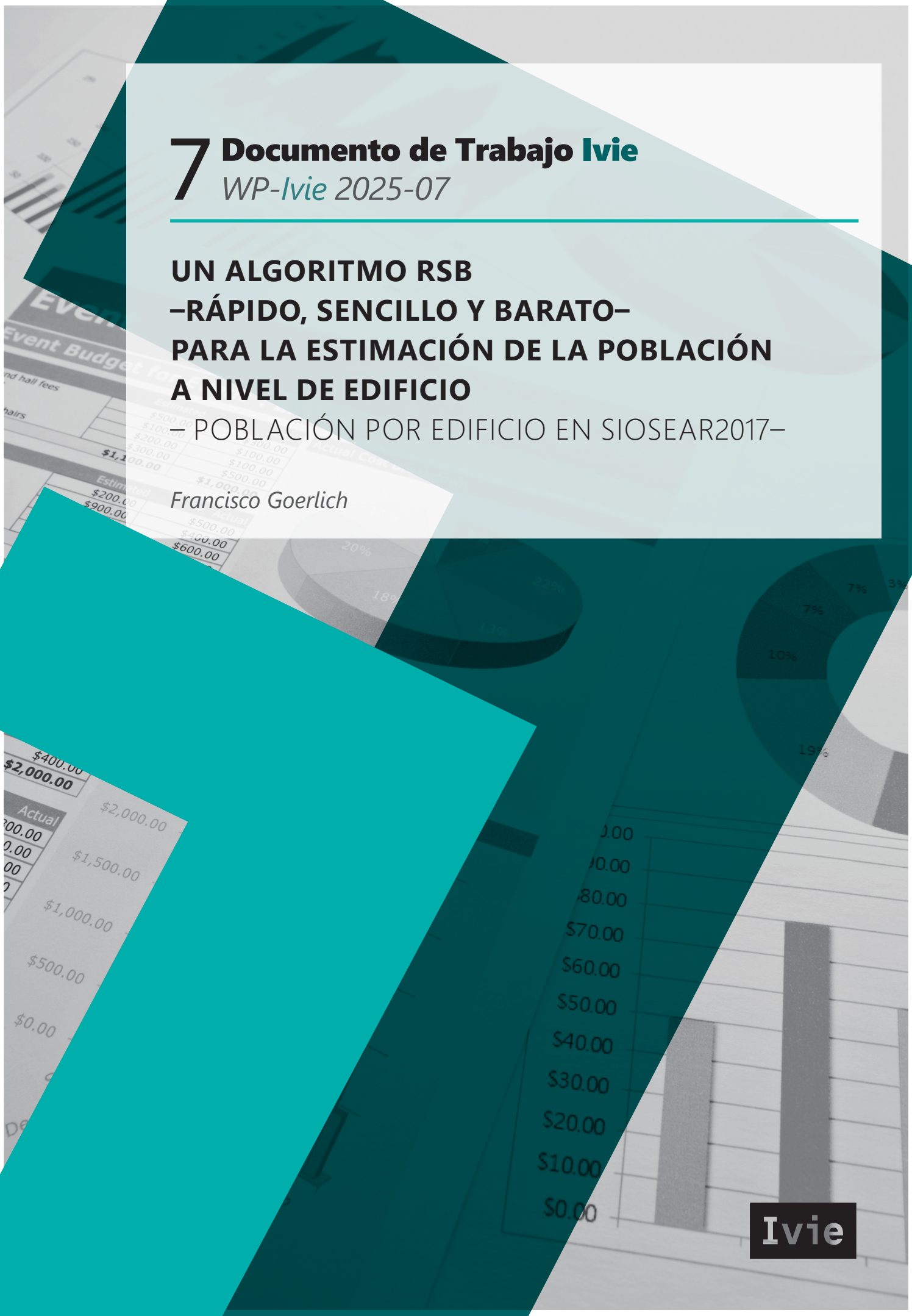


7 **Documento de Trabajo Ivie**
WP-Ivie 2025-07

UN ALGORITMO RSB –RÁPIDO, SENCILLO Y BARATO– PARA LA ESTIMACIÓN DE LA POBLACIÓN A NIVEL DE EDIFICIO

– POBLACIÓN POR EDIFICIO EN SIOSEAR2017–

Francisco Goerlich

Ivie

Los documentos de trabajo del Ivie ofrecen un avance de los resultados de las investigaciones económicas en curso o análisis específicos sobre debates de actualidad, con objeto de divulgar el conocimiento generado por diferentes investigadores.

Ivie working papers offer a preview of the results of economic research under way, as well as an analysis on current debate topics, with the aim of disseminating the knowledge generated by different researchers.

La edición y difusión de los documentos de trabajo del Ivie es una actividad subvencionada por la Generalitat Valenciana, Conselleria de Hacienda y Modelo Económico, en el marco del convenio de colaboración para la promoción y consolidación de las actividades de investigación económica básica y aplicada del Ivie.

The editing and dissemination process of Ivie working papers is funded by the Valencian Regional Government's Ministry for Finance and the Economic Model, through the cooperation agreement signed between both institutions to promote and consolidate the Ivie's basic and applied economic research activities.

Cómo citar/How to cite:
Goerlich Gisbert, F. « Un algoritmo RSB –Rápido, Sencillo y Barato– para la estimación de la población a nivel de edificio –Población por Edificio en SIOSEAR2017–». Working Papers Ivie n. º 2025-7. València: Ivie. https://doi.org/10.12842/WPIVIE_0725

Versión: Septiembre 2025 / Version: September 2025

Edita / Published by:

Instituto Valenciano de Investigaciones Económicas, S.A.

C/ Guardia Civil, 22 esc. 2 1º - 46020 València (Spain)

DOI: https://doi.org/10.12842/WPIVIE_0725

WP-Ivie 2025-7

Un algoritmo RSB –Rápido, Sencillo y Barato– para la estimación de la población a nivel de edificio

– Población por Edificio en SIOSEAR2017–

Francisco Goerlich[1]

Resumen

Para el estudio de la distribución de la población lo ideal sería disponer de un fichero de población georreferenciada a nivel de coordenada puntual a partir de su dirección postal. Dicho fichero podría ser agregado a la resolución que deseáramos para un ejercicio concreto, lo que proporcionaría una total flexibilidad. De esta forma podríamos obtener la población a nivel de edificio o manzana para análisis municipales, incluso de barrios en grandes ciudades, o podríamos generar *grids* de población con una elevada resolución, que nos permitieran hacer análisis tremendamente detallados. Esta información no está disponible públicamente.

En la investigación llevada a cabo por Goerlich y Mollá (2025) se desagrega la *grid* de población censal 2021 del Instituto Nacional de Estadística (INE), con resolución de 1 km x 1 km, a celdas de 100 m x 100 m, mediante métodos dasimétricos a partir del Sistema de Información de Ocupación del Suelo de España de Alta Resolución (SIOSEAR) referido a 2017.

El proceso de desagregación de Goerlich y Mollá (2025) utiliza como geografía intermedia los edificios residenciales de SIOSEAR2017, es plenamente consistente con la *grid* original –en el sentido de que agrega, celda a celda, la población de la *grid* censal–, y utiliza los dos tipos de información clave que la literatura sobre métodos dasimétricos de desagregación espacial ha señalado como relevantes: la tipología de los edificios –residenciales versus no residenciales– y su altura –la población vive en 3D–. Este estudio recupera una información no almacenada en su momento, y que creemos puede ser extraordinariamente útil en la práctica, la población de la geografía intermedia, es decir, la de los edificios residenciales, antes de agregar dicha población a las celdas de 100 m x 100 m efectuada por Goerlich y Mollá (2025). Esto proporciona una estimación de la población por edificio que puede ser de interés para múltiples aplicaciones.

Palabras clave: Población; Censo; *Grids* de población; Población por edificio.

Clasificación JEL: J11; R1

Abstract

For the study of population distribution, the ideal approach would be to have a georeferenced population file at the point-coordinate level, based on the postal addresses. This file could be aggregated at the desired resolution for a specific exercise, providing complete flexibility. In this way, we could obtain population data at the building or block level for municipal analysis—even for neighborhoods in large cities—, or generate high-resolution population grids, enabling extremely detailed analyses. However, this information is not publicly available.

In the research carried out by Goerlich and Mollá (2025), the 2021 census population grid from the Spanish National Institute of Statistics (INE), with a resolution of 1 km x 1 km, was disaggregated into 100 m x 100 m cells using dasymetric methods based on the Spanish High Resolution Land Use Information System (SIOSEAR) for 2017.

The disaggregation process by Goerlich and Mollá (2025) uses the residential buildings of SIOSEAR2017 as the intermediate geography. It is fully consistent with the original grid—aggregating the population of the census grid cell by cell—and uses the two key types of information that the literature on dasymetric spatial disaggregation methods identifies as relevant: the typology of the buildings (residential versus non-residential) and their height (as the population lives in 3D). This study recovers information that was not previously stored and that we believe may be extraordinarily useful in practice: the population of the intermediate geography—i.e., residential buildings—prior to aggregation into the 100 m x 100 m cells carried out by Goerlich and Mollá (2025). This provides an estimate of the population per building, that may be of interest for multiple applications.

Keywords: Population; Census; Population Grids; Population per building.

JEL classification: J11; R1

[1] F. Goerlich, Universitat de València e Ivie.

1.

Introducción y motivación

Este trabajo es un efecto colateral de Goerlich y Mollá (2025).

