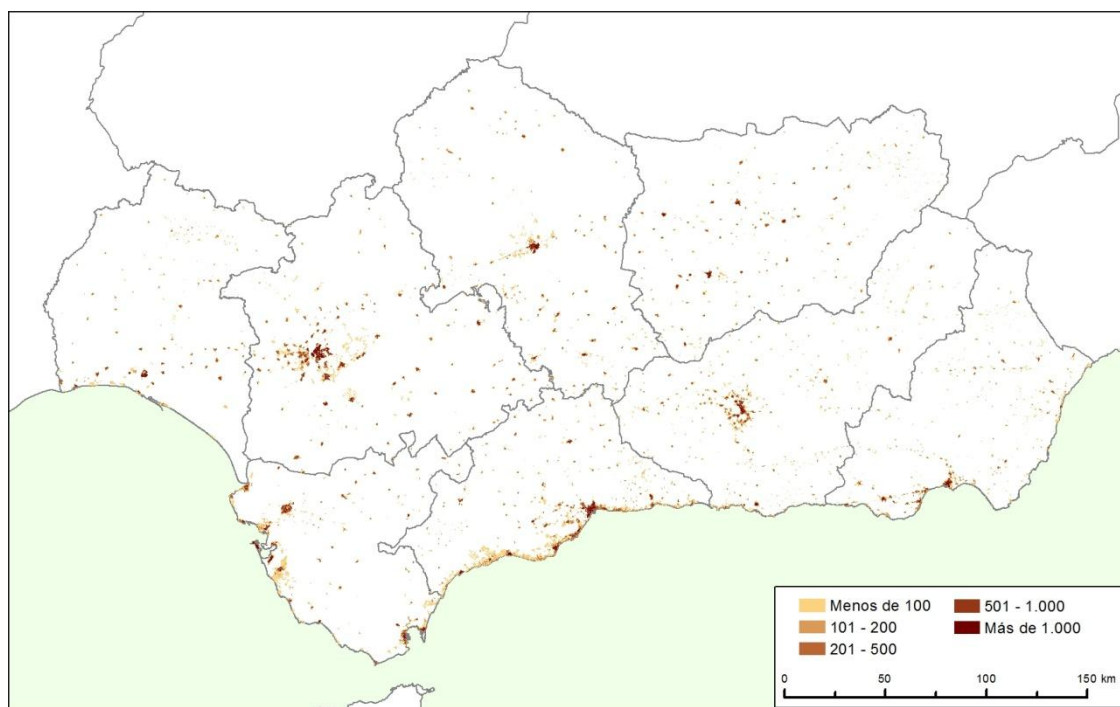


Figura 6. Distribución espacial de la población en Andalucía

6. Validación

La evaluación del funcionamiento de este enfoque híbrido se ha realizado seleccionando una muestra representativa de 330 secciones censales entre las 1.858 secciones completamente georreferenciadas de Andalucía (las cuales suponen el 32% del total de secciones censales de Andalucía). Esta muestra contiene el 18% de las secciones completamente georreferenciadas. La muestra se ha elegido de forma que sea representativa, ajustándose a una distribución por área y ocupación poblacional del territorio similar a la del conjunto de secciones censales de Andalucía.

Se ha generado una población no georreferenciada artificial, seleccionada aleatoriamente de entre los hogares de las 330 secciones de la muestra. En términos de intensidad de no georreferenciación, es decir, según tipo de modelo de regresión, la muestra de secciones de prueba queda distribuida como se presenta en la tabla 2.

Por otro lado, la forma en que la muestra de secciones censales se estructura en secciones/celdas y en celdas se presenta en la tabla 3, en tanto que la tabla 4 ofrece la distribución la población y los hogares no georreferenciados de esta muestra.

Tabla 2. Distribución de la muestra de secciones censales por intensidad de no georreferenciación

	Secciones de la muestra	%
Intensidad alta de no georreferenciación	13	3,9%
Intensidad media-alta de no georreferenciación	27	8,2%
Intensidad media-baja de no georreferenciación	33	10,0%
Intensidad baja de no georreferenciación	257	77,6%

Tabla 3. Secciones censales, sección/celdas y celdas en la muestra de secciones censales por área

Unidad espacial	Distribución
Secciones censales	330
Área (ha)	
≤ 252	91,5%
252 – 1.406	4,3%
>1.406	4,2%
Sección/celdas	16.962
Área (ha)	
≤ 2,9	21,4%
2,9 – 5	8,4%
5 – 6,3	70,2%
Celdas	16.643
Área (ha) ⁶	
≤ 2,9	19,8%
2,9 – 5	8,2%
5 – 6,3	72,0%

