Distribución espacial de las características de la población de Galicia por cuadrícula de 1km^2

METODOLOGÍA

Normativa y organización

En virtud del Convenio sobre cooperación estadística e intercambio de información entre el Instituto Nacional de Estadística (INE) y el Instituto Gallego de Estadística (IGE), firmado en mayo de 2020, el INE remite periódicamente al IGE:

- Estadística del Padrón Continuo: fichero final de la Estadística del Padrón Continuo con microdatos de población ajustada la cifras oficiales a 1 de enero de cada año, incluyendo nombre, apellidos, dirección, código postal, código de vía y número de hoja padronal para el ámbito de la Comunidad Autónoma (petición de carácter continuo); en este documento nos referiremos a él cómo Padrón estadístico
- Padrón Municipal Continuo: coordinación de padrones municipales (actividad no incluida en el Plan Estadístico Nacional): microdatos procedentes de las descargas nominales a 1 de enero y 1 de julio de cada año para el ámbito de la Comunidad Autónoma (petición de carácter continuo); en el que sigue nos referiremos a él cómo Padrón continuo

El Real Decreto 1314/1984, de 20 de junio , por el que se regula la estructura y competencias de la Tesorería General de la Seguridad Social, atribuye a este organismo la competencia relativa a la inscripción de empresas y la afiliación, altas y bajas de las personas trabajadoras.
En abril de 2011, el Instituto Galego de Estatística (IGE) firmó un convenio de colaboración en materia estadística con la Tesorería General de la Seguridad Social y con el Instituto Social de la Marina (ISM) por el cual recibirá trimestralmente la siguiente información extraída del fichero general de afiliación:

- fichero de afiliaciones en alta laboral
- fichero de cuentas de cotización en alta o baja producida en los tres meses

El Real Decreto 2583/1996, de 13 de diciembre, de estructura orgánica y funciones del Instituto Nacional de la Seguridad Social (INSS) le atribuye la gestión y funcionamiento del Registro de Prestaciones Sociales Públicas al INSS.
En julio de 2011 se firmó un convenio de colaboración entre el INSS, el ISM y el IGE en materia estadística (BOE núm. 219 de 12 de septiembre de 2011). En virtud de este convenio, el IGE recibirá anualmente de forma telemática información de las pensiones gestionadas por

el INSS y por el ISM que figuran en el Registro de Prestaciones Socias Públicas, del cuál es titular el INSS.

Introducción

En la actual organización administrativa del Estado español, los municipios constituyen las unidades administrativas menores en las que se divide el territorio y que tienen asignados límites precisos. Por esta razón, los estudios cuyo objetivo es la localización de la población suelen descender hasta el nivel municipal (Goerlich y Mas, 2009). De todas formas, diversos autores (Reher, 1994; Rúa y otros, 2003) recomendaron que para el estudio del asentamiento de la población sobre el territorio esta división es claramente insuficiente y es necesario aumentar la resolución geográfica de análisis.

Además, el tratamiento histórico dado a las direcciones postales conducía a la asignación de divisiones territoriales, como es el caso de la población en municipios, entidades colectivas o secciones censales. El efecto final de esta asignación es que los resultados estadísticos solo pueden obtenerse respeto de estas zonas o sus agregados, produciéndose una pérdida de información no deseable. Desde el punto de vista geográfico-estadístico, la división del territorio es relevante si los dominios de la partición son homogéneos, pero en el caso de los municipios y entidades colectivas, estamos ante una situación de falta de homogeneidad de superficie, población o concentración de unidades de producción.
Teniendo en cuenta esta problemática, hace unos años, algunos institutos de estadística de la Unión Europea con el apoyo de Eurostat lanzaron el European Forum of Geography and Geostatistics (EFGS), con el objetivo de armonizar estadísticas europeas sobre la base de una cuadrícula de 1 km de lado y en un sistema geodésico de referencia común. Este proyecto realiza tareas de estimación de la distribución de celdas de población para una amplia gama de países en todo el continente europeo.