Para el estudio de la distribución de la población lo ideal sería disponer de un fichero de **población georreferenciada a nivel de coordenada puntual** a partir de su dirección postal. Dicho fichero podría ser agregado a la resolución que deseáramos para un ejercicio concreto, lo que proporcionaría una total flexibilidad. De esta forma podríamos obtener la población a nivel de edificio o manzana para análisis municipales, incluso de barrios en grandes ciudades, o podríamos generar *grids* de población con una elevada resolución, que nos permitieran hacer análisis tremendamente detallados. Esta información, que en teoría existe por parte del Instituto Nacional de Estadística (INE) a partir del censo 2021 (INE 2023), no está —¡y lo que es peor, tampoco estará en el futuro, 🤦‍♀️!— disponible públicamente, entre otras razones por cuestiones de confidencialidad estadística.

En la investigación llevada a cabo por Goerlich y Mollá (2025) se desagrega la *grid* de población censal 2021 del INE —*GEOSTAT2021*—, con resolución de 1 km x 1 km , a celdas de 100 m x 100 m, mediante métodos dasimétricos (Eicher y Brewer 2001) a partir del Sistema de Información de Ocupación del Suelo de España de Alta Resolución (SIOSEAR) referido a 2017 —*SIOSEAR2017*–. El resultado es una distribución de la población tremendamente granular, con más de 1.3 millones de celdas[2]. Este formato para las estadísticas demográficas fue impulsado por Eurostat hace ya tiempo con objetivos múltiples, y el sistema de *grids* está normalizado a nivel europeo (INSPIRE 2023a).

El proceso de desagregación de Goerlich y Mollá (2025) utiliza como **geografía intermedia** los **edificios residenciales** de *SIOSEAR2017*, es plenamente consistente con la *grid* original —en el sentido de que agrega, celda a celda, la población de la *grid* censal—, y utiliza los dos tipos de información clave que la literatura sobre métodos dasimétricos de desagregación espacial ha señalado como relevantes: la tipología de los edificios —residenciales *versus* no residenciales— (Gallego 2010; Batista e Silva, Poelman, Martens y Lavalle 2013) y su altura —la población vive en 3D— (Goerlich 2016; Steinnocher, De Bono, Chatenoux, Tiede y Wendt 2019; Grippa, Linard, Lennert, Georganos, Mboga, Vanhuysse, Gadiana, y Wolff 2019; Schug, Frantz, van der Linden y Hostert 2021).

Este estudio recupera una información no almacenada en su momento, y que creemos puede ser extraordinariamente útil en la práctica, la población de la **geografía intermedia**, es decir, la de los **edificios residenciales**, antes de agregar dicha población a las celdas de 100 m x 100 m efectuada por Goerlich y Mollá (2025). Esto proporciona una estimación de la **población por edificio** que puede ser de interés para múltiples aplicaciones (Lwin y Murayama 2009, 2011), y que permite una agregación mucho más versátil por áreas arbitrarias que la de las celdas de 100 m x 100 m. Es en este sentido en el que se presenta un **algoritmo *RSB*** —***R***ápido, ***S***encillo y ***B***arato— para la estimación

[2] Exactamente 1.324.147 celdas.

de la **población a nivel de edificio residencial**. El algoritmo se basa en la interpolación por volúmenes edificados de la población de las celdas de la *grid* censal 2021 de 1 km x 1 km a los edificios residenciales contenidos en ellas, según la información disponible en *SIOSEAR2017*. Al igual que sucede en Goerlich y Mollá (2025) la población está restringida a las celdas con población, de forma que no hay población fuera de ellas, aunque haya edificios residenciales.

La estructura del trabajo es la siguiente. El apartado 2 describe los datos y el *software* utilizados. El apartado 3 se centra en los aspectos técnicos del proceso de desagregación que, si bien no aporta novedades respecto al seguido en Goerlich y Mollá (2025), sí presenta algunas peculiaridades que es necesario tener en cuenta. A continuación se hace un ejercicio de validación y consistencia limitada de las estimaciones obtenidas para tratar de evaluar hasta qué punto nuestras estimaciones son adecuadas. A este nivel de resolución conocer la precisión de las estimaciones es extremadamente difícil. El apartado 5 describe la base de datos y la información ofrecida. Finalmente, el apartado 6 recoge los comentarios finales.

2.

Datos y *software* utilizados

Con ocasión del Censo de Población y Viviendas 2021, y por mandato de Eurostat (INE 2023), el Instituto Nacional de Estadística (INE) publicó, a finales de enero de 2024, la población por celda sobre una *grid* regular de 1 km x 1 km de acuerdo con las especificaciones de la Directiva Comunitaria INSPIRE (2023a). La información publicada es un simple fichero Excel —o alternativamente de texto— con solo dos variables, el código estandarizado de celda y la población asociada. La población suma el total de la población del censo —47.400.798 residentes— y es la única variable ofrecida por el INE en este formato. La *grid* censal 2021 publicada por el INE —*GEOSTAT2021*—, tiene 115.410 celdas habitadas, lo que significa que solo el 22,6% del territorio está ocupado por población residente a esta escala. Las características básicas de *GEOSTAT2021* y la distribución actual de la población en este formato se describen extensamente en el capítulo 2 de Goerlich y Mollá (2025). Esta es la información básica de la que partimos.

La otra información relevante a nuestros efectos es la del Sistema de Información de Ocupación del Suelo de España de Alta Resolución (SIOSEAR) referida a 2017 —*SIOSEAR2017*— y generada por el Instituto Geográfico Nacional (IGN). *SIOSEAR2017* es una cartografía vectorial temática de ocupación y usos del suelo que toma la parcela catastral como unidad de referencia. En cierta forma podríamos decir que *SIOSEAR* es un Catastro temático. La versión de 2017 contiene unos 110 millones de polígonos sobre los que se describe con detalle su cobertura, lo que da idea de la resolución espacial de la base de datos. *SIOSEAR* adopta un modelo de datos muy detallado que diferencia entre coberturas y usos derivada de las recomendaciones de la Directiva Comunitaria INSPIRE sobre la cubierta terrestre (INSPIRE 2013) y los usos del suelo (INSPIRE 2023b). El rótulo SIOSE, ya sea para coberturas o para usos, se elabora a partir de un listado normalizado a nivel europeo sobre coberturas y usos en función de los porcentajes de ocupación, si bien, dada la elevada resolución de la base de datos, muchos de los polígonos tienen cobertura o uso único. La información sobre *SIOSEAR2017* está bien descrita en la documentación técnica que acompaña la base de datos (ETN SIOSE 2022a, 2022b, 2023) y en el capítulo 2 de Goerlich y Mollá (2025) en lo que hace referencia a nuestra aplicación.

La agregación a unidades administrativas —comunidades autónomas (CC. AA), provincias o municipios— requiere de sus contornos en formato vectorial. Estos proceden de la Base de Datos de Líneas Límite (*BDLL*) disponible en el Centro de Descargas del Centro Nacional de Información Geográfica (CNIG) dependiente del IGN (Goerlich y Pérez 2021).

A efectos de validación utilizamos los mismos datos que en Goerlich y Mollá (2025), el padrón georreferenciado de la Comunidad de Madrid a fecha 1 de enero de 2021[3]. Dicha geocodificación está realizada a nivel de coordenada puntual a partir de la dirección

[3] Dicha información fue amablemente facilitada por el **Instituto de Estadística de la Comunidad de Madrid (IEM)** para este tipo de ejercicios.

postal, es decir, la vía y el número de esta, lo que permite cualquier nivel de agregación[4].

Aunque en principio disponemos de coordenadas para los 6.751.251 residentes en Madrid a fecha 1 de enero de 2021, para 68.384 de estos residentes no se dispone de coordenada a partir de su dirección postal, por lo que se les asigna la coordenada del centroide de la sección censal en la que residen. Puesto que estos registros pueden distorsionar la verdadera distribución de la población, simplemente fueron eliminados del conjunto de datos de validación, lo que proporciona una población de referencia de 6.682.867 residentes. Así pues, el 99,0% de la población de padrón está perfectamente georreferenciada a partir del callejero.

Finalmente disponemos de un total de 453.003 coordenadas diferentes con población que oscila entre 1 y 1.395 habitantes, con un promedio de 15 residentes por coordenada, aunque la mediana es de solo 4 residentes, y el tercer cuartil es ligeramente inferior a la media, 14 residentes. La distribución de la población por coordenada es tremendamente asimétrica. Un 72% de las coordenadas tienen asignadas 10 o menos residentes, y solo disponemos de 2 coordenadas con más de 1.000 personas. A partir de este fichero puntual se asignó la población por coordenada al edificio más cercano en *SIOSEAR2017*, sin ninguna restricción adicional por tipo de edificio. La única condición es que el polígono al que se asigna la población debe tener cobertura EDF en *SIOSEAR2017*. Esta es nuestra principal información para validar los resultados, aunque hemos de admitir que a este nivel de detalle es extremadamente difícil conocer la precisión alcanzada si nos mantenemos a nivel de edificio.

Todos los procesos se han implementado en el sistema de cálculo estadístico *R* (R Core Team 2023). En concreto, las librerías de *tidyverse* (Wickham *et al* 2019) para la manipulación y tratamiento de datos –*data wrangling*–, la librería *sf* (Pebesma 2018) para el tratamiento de la información vectorial y las librerías *terra* (Hijmans 2025) y *stars* (Pebesma y Bivand 2023) para la manipulación de la información *ráster*, la librería *centr* (Zomorrodi 2025) para el cálculo de los centroides ponderados y la librería *areal* (Prener y Revord 2019) para los cálculos relacionados con la interpolación por áreas entre capas vectoriales.

[4] El fichero dispone de un registro por individuo, no por coordenada. Disponemos, además, del sexo, la nacionalidad —español/extranjero— y la edad en grandes grupos —menor de 16 años, de 16 a 64 años y de 65 y más años—.