Tabla 4. Población y hogares en la muestra de secciones censales

	Total	No georreferenciados	Tasa de incidencia de no georreferenciación
Población	465.769	39.781	8,5%
Hogares	167.597	13.885	8,3%

La validación se ha realizado sobre la muestra de secciones comparando las cifras de población y hogares a nivel de celda derivadas de nuestro enfoque híbrido

⁶ La unidad muestral seleccionada es la sección censal, por lo que no todas las celdas en la muestra alcanzan exactamente 250 x 250 m, ya que puede haber secciones censales vecinas que comparten celdas y que no se encuentran incluidas en la muestra.

(H) y del enfoque de transferencia zonal proporcional al área⁷ (PA) frente a las cifras de población y hogares completamente georreferenciados (GR) de la muestra. Las discrepancias se han medido para los enfoques híbrido y proporcional al área mediante el error absoluto total Δ (ecuaciones (1) y (2), respectivamente):

$$\Delta_H = \sum_i |Y_i^H - Y_i^{GR}| \quad (1)$$

$$\Delta_{PA} = \sum_i |Y_i^{PA} - Y_i^{GR}| \quad (2)$$

donde i indexa las celdas de la muestra de validación, e Y_i^j representa las cifras obtenidas en la celda i por el enfoque denotado con el superíndice $j = H, PA, GR$. Para facilitar la comparación directa, se ha calculado una tasa que relaciona el error absoluto total Δ con el total de hogares o de población, dependiendo del error medido. Finalmente, se ha obtenido un índice de discrepancia estandarizado dividiendo esta tasa por 2, lo que da un valor en el intervalo $[0, 1]$, recorriendo así el rango que va desde la asignación/desagregación espacial exacta a la completamente errónea.

Las asignaciones proporcionadas por los modelos H y PA están referidas a nivel de sección/celda. Esos datos son agregados por celdas; la tabla 5 muestra los índices de discrepancia en la división inicial de sección/celdas y en la división de celdas construidas por agregación de los datos de las sección/celdas. Se observa que el método híbrido supera al método proporcional al área tanto a nivel de sección/celda como a nivel de celdas de 250 x 250 m.

Tabla 5. Índices de discrepancia a nivel de sección/celda y celda (250 X 250 m). Hogares y población asignados

	Híbrido	Proporcional al área
Sección/celda		
Población	5,5%	36,3%
Hogares	5,3%	36,1%
Celdas		
Población	5,4%	35,3%
Hogares	5,2%	35,2%

Con el fin de comparar el método híbrido con otras experiencias de mallas de celdas se ha realizado una agregación geométrica de 1 km para obtener una

⁷ En este enfoque se transfiere la población de la sección censal a las correspondientes sección/celdas proporcionalmente al área de las mismas.

cuadrícula “simulada” de estas dimensiones⁸, ya que ésta se utiliza más frecuentemente en la literatura geoestadística.

Específicamente, para esta comparación se ha tomado el *grid* de densidad de población de España⁹, elaborado por Francisco Goerlich e Isidro Cantarino. La tabla 6 muestra los resultados del índice de discrepancia con celdas de 1 x 1 km.

Tabla 6. Índices de discrepancia a nivel de celdas de 1 x 1 km. Hogares y población asignados

	Regresión <i>log-log</i> de densidades ¹⁰	Proporcional al área	Híbrido
Población	4,4%	17,7%	1,8%
Hogares	--	17,8%	1,7%

El método híbrido supera al método proporcional al área y al de regresión *log-log* de densidades. Sin embargo los resultados difieren a lo largo del territorio. El enfoque híbrido supera al método proporcional al área en todos los grupos de población y por tamaño poblacional de las celdas, aunque ambos métodos se comportan de forma diferente. El método híbrido registra en zonas pobladas mejores resultados que el método proporcional al área. En términos de densidad de población, el enfoque híbrido obtiene mejores resultados en las zonas de densidad media mientras que el enfoque proporcional al área registra sus mejores resultados en áreas de alta densidad, como se muestra en las figuras 7 y 8¹¹.

Las áreas poco pobladas presentan también ciertas debilidades para el método híbrido, registrando en éstas y en las de baja densidad las mayores discrepancias.

Sin embargo el uso de información auxiliar mejora significativamente las discrepancias en la desagregación en casos extremos tales como las áreas no pobladas. Nuestras 330 secciones censales de la muestra, estructuradas en 16.643 celdas, tienen un 89,5% de celdas no pobladas. El método híbrido asigna población sólo al 2% de esas celdas no pobladas.