Además, la creciente demanda de estadísticas con un elevado detalle territorial para un análisis espacial más preciso se concreta normativamente en el Reglamento (UE) 2017/2391 del Parlamento Europeo y del Consejo de 12 de diciembre de 2017 por el que se modifica el Reglamento (CE) num. 1059/2003 en el que respeta a las tipologías territoriales (Tercet), publicado en el Diario Oficial de la Unión Europea de 29 de diciembre de 2017. Este reglamento señala que “debe aplicarse un sistema de mallas estadísticas para calcular y atribuir los tipos territoriales a las regiones y zonas en cuestión, ya que dichos tipos dependen de la distribución y de la densidad de la población en celdas de malla de un kilómetro cuadrado”. Este nuevo sistema zonal puede permitir combinar el máximo detalle territorial en la difusión con la preservación de la confidencialidad estadística.

Teniendo en cuenta todo el anterior, el objetivo de esta actividad es representar a la población de Galicia en un mapa buscando la homogeneidad espacial y un amplio desglose territorial. Concretamente se realiza esta actividad para analizar las posibilidades de representación de la información en unidades territoriales reducidas y con una unidad de observación homogénea. Se empleará una malla regular formada por celdas cuadradas de 1km de lado, en la que se pretende representar de manera **aproximada** la población de Galicia obtenida a partir de la explotación de diversos registros administrativos disponibles en el IGE.
Para generar la malla regular, se siguen las indicaciones derivadas de los experimentos realizados por el proyecto Geostat (proyecto ESSnet Geostat) del EFGS, que desarrolla la generación de una malla formada por celdas de 1 km de lado, empleando el mismo sistema de referencia espacial para toda Europa. Las celdas están codificadas con un sistema estándar que sigue las directrices de la Directiva INSPIRE y fueron generadas por el Instituto de Estudios del Territorio (IET).

Por otra parte, los registros administrativos constituyen una fuente vital de información para los institutos de estadística pública. El aprovechamiento de estas bases de datos reduce la necesidad de recurrir a otras fuentes, como los censos y las encuestas; este hecho trae aparejado la disminución, por una parte, de la carga sobre la población encuestada y, por otra, los costes que la elaboración de estas operaciones suponen para los institutos de estadística. Además, si los datos por los que se interroga a los ciudadanos ya están disponibles en registros administrativos, no tiene sentido tratar de volver a obtenerlos por medio de un censo o de una encuesta.

No obstante, no todo son ventajas a la hora de trabajar con registros administrativos. Hay que tener presente que la finalidad para la que se crean no es estadística, por lo que los conceptos y definiciones incluidos en ellos rara vez coincidirán con los de las estadísticas oficiales. A La hora de explotarlos estadísticamente habrá que tener en cuenta también la normativa que está detrás de la creación y gestión de los mismos, y los posibles cambios que afecten a dicha normativa.

En el IGE comenzamos con la tarea de fusionar registros administrativos de los que disponemos, con la finalidad de crear un sistema de información que contenga datos socioeconómicos de la población de Galicia. Estos trabajos comenzaron en el año 2019 con la ejecución de la actividad de interés estadístico “Desarrollo de una base de datos sociodemográfica empleando diversas fuentes administrativas y Estadísticas”, disponible en el Decreto 165/2018, de 27 de diciembre , por lo que se aprueba el Programa estadístico anual de la Comunidad Autónoma de Galicia para el año 2019.

El registro de partida en la fusión es el Padrón estadístico del INE, que contiene información demográfica sobre toda la población que reside en algún ayuntamiento gallego la una fecha

determinada. A priori, es lógico pensar que la variable de fusión por antonomasia en el caso de los registros administrativos es el DNI, ya que se trata de un documento que permite identificar "unívocamente" a cada persona. Pero, no siempre se puede recurrir a esta variable en el procedimiento de unión por dos motivos:

- En algunas bases de datos y registros administrativos no se dispone de esta variable, bien porque el usuario no está obligado a suministrarla cuando se inscribe en el registro, o bien porque el organismo gestor no la ofrece acogiéndose a la Ley de Protección de datos de Carácter Personal
- Se constató que el DNI no permite identificar unívocamente a cada persona, ya que, aunque en raras ocasiones, hay personas que han asignado el mismo número y también hay personas que no disponen de este identificador

Puesto que no siempre se puede recurrir al DNI en el enlace de bases de datos, cruzaremos los distintos registros administrativos empleando un conjunto de variables, como el nombre y los apellidos de la persona, su fecha de nacimiento, el ayuntamiento de residencia, etc; esto es, variables comunes a dos registros. En algunas ocasiones deberemos someter las variables de cruce a un procedimiento previo de depuración.

Objetivo

El objetivo de esta actividad estadística es representar la distribución de la población de Galicia, según diversas características sociodemográficas, en un mapa con un amplio desglose territorial. Se emplea una malla regular formada por celdas de 1km de lado, en la que se difundirá una **aproximación** de la población gallega a partir de la fusión y explotación de varios registros administrativos.

Conocer el emplazamiento concreto de residencia de la persona permitirá analizar relaciones entre localizaciones geográficas concretas, áreas o zonas de influencia, que vayan más allá de las delimitaciones que se suenen emplear en la estadística pública (ayuntamientos o, bajando un nivel, entidades de población). Esta información también puede ayudar a los poder públicos a trazar políticas de actuación focalizadas en territorios concretos.

Unidades estadísticas

La población objeto de estudio es la población residente en Galicia a 01/01 del año en curso.

Variables

La variable objeto de estudio es la población residente en Galicia, tanto total como clasificada según grupos de edad, sexo, lugar de nacimiento en relación al lugar de residencia, afiliación

a la Seguridad Social, pensiones contributivas de la Seguridad Social y su percepción. Las categorías contempladas son las siguientes:

Sexo

- Hombre
- Mujer

Grupos de edad

- Menos de 16 años
- De 16 a 64 años
- 65 e máis años

Lugar de nacimiento en relación al lugar de residencia

- Mismo municipio
- Distinto municipio dentro de Galicia
- Resto de España
- Estranjero

Afiliación a la Seguridad Social a 31/12

- Afiliaciones a la Seguridad Social
- Personas afiliadas a la Seguridad Social
- Porcentaje de hombres afiliados a la Seguridad Social
- Porcentaje de mujeres afiliadas a la Seguridad Social
- Ratio de feminidad de la población afiliada
- Porcentaje de población entre 16 e 34 años afiliada a la Seguridad Social
- Porcentaje de población entre 35 e 54 años afiliada a la Seguridad Social
- Porcentaje de población de 55 o más anos afiliada á Seguridad Social
- Porcentaje de población de nacionalidad extranjera afiliada a la Seguridad Social
- Porcentaje de afiliaciones no Régimen General y minería del carbón
- Porcentaje de afiliaciones no Régimen Especial de Autónomos
- Porcentaje de afiliaciones a la Seguridad Social en la agricultura y la pesca
- Porcentaje de afiliaciones a la Seguridad Social en la industria
- Porcentaje de afiliaciones a la Seguridad Social en la construcción
- Porcentaje de afiliaciones a la Seguridad Social en los servicios
-

Pensiones contributivas de la Seguridad Social

- Perceptores de pensiones contributivas
- Perceptores de pensiones contributivas hombre
- Perceptores de pensiones contributivas mujeres
- Ingresos por pensiones contributivas

Las fuentes empleadas para la elaboración de esta actividad son las siguientes:

- Padrón estadístico: fichero final de la Estadística del Padrón Continuo con microdatos de población ajustada a la cifras oficiales a 1 de enero de cada año. Organismo responsable: INE.
- Padrón continuo: microdatos procedentes de las descargas nominales a 1 de enero y 1 de julio de cada año del Padrón Municipal Continuo (coordinación de los padrones municipales) para el ámbito de Galicia. Organismo responsable: INE.
- Afiliaciones a la Seguridad Social en alta laboral: trabajadores en alta laboral en la Seguridad Social a 31/12 del año en curso. Organismo responsable: Tesorería General de la Seguridad Social.
- Pensiones de la Seguridad Social: pensiones gestionadas por el INSS y por el ISM que figuran en el Registro de Prestaciones Sociales Públicas, del cuál es titular el INSS. Organismo responsable: INSS y ISM
- Cartociudad: proyecto colaborativo liderado por el Instituto Geográfico Nacional (IGN) de producción y publicación mediante servicios web de datos espaciales de cobertura nacional. Contiene información de la red viaria continua con las calles, con portales, y las carreteras, con puntos kilométricos. Dispone de un servicio de procesamiento web basado en cálculos programados que operan sobre la información georreferenciada. Organismo responsable: IGN
- Planimetría de las secciones censales de Galicia, con fecha 1 de enero del año de referencia de los datos. Organismo responsable: INE.
- Modelo de Direcciones de la Administración General del Estado (MDAGE). Organismo responsable: INE.
- Catastro: información alfanumérica y cartografía catastral de todos los ayuntamientos de Galicia. Organismo responsable: Dirección General del Catastro.
- Callejero del censo electoral. Organismo responsable: INE
- Mapa de las parroquias de Galicia. Organismo responsable: IET
- Mapa de las entidades singulares de Galicia: elaborado a partir de los datos del Censo de Viviendas del año 2011 y de otros datos disponibles en el IET. Organismo responsable: IET
- Cartografía de los Lugares del Nomenclátor Galicia. Organismo responsable: IET

Definiciones y conceptos

Directiva INSPIRE. La Directiva INSPIRE (Infrastructure for Spatial Information in Europe) determina las reglas generales para el establecimiento de una Infraestructura de Información Espacial en la Unión Europea. Se inicia ante la necesidad de organizar y poner en común la información espacial de las diferentes Infraestructuras de datos Espaciales de los Estados Miembros y con el objetivo de superar los problemas de disponibilidad, calidad, gestión, accesibilidad y puesta en común de toda la geo-información..

Sistema de referencia espacial. Un sistema de referencia espacial permite asignar coordenadas a puntos sobre la superficie terrestre. Son utilizados en geodesia, navegación, cartografía y sistemas globales de navegación por satélite para la correcta georreferenciación de elementos en la superficie terrestre. Estos sistemas son necesarios dado que la Tierra no es una esfera perfecta.

Sistema de referencia ETRS89- LAEA. Sistema de referencia espacial que la Directiva INSPIRE recomienda para la generación de una capa vectorial de celdas uniformes de 1 km de lado que sea homogénea para toda Europa. Usa el sistema de coordenadas Lambert Azimutal Equal Area (LAEA).

Residente. Se considera residente a toda persona física que tiene su residencia habitual en uno de los ayuntamientos de la Comunidad Autónoma de Galicia

Personas afiliadas a la Seguridad Social. La afiliación al Sistema de la Seguridad Social es obligatoria para todas las personas incluidas en el campo de aplicación de la Seguridad Social y única para toda la vida del trabajador y de la trabajadora y para todo el sistema, sin perjuicio de las bajas, altas y demás variaciones que con posterioridad a la afiliación puedan producirse. Es decir, el trabajador o la trabajadora se afilian cuándo comienza su vida laboral y se da de alta en alguno de los regímenes del Sistema de la Seguridad Social. Esta situación se denomina alta inicial. Si cesa en su actividad será dado de baja pero seguirá afiliado/a en situación de baja laboral. Si retoma la actividad se producirá una alta denominada alta sucesiva a efectos estadísticos, pero no tendrá que afiliarse nuevamente, dado que, como ya se indicó, la afiliación es única para toda la vida del trabajador o de la trabajadora. En esta actividad estadística se difunde información de las personas en alta laboral en la Seguridad Social el 31/12 del año en curso.