3.
Aspectos técnicos

El proceso de desagregación es idéntico al utilizado en Goerlich y Mollá (2025) y está descrito de forma detallada en el capítulo 5 de dicho trabajo. Solo se introducen algunos detalles menores.

En este caso solo utilizamos 2 capas geométricas de información en lugar de las 3 utilizadas en dicho trabajo:

1. **Capa o geometría de origen**: celdas pobladas de *GEOSTAT2021*, con resolución de 1 km x 1 km, y para las que disponemos de una cifra de población para cada celda.

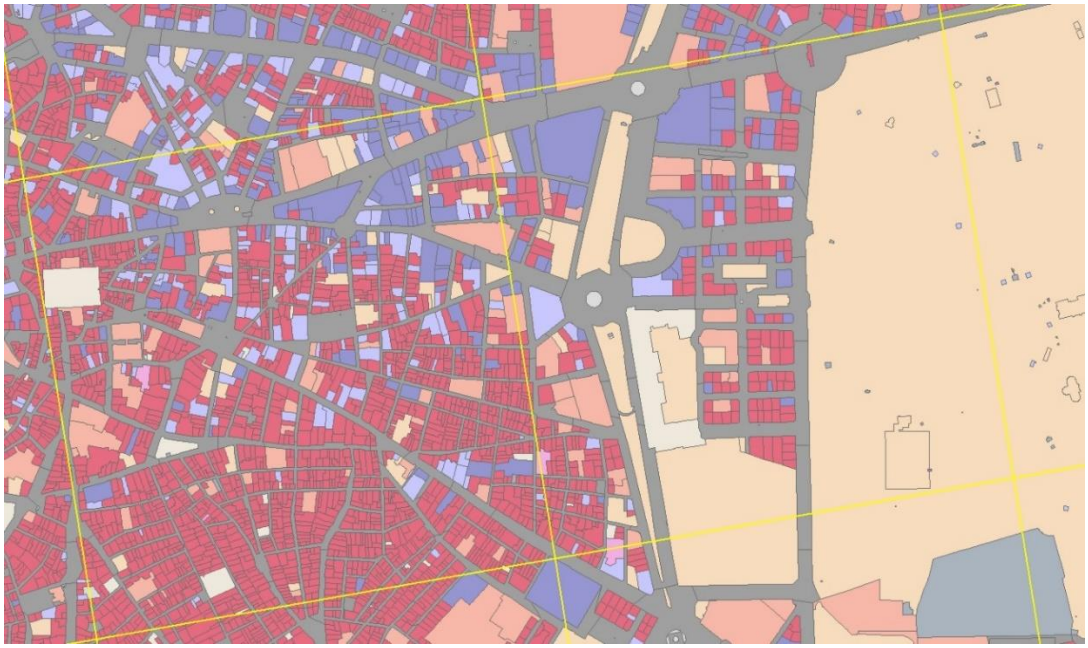
2. **Capa o geometría de destino**: polígonos de *SIOSEAR2017* con cobertura Edificación (*EDF*) y uso residencial (*RESID*), para los que disponemos de determinados atributos, como alturas y superficies, pero no disponemos de una cifra de población. Estos polígonos se superponen de forma incongruente con la capa de origen, y en el ejercicio de Goerlich y Mollá (2025) hicieron la función de geometría intermedia o transicional.

Los **mapas 1 y 2** muestran la información de ambas capas en un entorno urbano, la zona del Paseo del Prado de Madrid —con el Retiro a la derecha y la Puerta del Sol a la izquierda—. El **mapa 1** muestra los polígonos de coberturas —*T_POLIGONOS*— de *SIOSEAR2017* con las celdas de la *grid* con población sobreimpuestas. En **rojo** se muestran los edificios y en **amarillo** las celdas. Disponemos de una cifra de población para cada celda, que deberá distribuirse de forma consistente entre los edificios, pero no todos los edificios albergan población.

Mapa 1. Captura urbana de T_POLIGONOS y GEOSTAT2021: Madrid

Fuente: Instituto Geográfico Nacional, ETN SIOSE (2022a) –SIOSEAR2017–, Instituto Nacional de Estadística –Censo 2021– y elaboración propia.

Mapa 2. Captura urbana de T_USOS y GEOSTAT2021: Madrid

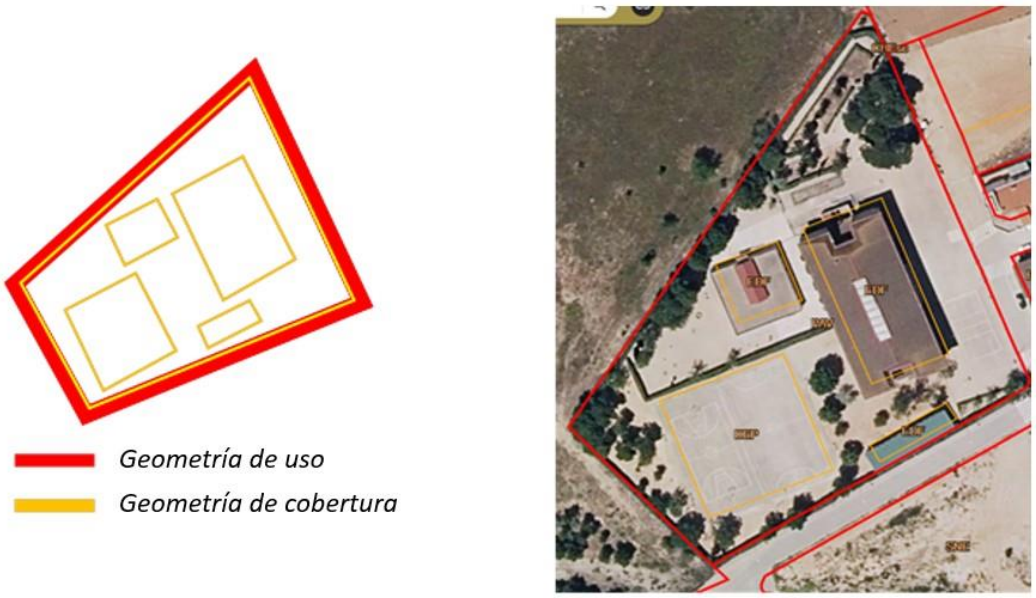


Fuente: Instituto Geográfico Nacional, ETN SIOSE (2022a) –SIOSEAR2017–, Instituto Nacional de Estadística –Censo 2021– y elaboración propia.

El **mapa 2** muestra la misma información desde el punto de vista geográfico, pero sustituye la capa de coberturas por la de usos, *T_USOS*, de *SIOSEAR2017*. En **rojo** se muestran los polígonos de uso residencial. Esta información sirve para filtrar aquellos edificios a los que se asignará población y a los que no. Nuestro algoritmo solo asignará población a los edificios residenciales.

Aunque la información sobre coberturas —*T_POLIGONOS*— y usos —*T_USOS*— de *SIOSEAR2017* están en diferentes geometrías es posible enlazarlas, de forma que sobre la capa de coberturas, que tiene mayor resolución, podemos incorporar la información de usos, que coincide, en principio, con la parcela catastral y tiene una menor resolución. Por otra parte, no existen superposiciones entre las geometrías de usos y las geometrías de coberturas, aunque naturalmente pueden compartir lindes. Así pues, los polígonos de coberturas están siempre contenidos en polígonos de usos, de forma que un polígono de cobertura no puede estar, geométricamente, en dos polígonos de usos. La relación entre ambas tablas es de 1 a muchos a partir del campo único en *T_USOS* que enlaza ambas tablas —*ID_PARCELA*— y la relación entre las geometrías de ambas capas puede observarse visualmente en la figura 1.

Figura 1. SIOSEAR: Relación entre las geometrías de usos y coberturas



Fuente: ETN SIOSE (2022a) –SIOSEAR2017–

De esta forma mantenemos la mayor resolución espacial disponible en la capa de coberturas —*T_POLIGONOS*—, pero incorporando información sobre usos —*T_USOS*—. Como se explica en Goerlich y Mollá (2025) a cada uno de estos polígonos podemos atribuirles un dato de altura de los edificios, resumiendo así, en una sola capa la información más relevante para la desagregación espacial de la población a nivel de edificio. Sobre lo que actualmente no disponemos de información a nivel de edificio, y puede ser especialmente relevante en el caso de España, es la distinción entre viviendas principales —primeras residencias— y viviendas secundarias y vacías. Dado que nuestro objetivo es redistribuir la población residente —lo que se conoce como población nocturna— entonces solo las viviendas principales serían de interés, sin embargo, en la práctica todos los edificios con uso residencial recibirán población, ya que no hay forma de ajustar por la tipología de viviendas —principales *versus* secundarias o vacías— dentro de los edificios. La consecuencia lógica de esto es que tenderemos a dispersar en exceso la población.