Aunque es complicado lograr una medida de la discrepancia espacial, es decir, cuán lejos de la ubicación correcta queda asignada la población a través del

⁸ La unidad de muestra ha sido la sección censal, por lo que en la muestra no todas las celdas tienen una dimensión de 1 x 1 km, ya que puede haber secciones censales vecinas que comparten celdas y que no se encuentran incluidas en la muestra.

⁹ GOERLICH, Francisco J., y CANTARINO, Isidro (2012). Una grid de densidad de población para España. Bilbao, Fundación BBVA. ISBN: 978-84-92937-39-4, 138 pp.
http://www.fbbva.es/TLFU/dat/Una_grid_de_densidad.pdf

¹⁰ Regresión *log-log* de densidades se refiere al método aplicado en “Una grid de densidad de población para España” por Francisco Goerlich e Isidro Cantarino.

¹¹ Se representan los índices de discrepancia a nivel de sección/celda. Una sección/celda corresponde a una única sección censal mientras que una celda puede cubrir áreas de varias secciones censales. Ello hace que el análisis de las características de las secciones censales sea sólo posible a nivel de sección/celda.

modelo, la autocorrelación espacial de los errores (hogares asignados – hogares registrados) puede proporcionar cierta información que sirva como primera aproximación para medir la discrepancia espacial, al menos la relacionada con el signo de los errores vecinos (ver figuras 9 y 10).

FIGURA 7. Índice de discrepancia por densidad de población en las secciones censales (población por ha)

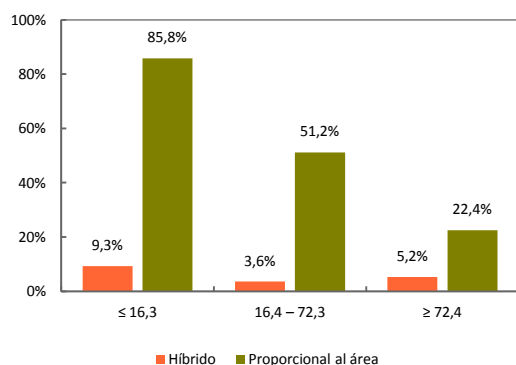


FIGURA 8. Índice de discrepancia por población en las secciones censales

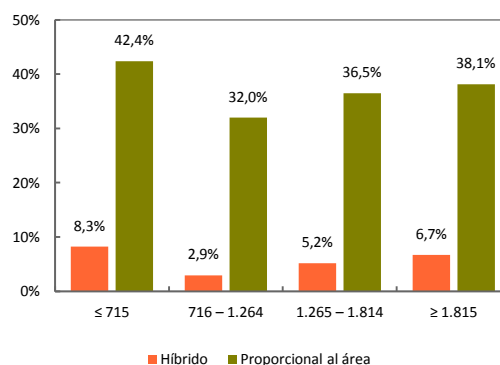


FIGURA 9. Autocorrelación espacial de los errores en el enfoque híbrido. Pesos según el criterio de la torre

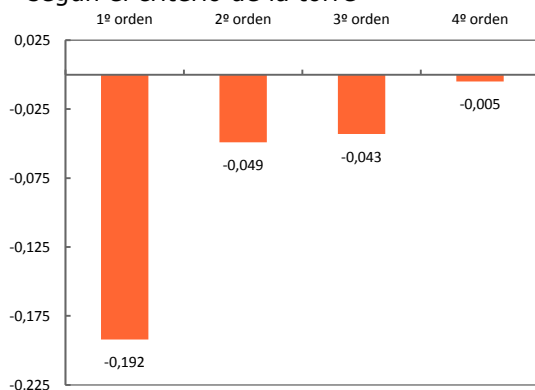
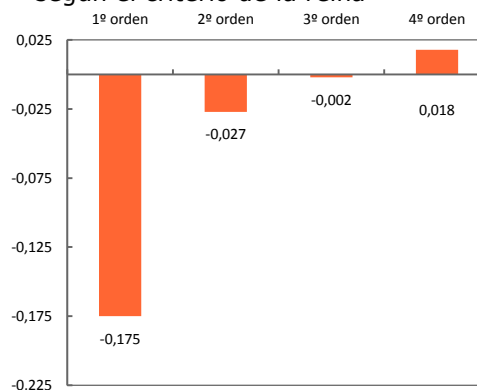


FIGURA 10. Autocorrelación espacial de los errores en el enfoque híbrido. Pesos según el criterio de la reina



Las correlaciones de primer orden¹² son considerablemente más altas que otras correlaciones con mayor orden de contigüidad. Así, las diferencias entre los hogares asignados y el número real de hogares en las sección/celdas vecinas son de signo opuesto y mayores a las que se dan con las sección/celdas vecinas de "larga distancia".