Pensiones contributivas de la Seguridad Social. Son prestaciones económicas y de duración indefinida, aunque no siempre, en las que la concesión está generalmente supeditada a una previa relación jurídica con la Seguridad Social (acreditar un período mínimo de cotización en determinados casos, ...), siempre que se cumplan los demás requisitos exigidos. Su cuantía se determina en función de las aportaciones efectuadas por el trabajador y el empresario, si se trata de trabajadores por cuenta ajena, durante el período considerado a los efectos de la base reguladora de la pensión de que se trate. Las clases de pensiones son: incapacidad permanente, jubilación, viudedad, orfandad y a favor de familiares.

Procesamiento de los datos

Para la realización de esta actividad estadística se realizaron los siguientes pasos:

1. Depuración de los códigos que identifican el distrito y la sección censales de residencia en el Padrón continuo
2. Enlace del Padrón estadístico y el Padrón continuo
3. Georreferenciación de los portales del Padrón estadístico
4. Cruce con el fichero de afiliaciones en alta laboral a la Seguridad Social

5. Cruce con el fichero de prestaciones de la Seguridad Social.

1º Paso: depuración de las secciones y de los distritos censuales en el padrón continuo

La primera tarea que hay que realizar para la fusión de los registros es la depuración de los códigos que identifican el distrito y la sección censales de residencia en el Padrón continuo de habitantes.

El Padrón continuo constituye una primera aproximación a la población de Galicia. Este registro, que nos remite el INE dos veces al año, con referencia de los datos el 1 de enero y el 1 de julio de cada año, contiene el DNI de la persona. No obstante, se trata de una primera aproximación que puede contener errores. El INE somete esta base de datos a un proceso de depuración y remiten el resultado de esta depuración, con fecha de referencia el 1 de enero. Este segundo "Padrón" es lo que se denomina como Padrón estadístico. Esta base no contiene el DNI de la persona, que es una variable muy relevante a la hora de cruzar registros administrativos; por lo que tratamos de vincular el Padrón estadístico con el continuo a 1 de enero para poder contar con el DNI de la persona.

El objetivo en este punto es depurar las variables que identifican el distrito y la sección de residencia de cada persona en el Padrón continuo, con el objeto de maximizar, a posteriori, el número de registros que ofrece lo cruce de los Padrones continuo y estadístico. Para ello, recurrimos al Callejero del censo electoral, que ofrece una relación sistemática y actualizada a 1 de enero del año en curso de las vías y tramos de vía que pertenecen a cada sección censal. En particular, se detectó que el Padrón continuo contiene erratas en estas dos variables:

- En las zonas urbanas, donde las vías suelen disponer de un código identificativo (*cvia*) y las viviendas suelen estar numeradas (*numer*), la mayor parte de las erratas advertidas tienen que ver con...
    - Registros asignados a una vía (*cvia*) y ayuntamiento concretos, cuando en el Callejero no existe tal código de vía en ese ayuntamiento
    - Registros asignados a una vía (*cvia*) con una numeración de vivienda (*numer*) que no se encuentra entre los extremos inferior y superior de numeración de esa vía en el Callejero (*ein* y *esn*)
- En las zonas rurales, donde las vías no suelen disponer de código identificativo (*cvia*=0) y las viviendas, muchas veces, no se encuentran numeradas (*numer* en blanco), la mayor parte de las erratas se deben a que el código que identifica la entidad colectiva (*cun*) figura asociado con códigos distintos para el distrito y la sección (*dist+ secc*) en ambas bases de datos (Padrón continuo y Callejero).

Las variables fundamentales con las que se trabajará para intentar depurar el distrito y la sección del Padrón continuo a partir del Callejero del censo electoral son las siguientes:

- *cpro*: código de la provincia de residencia (se llama igual en el Padrón y más en el Callejero)
- *cmun*: código del ayuntamiento de residencia (se llama igual en el Padrón y más en el Callejero)
- *cun*: código que identifica la entidad de residencia (se llama igual en el Padrón y más en el Callejero); se trata de una variable creada por el INE concatenando...
  - El código de la entidad colectiva (2 dígitos)
  - El código de la entidad singular (2 dígitos)
  - Un dígito de control
  - El código del núcleo (99 se es diseminado)
- *cvia*: código que identifica la vía (se llama igual en el Padrón y en el Callejero)
- *tinum*: variable que toma valor 0 si la vía no se encuentra numerada, 1 para el tramo de numeración impar y 2 para el tramo de numeración par (se llama igual en el Padrón y más en el Callejero)
- *numer*: número de la vivienda de residencia (solo se encuentra en el Padrón continuo)
- *ein*: extremo inferior de numeración de la vía (solo se encuentra en el Callejero)
- *esn*: extremo superior de numeración de la vía (solo se encuentra en el Callejero)
- *cein* (Callejero)/*cnumer* (Padrón): cualificador del extremo inferior de numeración
- cesn (Callejero)/*cnumers* (Padrón): cualificador del extremo superior de numeración
- *nviac*: nombre corto de la vía (se encuentra en ambas bases de datos, pero solo vamos a emplear los valores del Padrón continuo)
- *nentsic*: nombre corto de la entidad singular de población (se encuentra en ambas bases de datos, pero solo vamos a emplear los valores del Callejero)

2º Paso: enlace del Padrón estadístico y el Padrón continuo

El objetivo en este punto es tomar como base de la fusión el Padrón estadístico y cruzar con el Padrón continuo (con el distrito y sección depurados), con la finalidad de crear una tabla auxiliar que sirva de nexo entre ambos registros. De esta forma, se podrá recuperar de forma rápida y sencilla el documento identificativo de la persona (ya se trate del DNI, del pasaporte/DNI UE o del NIE) del Padrón continuo, cuando sea útil emplear esta variable en el cruce entre el Padrón estadístico y otros registros administrativos.

Las variables fundamentales con las que trabajamos para cruzar son las siguientes:

- *cpro*: código de la provincia de residencia
- *cmun*: código del ayuntamiento de residencia
- *cvia*: código que identifica la vía de residencia
- *dist* (Padrón estadístico) y *dist_ dep* (Padrón continuo): código del distrito censal de residencia

- *secc* (Padrón estadístico) y *secc_ dep* (Padrón continuo): código de la sección censal de residencia
- *sexo*: código identificativo del sexo de la persona (1 si es hombre, 6 si es mujer)
- *anno+mes+día* (Padrón estadístico) y *fnac* (Padrón continuo): fecha de nacimiento
- *nomb*: nombre de la persona
- *ape1*: primer apellido de la persona
- *ape2*: según apellido de la persona
- *cpron*: código de la provincia de nacimiento
- *cmunn*: código del ayuntamiento de nacimiento

En algunas, como es el caso de la variable sexo, la unión es automática, puesto que se trata de una variable categórica que solo puede tomar dos valores, 1 o 6. En este caso, el cruce se realiza directamente con el programa SQL Server, en el entorno de la propia base de datos en la que se graba la información. No obstante, para cruzar por medio de variables de tipo carácter, como el nombre, que pueden diferir para una misma persona de una tabla a otra (por ejemplo, en un registro puede figurar el nombre simple de la persona, "MARÍA", y en el otro el compuesto, "MARÍA DO CARME"), es preciso recurrir a cruces indirectos, que no busquen similitud exacta, sino grado de semejanza. Realizar este tipo de cruces con las herramientas de consulta de SQL es complicado, por lo que se emplean las librerías stringdist y fuzzyjoin (Robinson, 2016) del software libre de programación R. Ambas permiten comparar el grado de similitud que existe entre cadenas de texto o entre variables numéricas de una manera rápida y sencilla, vinculando cada registro de una base con el registro de la otra base de datos con el que guarde mayor grado de semejanza. Por lo tanto, una parte de los cruces se realizará en SQL y a otra en R.