La capa de origen —*GEOSTAT2021*— no cubre la totalidad del territorio, sino solo algo más de la quinta parte del mismo, un 22,6%. Los ejercicios de desagregación espacial de la población suelen partir de población por unidades administrativas que son exhaustivas del territorio bajo consideración. Así, por ejemplo, Gallego (2010) parte de poblaciones a nivel municipal y Goerlich y Cantarino (2012, 2013) parten de poblaciones a nivel de sección censal. En nuestro caso, existe una primera acotación espacial de donde reside la población que nos viene dada y que debemos respetar. No habrá población fuera de las celdas habitadas de *GEOSTAT2021*, porque nuestra desagregación debe ser consistente con la *grid* de referencia, una *grid bottom-up*. Naturalmente hay edificios en SIOSEAR*2017* fuera de *GEOSTAT2021*. En nuestro contexto se trata de

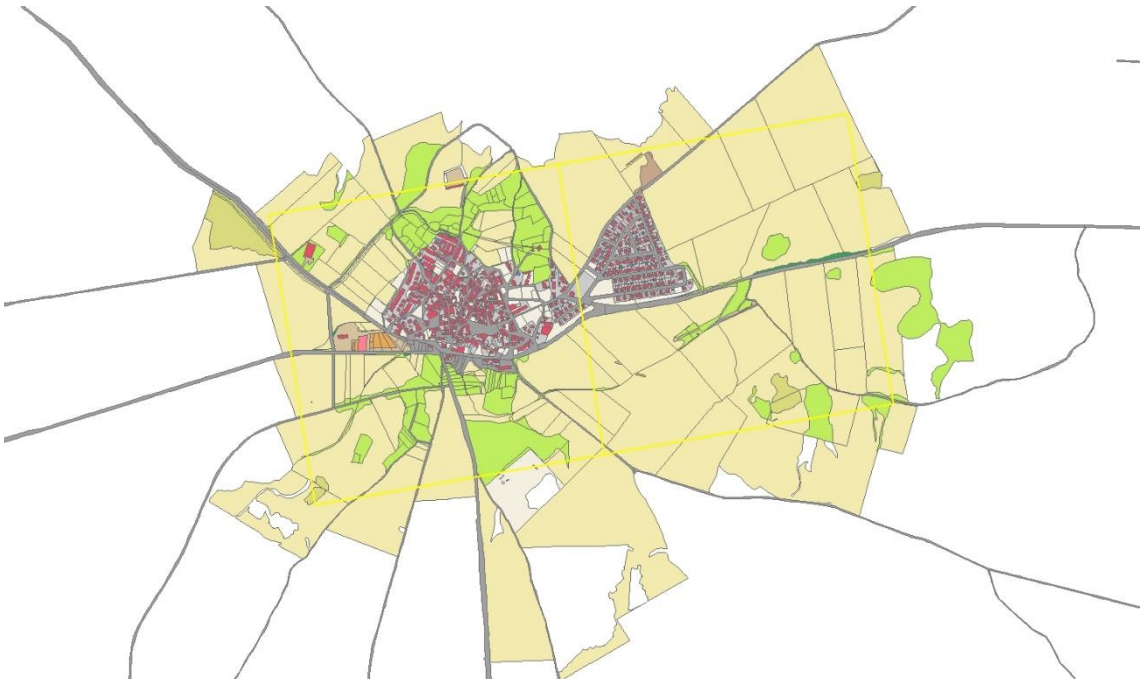
edificios vacíos, en el sentido de que no albergarán población residente por construcción, de la misma forma que no lo hacen en la *grid* censal 2021.

Los **mapas 3 y 4** muestran la misma información que los mapas 1 y 2 pero en un contexto rural, y una vez los polígonos de *SIOSEAR2017* han sido filtrados por las celdas de *GEOSTAT2021*, de forma que solo los edificios con intersección no nula con estas celdas son considerados en el proceso de desagregación, es decir, la población se distribuirá entre los edificios residenciales que se encuentran en las celdas habitadas de *GEOSTAT2021*.

Al objeto de darnos cuenta de la resolución disponible, los **mapas 5 y 6** muestran la misma información que los mapas 1 y 2 pero para una urbanización con viviendas unifamiliares. El **mapa 5** muestra los polígonos de coberturas —*T_POLIGONOS*— de *SIOSEAR2017* con los edificios en **rojo**, mientras que el **mapa 6** muestra los usos, *T_USOS*, de *SIOSEAR2017* con el residencial en **rojo**. La población de cada celda es distribuida a los edificios residenciales del mapa 5, lo que da una idea de la resolución de la capa final obtenida.

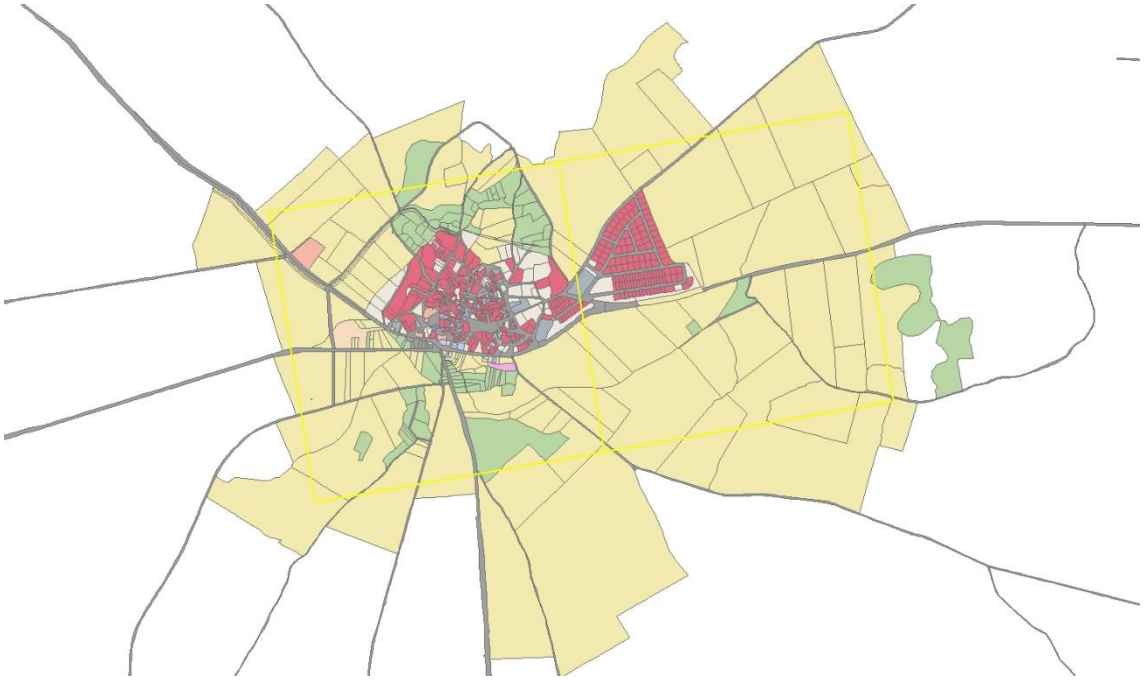
Al igual que en Goerlich y Mollá (2025) el proceso de desagregación elimina como soporte aquellos edificios con atributo específico nave (*EDFvn*) y en ruinas (*EDFer*), aunque mantiene los edificios en construcción (*EDFec*). Todo ello, una vez filtrados los polígonos con edificaciones (*EDF*) residenciales (*RESID*) en *SIOSEAR2017* por las celdas de *GEOSTAT2021*.

Mapa 3. Captura rural de T_POLIGONOS filtrado espacialmente por GEOSTAT2021



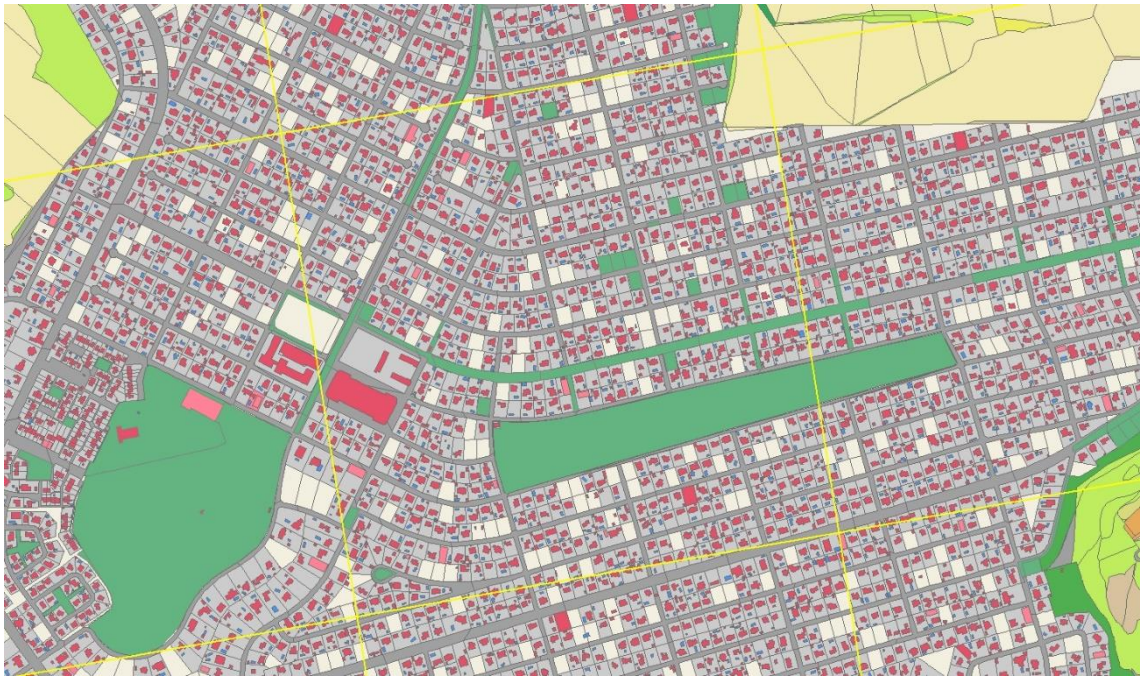
Fuente: Instituto Geográfico Nacional, ETN SIOSE (2022a) –SIOSEAR2017–, Instituto Nacional de Estadística –Censo 2021– y elaboración propia.