Este procedimiento de validación comprueba el funcionamiento del método híbrido aunque no valida ni prueba el funcionamiento del modelo de regresión. El modelo se construye con base en una subpoblación de sección/celdas que

¹² Correlación espacial basada en pesos de contigüidad, según los criterios de la torre y de la reina, de primer, segundo, tercer y cuarto orden, excluyendo los órdenes más bajos en cada caso. Los errores corresponden a las diferencias en los valores de los hogares en las sección/celdas.

presentan un patrón característico: problemas de georreferenciación. Por lo tanto, una muestra artificial de población no georreferenciada no se ajustará a la población modelizada en la regresión. Para probar el funcionamiento del modelo de regresión por separado, se podría plantear una selección de casos reales de población no georreferenciada, sobre la que estudiar en profundidad su localización a fin de obtener una georreferenciación correcta y, a partir de ella, comparar estos resultados con los asignados por la regresión.

7. Salvaguarda del secreto estadístico y la confidencialidad

Como ya se ha mencionado, una vez que el 87,4% de la población ha sido georreferenciado con exactitud, se ha añadido a la malla regular el 12,6% restante asignado con la operativa top-down. Antes de llevar a cabo la difusión de los datos, estos han sido sometidos a un proceso de Control de Divulgación Estadística (SDC, en su acrónimo inglés).

Los métodos de SDC reducen el riesgo de difundir datos sensibles a un nivel aceptable, a la vez que permiten publicar el máximo de información posible. En este proyecto se ha fijado un determinado umbral a partir del cual se difunden los datos, de manera que los datos quedan protegidos al impedirse tanto la difusión de datos reales por debajo de dicho umbral como la posibilidad de su deducción exacta.

Se han aplicado métodos no perturbativos para proteger la confidencialidad de los datos, de manera que se ha reducido la cantidad de información difundida debido a operaciones de supresión:

- Supresión primaria de los valores que están por debajo del umbral fijado.
- Supresión secundaria de valores para proteger la supresión primaria. Las tablas incluyen los totales generales junto a los valores de las categorías, existiendo una relación lineal entre estos valores. La supresión secundaria de datos es necesaria para evitar que los valores sensibles sean recalculados por diferencia.

La aplicación del control de divulgación se ha realizado mediante un enfoque doble:

- Un algoritmo específico diseñado con este propósito.
- Tau-Argus: software diseñado para prestar asistencia en la protección de datos y en la producción de tablas seguras. Este software fue desarrollado como parte del proyecto CASC¹³ (Aspectos Computacionales del Secreto Estadístico), que formó parte del Quinto Programa Marco de la Unión Europea.

¹³ <http://neon.vb.cbs.nl/casc/index.htm>

8. Difusión de los datos

Dentro del grupo de capas "G07 Sistema Urbano"¹⁴ de la colección de datos espaciales DERA (Datos Espaciales de Referencia de Andalucía) del IECA, se encuentran accesibles para su descarga en formato de archivo de formas los datos del proyecto resultantes tras el proceso de protección SDC, su09_grid_poblacion_250.shp.

Igualmente, se ha incorporado al proyecto la estrategia de difusión, acceso y reutilización de la información del Plan Estadístico y Cartográfico de Andalucía 2013-2017, al ofrecer en la infraestructura de datos espaciales IDEAndalucía las capas de información espacial generadas como servicios interoperables. Ello permitirá a los usuarios consumir servicios de visualización WMS¹⁵, con lo que se facilitará su uso en combinación con otras capas de información espacial, y se contribuirá así al desarrollo de procesos de generación de valor añadido basados en la reutilización de la información por parte de la Administración Pública, los agentes económicos, sociales y del conocimiento, así como por la ciudadanía en general.

Por otro lado, la página web del IECA ofrece también un visualizador cartográfico, descrito en el apartado siguiente, el cual permite realizar consultas interactivas. Los datos son representados en tres mapas de celdas: un mapa de coropletas, "Población total", y dos mapas de densidad de puntos, "Población por nacionalidad" y "Población por grupos de edad", como se muestran en la figura 11.