Mapa 4. Captura rural de T_USOS filtrado espacialmente por GEOSTAT2021



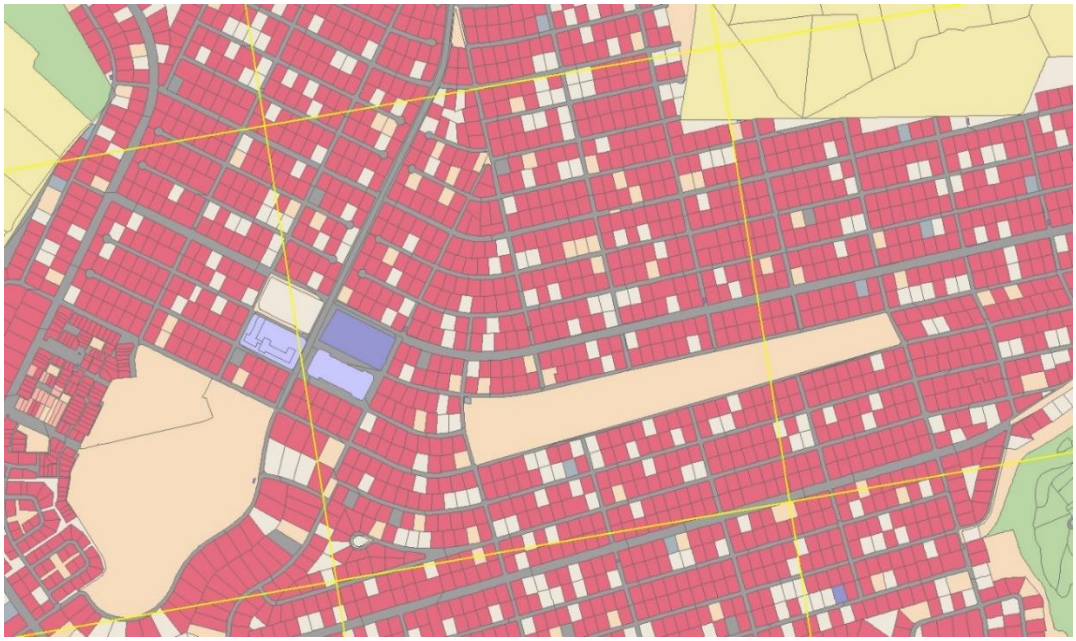
Fuente: Instituto Geográfico Nacional, ETN SIOSE (2022a) –SIOSEAR2017–, Instituto Nacional de Estadística –Censo 2021– y elaboración propia.

Mapa 5. Captura urbana de viviendas unifamiliares de T_POLIGONOS y GEOSTAT2021



Fuente: Instituto Geográfico Nacional, ETN SIOSE (2022a) –SIOSEAR2017–, Instituto Nacional de Estadística –Censo 2021– y elaboración propia.

Mapa 6. Captura urbana de viviendas unifamiliares de T_USOS y GEOSTAT2021



Fuente: Instituto Geográfico Nacional, ETN SIOSE (2022a) –SIOSEAR2017–, Instituto Nacional de Estadística –Censo 2021– y elaboración propia.

Al contrario de lo que sucede en Goerlich y Mollá (2025) la población en la capa de destino no coincidirá con la población en la capa de origen. La razón es que algunas celdas carecen de soporte donde localizar la población, es decir, no hay edificios residenciales en ellas —tablas 5.2 y 5.3 en Goerlich y Mollá (2025)—. En los casos en los que hay edificios, pero su uso no es residencial, entonces distribuiremos la población a los edificios existentes en la celda independientemente de su uso, pero si ni siquiera hay edificios, entonces naturalmente perderemos la población. Estos casos son marginales, ya que el 99.4% de la población de *GEOSTAT2021* reside en celdas que tienen edificios residenciales en ellas según *SIOSEAR2017*.

Finalmente, **dos precisiones técnicas**.

En **primer lugar**, *SIOSEAR* incorpora valores sobre las superficies de todos los polígonos, calculadas en la proyección de distribución del fichero, y porcentajes de ocupación en el caso de que el ROTULO SIOSE incluya varias coberturas. Sin embargo, todas las superficies utilizadas en el proceso de desagregación son de elaboración propia después de la reproyección de *SIOSEAR* a ETRS89-LAEA, de forma que la nueva proyección coincide con la de la *grid*, la capa geométrica de partida. Los porcentajes del ROTULO SIOSE, cuando son necesarios, se aplican sobre estas superficies calculadas.

En **segundo lugar**, puesto que la cifra de población es un número natural, el proceso de desagregación debe respetar esta característica. Así lo hicimos en la generación de la *grid* de 100 m x 100 m, y así debemos hacerlo ahora con mayor motivo. La literatura es silenciosa sobre este punto y, hasta donde nosotros conocemos, solo Goerlich y Cantarino (2012, 2013) y Bastista e Silva y Poelman (2016) ofrecen un proceso de desagregación espacial de la población a números naturales. Pudiera parecer una cuestión estética o trivial, pero no lo es en absoluto. Existe cierta evidencia de que distribuir la población en

reales aumenta la dispersión de esta sobre la capa de destino. Goerlich (2025) genera *grids* de población históricas para todos los años censales desde 1900 hasta 2021. En todos los casos, excepto en 2011, la desagregación se efectúa a números naturales. El motivo es que en 2011 el INE ofreció cifras de población en reales, lo que el propio INE (2023) ha reconocido que no fue una buena idea. En dicho año se observa un pico en el número de celdas habitadas respecto a la tendencia observada, que se revierte en 2021, y que, al menos parcialmente, se puede rastrear a un número no despreciable de celdas con población inferior a 1 habitante, de forma que muchas de ellas desaparecen cuando se efectúa un redondeo a enteros. Algo similar sucede con las *grids* de población de la *Global Human Settlement Layer* (*GHSL*, Freire, MacManus, Pesaresi, Doxsey-Whitfield y Mills 2016) cuyo número de celdas habitadas se reduce considerablemente si redondeamos sus valores a enteros (Goerlich 2025).

Este efecto de aumento artificial de la dispersión cuando la desagregación se efectúa a números reales es especialmente acusado cuando las unidades en la capa de origen son pequeñas en términos demográficos, como sucede en este caso. El 76% de las celdas de *GEOSTAT2021* no superan los 100 habitantes y el 39% no supera los 10 habitantes. También cuando las unidades en la capa de destino son pequeñas y numerosas dentro de cada una de las unidades en la capa de origen, como los mapas anteriores muestran claramente en nuestra aplicación. En este contexto es fácil entender como una estimación en reales tiene a generar una excesiva dispersión de la población. Por ejemplo, consideremos una celda con solo un habitante. Existen 8,486 celdas con esta población en *GEOSTAT2021*. Resulta obvio que este habitante solo puede ocupar una vivienda en la capa de destino, es decir, debe ser asignado a un edificio, pero si hay varios polígonos con edificios en la celda, y no restringimos el resultado a números naturales, ese habitante será repartido entre varios edificios en la capa de destino. La conclusión lógica es que es necesario efectuar una desagregación a números naturales, en lugar de a números reales, aun a costa de aumentar la complejidad y el tiempo de cálculo de los procesos. Con seguridad la distribución espacial de la población conseguida será mucho más realista. Esta distribución a números naturales se efectuó a las celdas de 100 m x 100 m en Goerlich y Mollá (2025), y no se aplicó a la geometría intermedia, pero puesto que esta geometría es ahora la capa de destino debemos aplicarla a este nivel, de forma que la población de cada polígono sea un número natural.

El algoritmo de ajuste a enteros, efectuado a nivel de celda de 1 km x 1 km, es el método de los restos mayores (Cox 1987; Balinski y Rachev 1993, 1997), que es básicamente el implementado por Bastista e Silva y Poelman (2016), y que ya fue utilizado anteriormente por Goerlich y Cantarino (2012, 2013). Nuestra distribución de la población a nivel de edificios será pues una distribución en números naturales.

El resultado de implementar el proceso anterior es una **distribución de la población sobre más de 8 millones de edificios**[5] para un total de algo más de 47 millones de personas[6].

[5] Exactamente 8.654.732 edificios.

[6] Exactamente 47.399.524 personas. Perdemos 1.274 personas respecto a la población del censo 2021, que son aquellas que residen en celdas para las que no hay cobertura *EDF* en *SIOSEAR2017*, ni tampoco polígonos de uso residencial –*RESID*–

En principio, el *output* del proceso es un fichero vectorial poligonal con los polígonos de *SIOSEAR2017* a los que se les asigna población. El **mapa 7** muestra un ejemplo del resultado obtenido para un municipio de La Rioja.

A partir de este resultado poligonal se generó un fichero puntual con los centroides geométricos de dichos polígonos, que resulta más ligero de manipular y más operativo para algunas aplicaciones. Esto es lo más cerca que podemos estar del fichero de **población georreferenciada a nivel de coordenada puntual** al que se hacía referencia al principio del trabajo. Es este fichero el que constituye el objeto de distribución.

Como producto derivado se generó un fichero vectorial puntual de **centroides municipales ponderados por la población** para cada uno de los municipios existentes a fecha del censo 2021. Estas coordenadas representan el punto más representativo del municipio donde situar a la población del mismo en el caso de que se necesiten coordenadas puntuales a nivel municipal[7]. Al igual que sucede con los centroides geométricos (Goerlich 2023) estas coordenadas no tienen por qué situarse dentro del término municipal, a pesar de que este sigue siendo el punto más representativo donde situar a la población del municipio en estos casos. De hecho, esto no sucede en 6 municipios. Si se necesitara un centroide ponderado que cayera necesariamente dentro del término municipal siempre se puede tomar, en estos casos, el centroide del edificio más poblado interior al contorno del municipio.

Mapa 7. Ejemplo de la información obtenida

Fuente: Instituto Geográfico Nacional, ETN SIOSE (2022a) –SIOSEAR2017–, Instituto Nacional de Estadística –Censo 2021– y elaboración propia.

aunque no haya edificios. Los detalles pueden verse en la tabla 5.3 de Goerlich y Mollá (2025).

[7] Unos pocos edificios —29— caen fuera de nuestro vectorial de contornos municipales y todavía menos edificios —7— caen en condominios, a los que no se les asigna población. Por esta razón, para no perder edificios con población estimada en el proceso de generación de los centroides municipales, la asignación de edificios en estos casos se hizo al término municipal más cercano.

4.

Validación: ¿Son fiables nuestras estimaciones?

Esta es la pregunta del millón —💰— en este tipo de trabajos, —😄—. La respuesta es que no lo sabemos con certidumbre.

La precisión de las estimaciones de población a nivel de edificio es extremadamente difícil de evaluar. La razón fundamental es que no hay una fuente de datos independiente que permita contrastar de forma precisa y fiable nuestras estimaciones. El conjunto de validación utilizado en Goerlich y Mollá (2025) y descrito anteriormente —el padrón georreferenciado de la Comunidad de Madrid a fecha 1 de enero de 2021— es solo de utilidad limitada a este nivel de resolución, donde unos pocos metros —muy pocos en algunos casos— pueden representar una asignación a un edificio incorrecta. Esta afirmación es válida tanto para las coordenadas del conjunto de validación, que son asignadas a los edificios de *SIOSE2017*, como para el resultado del proceso de desagregación seguido.

Incluso a nivel de celdas de 1 km x 1 km, donde no deberían haber diferencias entre la *grid* censal del INE y la derivada por agregación a partir de las coordenadas del IEM, se detectan discrepancias apreciables entre *GEOSTAT2021* y los resultados derivados del IEM (Goerlich y Mollá 2025, mapa 1.7). Tanto por la diferencia de fuente para la población, censo *versus* padrón, como por el hecho de que no todas las coordenadas del padrón 2021 del IEM corresponden a una dirección postal, aunque estas hayan sido eliminadas en la construcción del conjunto de validación, como por cuestiones metodológicas relacionadas con los procesos de georreferenciación de la población. Los datos del IEM proceden de la geocodificación a partir de direcciones postales, y sitúan las coordenadas en los puntos de entrada de los edificios de un callejero para el que no disponemos de la cartografía[8], mientras que nuestras estimaciones a nivel de edificio se asignan a polígonos de *SIOSEAR2017*[9] y las estimaciones puntuales se derivan como el centroide geométrico de dichos polígonos.

Aunque todos estos pequeños desajustes de fuentes, junto con la elevada resolución de la información generada, hacen difícil validar los resultados con precisión, si es posible una evaluación aproximada. Empezamos comparando algunos estadísticos descriptivos entre el conjunto de validación y los datos generados, para a continuación ofrecer una medida cuantitativa de la precisión de la información generada.

En nuestro conjunto de validación, para una población de 6.682.867, cuando dejamos caer las coordenadas de padrón sobre los edificios de SIOSEAR*2017* encontramos 385.285 edificios con población. En nuestros datos, para una población de 6.726.482[10], encontramos 523.314 edificios habitados.

[8] La cartografía de los edificios donde se asigna la población de las coordenadas de padrón del IEM proceden de *SIOSEAR2017*.

[9] Cuyo origen es la **cartografía catastral**.

[10] Esta cifra es prácticamente idéntica a la población del censo 2021 para la Comunidad de Madrid, 6.726.640 residentes. La pequeña discrepancia es debida a las celdas borde, es decir, a aquellas que están en los lindes de la Comunidad de Madrid con comunidades vecinas.

Así pues encontramos, en nuestros resultados, un 37% más de edificios habitados que en nuestro conjunto de validación. Examinando la tabulación cruzada para aquellos edificios que tienen población en alguno de los dos conjuntos de datos, observamos que hay 14.705 edificios habitados en el conjunto de validación que no lo están en nuestras estimaciones, lo que representan 187.217 personas del padrón 2021 de Madrid –un 2,8%–, mientras que nuestra desagregación atribuye población a 153.734 edificios que no aparecen como habitados en el conjunto de validación, lo que representan 694.860 personas del censo 2021 de Madrid –un 10,3%–.Esta excesiva dispersión de la población se debe, en gran parte, a la falta de información sobre segundas residencias y viviendas vacías. El número medio de personas por edificio en el conjunto de validación es de 17 residentes, mientras que en los datos estimados es solo de 13. Esta diferencia también la encontramos para el número mediano de personas por edificio, aunque las discrepancias son menos abultadas, 4 personas para el conjunto de datos de validación frente a 3 en los datos estimados.

La tabla 1 muestra, para determinados intervalos de tamaño, las celdas y la población estimada en cada uno de ellos, tanto para el conjunto de datos de validación, procedentes del IEM, como para los datos desagregados a nivel de edificio en este trabajo. En el conjunto de datos de validación el 10,7% de los edificios tiene un solo residente, mientras que este porcentaje asciende al 18,3% en nuestras estimaciones. En el otro extremo de la distribución encontramos el resultado contrario, en el conjunto de datos de validación un 3,2% de los edificios tienen más de 100 residentes, mientras que en nuestras estimaciones esto solo sucede en un 2,1% de los edificios.

Tabla 1. Edificios habitados y población por intervalos de tamaño

Intervalo	Datos de Validación (IEM)				Datos de Estimados			
	Edificios	%	Población	%	Edificios	%	Población	%
1	41.024	10,7	41.024	0,6	95.512	18,3	95.512	1,4
2	62.416	16,3	124.832	1,9	92.414	17,7	184.828	2,7
3	55.655	14,5	166.965	2,5	90.892	17,4	272.676	4,1
4	61.058	15,9	244.232	3,7	61.888	11,8	247.552	3,7
5	24.943	6,5	124.715	1,9	31.344	6,0	156.720	2,3
6 o 7	20.180	5,3	128.943	1,9	25.509	4,9	162.766	2,4
[8, 10]	13.869	3,6	122.559	1,8	17.071	3,3	151.757	2,3
[11, 15]	14.516	3,8	187.194	2,8	20.424	3,9	264.122	3,9
[16, 25]	23.702	6,2	479.607	7,2	27.168	5,2	541.596	8,1
[26, 50]	33.252	8,7	1.217.612	18,2	32.595	6,2	1.167.463	17,4
[51, 100]	20.312	5,3	1.408.948	21,1	17.333	3,3	1.220.697	18,1
[101, 200]	8.185	2,1	1.114.679	16,7	7.308	1,4	1.006.797	15,0
Más de 200	4.173	1,1	1.321.557	19,8	3.856	0,7	1.253.996	18,6
Total	**383.285**	**100,0**	**6.682.867**	**100,0**	**523.314**	**100,0**	**6.726.482**	**100,0**

Fuente: IEM -Padrón 2021-, INE (2023), IGN -SIOSEAR2017-, Goerlich y Mollá (2025) y elaboración propia

Algo similar ocurre con la población. La población residente en edificios con hasta 5 habitantes es el 10,5% del total en el conjunto de datos de validación, mientras que asciende al 14,2% en nuestras estimaciones. Por el contrario, el 57,5% de la población reside en edificios con más de 50 residentes en el conjunto de validación, porcentaje que se ve reducido al 51,8% en nuestras estimaciones.

Un estadístico sencillo para medir la similitud entre dos estructuras porcentuales es

$$\zeta = \frac{1}{2}\Sigma_j \left| S_{1j} - S_{2j} \right| \qquad (1)$$

donde S_{1j} y S_{2j} representan las estructuras porcentuales bajo comparación y *j* indexa el número de elementos en dicha estructura. Este estadístico, ζ, varía entre 0, si ambas estructuras porcentuales son idénticas, y 1, en el caso de máxima discrepancia, cuando las estructuras porcentuales no se solapan.

El valor de ζ, expresado en términos porcentuales, 100× ζ, es de 11.9% para las distribuciones relativas de los edificios de la tabla 1, y algo inferior, 6.6%, cuando comparamos las estructuras porcentuales de las poblaciones por tamaños demográficos de los edificios.

Una generalización del estadístico de similitud (1) permite medir la discrepancia entre ambos conjuntos de datos, el de validación y nuestras estimaciones, con una interpretación relativamente sencilla. Dicha métrica es utilizada habitualmente en esta literatura (Goerlich y Cantarino 2012, 2013; Goerlich 2025; Goerlich y Mollá 2025) y parte de un indicador absoluto de discrepancias definido a nivel de edificio, que es nuestra unidad geográfica de destino, como

$$\Delta = \Sigma_j \left| P_j - P_j^{ref} \right| \qquad (2)$$

donde *j* indexa los edificios, y el superíndice *ref* se refiere a la población utilizada como referencia en el conjunto de datos de validación.

El valor de Δ oscila entre 0, cuando la distribución de la población estimada a nivel de edificio es idéntica a la de la población de referencia, y 2 veces la población a distribuir, cuando toda la población está en edificios diferentes en ambos conjuntos de datos y no hay solapamientos. Obsérvese que se trata de un indicador de desajuste espacial un tanto ingenuo, ya que la contribución a la discrepancia es independiente del error en la distancia que cometemos al localizar incorrectamente la población, es decir, la contribución es la misma si a una persona la situamos en el edificio contiguo de donde debe estar o en la parte diametralmente opuesta de nuestra área de análisis. Aunque la distancia es esencial en un indicador de discrepancias con contenido geográfico (O´Sullivan y Wong 2007) mantendremos (2) como métrica a utilizar, tanto por su fácil interpretabilidad una vez normalizado, como por comparabilidad con otros estudios.