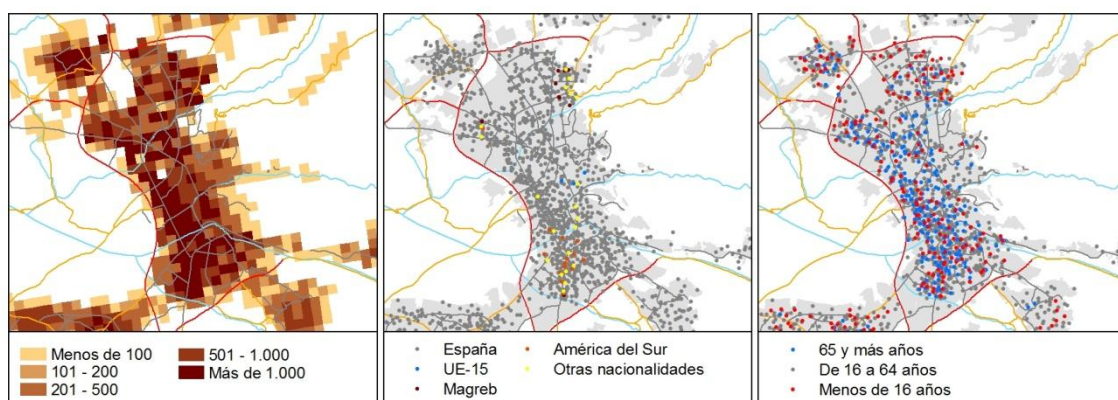


Figura 11. Mapas de celdas interactivos: población total, población por nacionalidad y población por grupos de edad

En los dos últimos, la representación de los puntos se realiza en cada celda de una forma aleatoria, de tal modo que se evita la localización exacta de las direcciones postales y de la población. Esto también facilita la representación de varias categorías simultáneamente, por ejemplo, si aparecen tres categorías en una

¹⁴http://www.juntadeandalucia.es/institutodeestadisticaycartografia/DERA/ficheros/G07_Sistema_Urbano.zip

¹⁵<http://www.juntadeandalucia.es/institutodeestadisticaycartografia/geoserver-ieca/grid/wms?>

misma dirección (portal) los tres puntos quedarían superpuestos en el lugar exacto que indican sus coordenadas geográficas.

En la generación de estas capas se ha utilizado ArcGIS 10. El comando para representar la información es *Propiedades de la capa/Simbología/Cantidades/Densidad de puntos*.

Por otro lado, para transferir la información a la aplicación de visualización de internet (visualizador cartográfico) se ha utilizado la herramienta de ArcGIS *Herramientas de administración/Clase de entidad/Crear puntos aleatorios*. Esta herramienta genera una capa nueva con topología puntual a partir de los datos de cada una de las celdas. Cada punto representa un número determinado de habitantes, que cambia según la escala de representación.

8.1 Visualizador cartográfico

El visualizador cartográfico¹⁶ ha sido desarrollado para mostrar la información en internet y hacer posible la consulta de este tipo de datos. Por otro lado, se ha utilizado en esta aplicación la información cartográfica de base generada por la Subdirección de Cartografía del IECA. Se han diseñado, a su vez, los componentes principales del visualizador con el fin de hacerlo sencillo y facilitar al usuario el acceso a la información a través de mapas. Se ha obtenido así un visualizador que no requiere por parte del usuario conocimientos específicos en tecnologías de la información.

Tecnológicamente, el proyecto ha sido desarrollado en Open Source, en particular el visualizador cartográfico web OpenLayers y Geoserver para proveer servicios cartográficos.

El visualizador cartográfico ha sido personalizado y se han desarrollado funcionalidades específicas para responder a las necesidades particulares del proyecto, por ejemplo, controles y comandos de leyenda, visualización de la información alfanumérica para un área específica del territorio, etc.

Hitos en el desarrollo:

- Leyenda dinámica para mostrar sólo la información relevante del mapa.
- Herramienta para la consulta de datos alfanuméricos a nivel de celda. Formato de datos alfanuméricos de los servicios cartográficos con el fin de personalizar los datos mostrados.
- Ventana emergente para presentar los datos alfanuméricos mencionados antes.
- Mantenimiento de la escala de representación y zoom en el cambio de un mapa a otro.

¹⁶ <http://www.juntadeandalucia.es/institutodeestadisticaycartografia/bd/GRIDvisor/densidad.jsp>

- Carga de datos de servicios cartográficos para utilizar esta información desde el visualizador cartográfico.
- Pruebas de usabilidad, pruebas de velocidad de carga, pruebas del visualizador y de los servicios cartográficos.



Figura 11. Visualizador cartográfico: población total



Figura 13. Visualizador cartográfico: población por nacionalidad