Dada la dependencia de (2) respecto al tamaño de la población resulta útil re-escalarlo al intervalo [0, 1], y definir el indicador de discrepancias relativo como

$$\delta = \frac{\Delta}{2\times\Sigma_j P_j} = \frac{\Sigma_j \left| P_j - P_j^{ref} \right|}{2\times\Sigma_j P_j} \qquad (3)$$

De esta forma 100× δ puede interpretarse como el porcentaje de población que situamos de forma incorrecta sobre el territorio, donde la precisión de la incorrección en la localización viene determinada por las unidades de análisis, en nuestro caso edificios.

Aunque las poblaciones del conjunto de validación —padrón 2021 georreferenciado del IEM— y la que hemos desagregado

—*grid* censal del INE procedente de la intersección de *GEOSTAT2021* con el fichero provincial de Madrid de *SIOSEAR2017*— son muy similares —6.682.867 y 6.726.482 respectivamente—, no son exactamente las mismas, $\sum_j P_j \neq \sum_j P_j^{ref}$. En estas condiciones, la normalización de Δ para que esté acotado entre 0 y 1 es la siguiente

$$\delta' = \frac{\Delta - \Delta_{min}}{\Delta_{max} - \Delta_{min}} \qquad (4)$$

La razón es sencilla. El valor máximo de (2) es 2 veces la población cuando la población a desagregar procedente de la *grid* y la de referencia son exactamente las mismas. En caso contrario, el valor máximo resulta ser la suma de las poblaciones, $\Delta_{max} = \sum_j P_j + \sum_j P_j^{ref}$. El valor mínimo ahora no es cero, porque si situamos toda la población correctamente en los edificios del conjunto de validación, no todos los sumandos en (2) se anulan. El valor mínimo resulta ser la diferencia (en valor absoluto) de las poblaciones, $\Delta_{min} = \left|\sum_j P_j - \sum_j P_j^{ref}\right|$. En estas circunstancias la normalización que acota el índice Δ en el intervalo [0, 1] es precisamente (4), lo que implica tanto un cambio de origen como de escala.

La métrica (4), δ', expresada en términos porcentuales, $100\times\delta'$, toma un valor del **22,8%** para nuestro ejercicio. Lo que podemos interpretar como que casi el 23% de la población es asignada a un edificio al que no le corresponde según el conjunto de datos de validación. Sin duda alguna no es un error bajo, pero tampoco es excesivamente elevado. Goerlich y Mollá (2025) obtienen un error del 16,9% en la desagregación de *GEOSTAT2021* a una *grid* con celdas de 100 m x 100 m, lo que representa una resolución mucho menor que a nivel de edificio[11]. El coeficiente de correlación entre ambas series resultó ser de 0,86, lo que representa una concordancia razonablemente elevada.

Este error no sería objeto de gran preocupación si se debiera a personas que son asignadas a edificios colindantes o muy cercanos en vez de al que realmente les corresponde. Goerlich y Mollá (2025) encuentran que el error disminuye rápidamente conforme disminuimos la resolución de la *grid* de destino[12]. Investigar esta cuestión requiere disminuir la resolución de la capa de destino —edificios—. Para ello rasterizamos tanto el conjunto de datos de validación como las estimaciones obtenidas a nivel de edificio a una *grid* con tamaño de celda de 50 m x 50 m, y una vez hemos obtenido los centroides de los edificios en ambos casos. El valor del estadístico de discrepancia, $100\times\delta'$100×, es en este caso del 17,7%[13].

[11] 1.324.147 puntos donde localizar la población a partir de la *grid* con celdas de 100 m x 100 m frente a 8.654.732 puntos donde localizar la población a partir de la asignación de la misma a los edificios de *SIOSEAR2017*.

[12] De un error del 16,9% en celdas de 100 m x 100 m se pasa a uno del 10,9% en celdas de 200 m x 200 m y de 6,1% en celdas de 500 m x 500 m.

[13] Resulta ilustrativo examinar como disminuye el error conforme disminuimos la resolución de la capa de destino, es decir, conforme aumentamos el tamaño de celda en el ejercicio de rasterificación que acabamos de mencionar. Para un tamaño de celda de 25 m x 25 m el error disminuye poco respecto a la validación efectuada directamente a nivel de edificio, siendo de un 20,8%. Ello es razonable, puesto que este tamaño de celda representa solo 625 m² y agrupa pocos edificios contiguos. Para un tamaño de celda de 50 m x 50 m el error disminuye hasta el 17,7%. Cuando aumentamos el tamaño de celda a 75 m x 75 m el error disminuye hasta el 15,4% y para un tamaño de celda de 100 m x 100 m baja hasta el 13,7%.

Resulta llamativo que para la misma resolución que la utilizada en Goerlich y Mollá (2025), utilizando los mismos datos y el mismo algortimo de desagregación espacial, obtengamos ahora un error de validación significativamente menor –del orden de 3 puntos porcentuales inferior en nuestra aplicación–. Es cierto que hay algunas diferencias de implementación. Por ejemplo, el conjunto de validación en Goerlich y Mollá (2025) se construyó dejando caer directamente las coordenadas de padrón sobre las celdas de la *grid* de 100 m x 100 m, mientras que ahora les hemos asignado el edificio más cercano en *SIOSEAR2017*, hemos calculado el centroide de dichos edificios y hemos rasterizado estos puntos a una *grid* con las mismas características –Sistema de Referencia de Coordenadas (*CRS*), origen y resolución– que la utilizada en Goerlich y Mollá (2025). Pero todas estas diferencias son pequeñas y no parece que sean

El **resumen** es que nuestra desagregación no es perfecta, pero supone una representación tremendamente granular de la distribución de la población, muy flexible en términos de la agregación por áreas geográficas cualesquiera y, en nuestra opinión, muy útil para análisis locales 😁. Los datos deben ser entendidos de forma similar a los de una encuesta. Un solo registro no es representativo de la población, pero tomados en conjunto pueden ser agregados en múltiples direcciones bajo diferentes criterios. Algo similar ocurre con la población a nivel de edificio estimada en este trabajo. Para un edificio particular la estimación puede no ser razonable, pero para pequeñas áreas, que nada tengan que ver con los lindes administrativos a partir de los cuales se recoge la información demográfica, nuestras estimaciones pueden ser muy razonables.

responsables de una diferencia de 3 puntos porcentuales en los errores de validación en ambos ejercicios.

La principal diferencia de implementación entre el algoritmo utilizado en Goerlich y Mollá (2025) y el empleado en este trabajo radica en el momento en el que se efectúa el *downscaling* y el redondeo a enteros. En Goerlich y Mollá (2025) la desagregación de la población de cada celda de 1 km x 1 km se efectúa a partir del volumen residencial edificado en las celdas de 100 m x 100 m y el redondeo se efectúa a nivel de celda de 100 m x 100 m, que es la resolución a la que deseamos una cifra de población en enteros. Aunque la geografía intermedia son los edificios residenciales de *SIOSEAR2017* no se llegó a estimar una cifra de población por edificio. Este trabajo lo que hace es, precisamente, aprovechar la geografía intermedia para asignar una cifra de población a cada edificio, y el redondeo se hace a nivel de edificio. Nuestra intuición es que la mejora en el estadístico de discrepancia en este caso, respecto a lo que se obtiene en Goerlich y Mollá (2025) se debe, en gran parte, al momento en el que se efectúa el redondeo a números naturales. ¡Una intuición que merecería la pena investigar, 👍!

En apoyo de nuestra intuición observamos que cuando aumentamos el tamaño de las celdas a las de *GEOSTAT2021*, 1 km x 1 km, entonces el error que encontramos es de solo 1,3%, idéntico al obtenido por Goerlich y Mollá (2025, capítulo 2) cuando comparan el conjunto de datos de validación con las celdas de *GEOSTAT2021* para Madrid.

Si esta intuición es cierta, parece que podríamos mejorar la calidad de la *grid* de Goerlich y Mollá (2025) simplemente agregando el fichero puntual de este trabajo a la *grid* de 100 m x 100 m. Naturalmente para que la *grid* fuera completa deberíamos recuperar las 1.274 personas residentes en las 441 celdas de *GEOSTAT2021* para las que no existe soporte en *SIOSEAR2017*, y que ahora hemos perdido. Esto no hace sino aumentar la importancia de en qué momento efectuar el redondeo a números naturales en los procesos de desagregación de la población, algo sobre lo que se enfatiza en **Goerlich (2025)** y Goerlich y Mollá (2025).

5.
Base de datos e información ofrecida

Este apartado describe la información disponible a partir de este trabajo y que puede ser descargada de **zenodo**[14]. Toda la información geográfica se distribuye en coordenadas geográficas y sistema de referencia geodésico ETRS89 —EPSG 4258—[15].

El *output* básico del trabajo es un fichero de texto plano, POBEDFSIOSEAR2017.csv, con 3 campos:

- **CodProv**: Código de provincia.
- **ID_POLYGON**: Identificador de polígono en la tabla *T_POLIGONOS* –o *T_VALORES*– de *SIOSEAR2017*.
- **POB**: Población del polígono.

Esta tabla deber ser enlazada con la tabla *T_POLIGONOS* —o *T_VALORES*— de *SIOSEAR2017* mediante el campo **ID_POLYGON** para representación gráfica y/o análisis adicional. El campo **CodProv** se corresponde con el fichero provincial de *SIOSEAR2017*, ya que esta información se distribuye por provincias, y se facilita para no tener que leer *SIOSEAR2017* entero al efectuar la asignación de la población a los polígonos en el caso de que desee realizarse un análisis local.

Esta misma información está disponible en un fichero **vectorial de puntos** —centroides de los polígonos con edificios a los que se ha asignado la población— en formato *GeoPackage* y con los mismos campos que el fichero de texto plano, POBEDFSIOSEAR2017.gpkg, en la capa POBEDFSIOSEAR2017.

Adicionalmente, en POBEDFSIOSEAR2017.gpkg hay una **capa puntual de centroides municipales ponderados por la población a nivel de edificio** —Centroides— para los municipios existentes en el censo 2021 derivada de la capa puntual de población por edificio y los contornos municipales (Goerlich y Pérez 2021).

[14] Información adicional elaborada, no disponible para descarga en **zenodo**, puede obtenerse si se solicita al **autor**.

[15] La razón de utilizar coordenadas geográficas es que *SIOSEAR2017* se distribuye con proyección UTM en el huso correspondiente, además de que Canarias se distribuye en sistema de referencia geodésico REGCAN, que es compatible con ETRS89 –y a nuestros efectos, en realidad, idénticos–. Por tanto, no era posible generar un fichero geográfico proyectado único que fuera compatible con todo *SIOSEAR2017*, por lo que se decidió mantener coordenadas geográficas en la información distribuida de esta naturaleza.

6.

Comentarios finales

Este trabajo ofrece una estimación de la población por edificio consistente con la población del censo 2021 para toda España y derivada a partir de la *grid* de población censal 2021 del INE —*GEOSTAT2021*— y el Sistema de Información de Ocupación del Suelo de España de Alta Resolución (SIOSEAR) referido a 2017 —*SIOSEAR2017*—. El resultado es una distribución de la población tremendamente granular, con algo más de 8.6 millones de puntos —edificios— con población, y que además es consistente con la *grid* censal 2021 (INE 2023).

Los ejercicios de validación, aunque limitados, avalan la utilidad de los resultados obtenidos, 👍.

Referencias

Balinski, M. L. y Rachev, S. T. (1993) "Rounding Proportions: Rules of Rounding". *Numerical Functional Analysis and Optimization* 14, 5-6, 475-501. doi: 10.1080/01630569308816535.

Balinski, M. L. y Rachev, S. T. (1997) "Rounding Proportions: Methods of Rounding". *Mathematical Scientist* 22, 1-26.

Batista e Silva, F. y Poelman, H. (2016) *Mapping population density in Functional Urban Areas*. JRC Technical Reports. Joint Research Centre. European Commission. EUR 28194 EN.

Batista e Silva, F.; Poelman, H.; Martens, V. y Lavalle, C. (2013) *Population Estimation for the Urban Atlas Polygons*. JRC Technical Reports. Joint Research Centre. European Commission. EUR 26437 EN.

Cox, L. H. (1987) "A Constructive Procedure for Unbiased Controlled Rounding". *Journal of the American Statistical Association* 82, 398, (junio), 520-524. doi: 10.2307/2289455.

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

ETN SIOSE (Equipo Técnico Nacional del Sistema de Información de Ocupación del Suelo en España) (2022a) "Documento Técnico SIOSE AR". Sistema de Información de Ocupación del Suelo en España de Alta Resolución. Versión 1.3. D. G. Instituto Geográfico Nacional. Subdirección General de Cartografía y Observación del Territorio. Servicio de Ocupación del Suelo (23 marzo 2022).

ETN SIOSE (Equipo Técnico Nacional del Sistema de Información de Ocupación del Suelo en España) (2022b) "Tablas de Usos, Coberturas y Atributos". Sistema de Información de Ocupación del Suelo en España de Alta Resolución. Versión 1.0. D. G. Instituto Geográfico Nacional. Subdirección General de Cartografía y Observación del Territorio. Servicio de Ocupación del Suelo (24 marzo 2022).

ETN SIOSE (2023) "Estructura de la base de datos SIOSE AR". Sistema de Información de Ocupación del Suelo en España de Alta Resolución. Versión 3.3. D. G. Instituto Geográfico Nacional. Subdirección General de Cartografía y Observación del Territorio. Servicio de Ocupación del Suelo (30 marzo 2023).

Freire S.; MacManus K.; Pesaresi M.; Doxsey-Whitfield E. y Mills J. (2016) "Development of new open and free multi-temporal global population grids at 250m resolution". Geospatial Data in a Changing World. Association of Geographic Information Laboratories in Europe (AGILE), JRC100523. ISBN: JRC100523.

Gallego, F. J. (2010) "A population density grid of the European Union". *Population & Environment* 31, 6, (julio), 460-473. doi: 10.1007/s11111-010-0108-y.

Goerlich, F. J. (2016) "Una aproximación volumétrica a la desagregación espacial de la población combinando cartografía temática y datos LIDAR". *Revista de Teledetección* 46, (junio), 147-163. ISSN: 1133-0953. EISSN: 1988-8740. doi: 10.4995/raet.2016.4710.

Goerlich, F. J. (2023) *¿Dónde está Wally? –Como y donde situar a la población–* On-line. Versión: 30/03/2023.

Goerlich, F.J. (2025) *HIPGDAC-ES: historical population grid data compilation for Spain (1900-2021)*. *Scientific Data* 12, 280. doi: 10.1038/s41597-025-04533-8.

Goerlich, F. J. y Cantarino, I. (2012) *Una grid de densidad de población para España*. Informes. Economía y Sociedad. Fundación BBVA (noviembre).

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

Goerlich, F. J. y Mollá, S. (2025) *La población española en alta resolución. De la grid de 1km x 1km a una de 100m x 100m. Metodología, análisis y aplicaciones*. Bilbao: Fundación BBVA, en prensa.

Goerlich, F. J. y Pérez, P. (2021) *LAU2boundaries4spain: R package providing LAU2 (municipalities) data geometries for Spain for 2002-2021*. ROpenSpain. Repositorio en GitHub.

Grippa, T.; Linard, C.; Lennert, M.; Georganos, S.; Mboga, N.; Vanhuysse, S.; Gadiana, A. y Wolff, E. (2019) "Improving Urban Population Distribution Models with Very-High Resolution Satellite Information". *Data* 4, 1, 13. doi: 10.3390/data4010013.

Hijmans R. (2025) *terra: Spatial Data Analysis*. *R* package version 1.8-21.

INE (2023) "Censos de Población y Viviendas 2021. Metodología. Versión provisional." Instituto Nacional de Estadística. Subdirección General de Estadísticas Demográficas. Junio.

INSPIRE (2013) *D2.8.II.2 Data Specification on Land Cover - Technical Guidelines* v3.1.0 INSPIRE Infrastructure for Spatial Information in Europe. Comisión Europea.

INSPIRE (2023a) *D2.8.I.2 Data Specification on Geographical Grid Systems - Technical Guidelines* v3.1 INSPIRE Infrastructure for Spatial Information in Europe. Comisión Europea.

INSPIRE (2023b) *D2.8.III.4 Data Specification on Land Use - Technical Guidelines* v3.1.0 INSPIRE Infrastructure for Spatial Information in Europe. Comisión Europea.

Lwin, K. y Murayama, Y. (2009) "A GIS approach to estimation of building population for micro-spatial analysis". *Transactions in GIS* 13, 4, 401-414. doi: 10.1111/j.1467-9671.2009.01171.x.

Lwin, K. y Murayama, Y. (2011) "Estimation of Building Population from LIDAR Derived Digital Volume Model". En Y. Murayama y R. B. Thapa (Eds.), *Spatial Analysis and Modelling in Geographical Transformation Process: GIS-based Applications* 87-98. doi: 10.1007/978-94-007-0671-2.

O´Sullivan, D. y Wong, D. W. S. (2007) "A surface-based approach to measuring spatial segregation". *Geographical Analysis* 39, 2, (abril), 147-168. doi: 10.1111/j.1538-4632.2007.00699.x.

Pebesma, E. (2018) "Simple Features for R: Standardized Support for Spatial Vector Data". *The R Journal* 10, 1, 439-446. doi: 10.32614/RJ-2018-009.

Pebesma, E. & Bivand, R. (2023) *Spatial Data Science: With Applications in R*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016

Prener, C. & Revord, C. (2019) "areal: An R package for areal weighted interpolation". *Journal of Open Source Software* 4, 37, 1221. doi: 10.21105/joss.01221.

R Core Team (2023) *R: A Language and Environment for Statistical Computing*. Vienna: *R* Foundation for Statistical Computing.

Schug, F.; Frantz, D.; van der Linden, S. y Hostert, P. (2021) "Gridded population mapping for Germany based on building density, height and type from Earth Observation data using census disaggregation and bottom-up estimates". *PLoS ONE*. 16-3. doi: 10.1371/journal.pone.0249044.

Steinnocher, K.; De Bono, A.; Chatenoux, B.; Tiede, D. y Wendt, L. (2019) "Estimating urban population patterns from stereo-satellite imagery". *European Journal of Remote Sensing* 52, 12-25. doi: 10.1080/22797254.2019.1604081.

Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L. D.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; Kuhn, M.; Pedersen, T. L.; Miller, E.; Bache, S. M.; Müller, K.; Ooms, J.; Robinson, D.; Seidel, D. P.; Spinu, V.; Takahashi, K.; Vaughan, D.; Wilke, C.; Woo, K. y Yutani, H. (2019) "Welcome to the tidyverse". *Journal of Open Source Software* 4, 43, 1686. doi: 10.21105/joss.01686.

Zomorrodi, R. (2025) *centr: Weighted and Unweighted Spatial Centers*. *R* package version 0.2.2.9000


Event
Event Budget for Event : EXPENSES
Actual Cost Breakdown
Estimated vs. Actual
Site
Decorations
Publicity
Miscellaneous
Refreshments
Program
Prizes
$0.00
$10.00
$20.00
$30.00
$40.00
$50.00
$60.00
$70.00
$80.00
$90.00
$100.00
$1,100.00
$1,000.00
$2,300.00
$2,500.00
$1,500.00
$2,000.00
10%
19%
7%
7%

Ivie