El procedimiento de cruce consta de varios pasos; en el primero, se procede a la unión de registros por medio de todas aquellas variables comunes en las dos tablas, las mencionadas en el párrafo precedente. Con este primer paso se logra vincular a más del 80% de los registros de ambos Padrones, continuo y estadístico. Para los restantes, se procede a realizar una nueva unión, prescindiendo en este segundo paso de una de las variables de cruce: el sexo de la persona. El procedimiento continúa, relajando criterios de semejanza en cada nuevo paso (el lugar de residencia, el lugar de nacimiento, etc.), hasta que se consigue completar la unión de más del 98% de los registros del Padrón estadístico. En todo paso, se exige que el cruce sea biunívoco, esto es: si dos o más registros del Padrón continuo cumplen los requisitos de semejanza con uno de los registros del estadístico, se prescinde del cruce. Además, a cada paso del procedimiento de unión se le asigna una calidad de precisión, que va de 1 (máxima precisión) a 12 (mínima precisión aceptable). Después de completado, se comprueba que las uniones son correctas, en particular los cruces con precisión baja (valores elevados en la variable que muestra la calidad), para corroborar que se trata de la misma persona en el Padrón estadístico y en el continuo. Para realizar estas comprobaciones, se

recurre también a la búsqueda de la información de la persona en los registros de años precedentes, en particular cuando se trata de uniones donde la coincidencia en variables clave como el nombre y los apellidos no es exacta o cuando cambia la fecha de nacimiento de un Padrón a otro. Este tipo de contraste del historial de la persona en los registros administrativos resultó ser muy útil para conseguir un alto grado de fiabilidad de la unión realizada entre las bases de datos de población.

3º Paso: georreferenciación de los portales

En este punto se exponen la metodología empleada para georreferenciar los portales donde reside la población de Galicia. A cada uno de los portales se le asignará una coordenada X- Y en un sistema de Referencia estándar.

En este procedimiento de georreferenciación se echó mano también del software libre R y de los diversos paquetes cartográficos que ofrece. El objetivo en este punto es georreferenciar la población disponible en los portales del Padrón estadístico.

La fuente principal que se emplea para georreferenciar es la base de datos proporcionada por el INE MDAGE. En esta base de datos están disponibles las aproximaciones postales (portales) de Galicia con sus coordenadas UTM X, Y en el HUSO 30.

Las variables que se emplearán para georreferenciar son:

- cvia: código de vía
- tvia: tipo de vía
- nvia: nombre de la vía
- num: número del portal
- calificador: calificador del portal
- cec: código de entidad colectiva
- ces: código de entidad singular
- cnuc: código de núcleo o diseminado
- dices: código de distrito censal
- secc: código de sección censal

Estas variables están disponibles en las dos bases de datos.

Hay que resaltar que la base de datos MDAGE no es completa, existen portales que no tienen unas coordenadas X, Y válidas o que no están georreferenciados. De todas formas, esta será la fuente de información principal, aunque después se complete con otras fuentes auxiliares.

Por lo tanto, la estrategia que se empleará para georreferenciar la población será realizar un procedimiento ayuntamiento a ayuntamiento. Primero se georreferenciará las vías que tienen *cvia*>0 y que se corresponde (aproximadamente) con la parte más urbana de Galicia y a continuación se georreferenciarán las vías con *cvia*=0, que se corresponde (aproximadamente)

con la parte más rural de Galicia. Dentro de cada ayuntamiento los pasos a seguir se describen a continuación:

Parte urbana:

Paso 1.- Se georreferencian las calles que tienen disponible el código de vía, tipo de vía, nombre de vía y número. En este caso se cruza la base de datos MDAGE con el Padrón estadístico empleando las variables anteriores y se le asigna las coordenadas correspondientes:

- En este proceso surge el problema de que en la MDAGE existen casos donde dos portales tienen los mismos códigos de vía y número, pero distintas coordenadas. Para resolver este problema se calculó el promedio de las coordenadas y se le asignó a ambos portales este valor medio.
- Para las calles que tienen el mismo código de vía pero el número del Padrón estadístico no está disponible en el MDAGE lo que se hizo fue aplicar la regresión con splines. Se tendrá en cuenta si los números del portal son pares o impares. Aplicando el siguiente modelo con splines se obtendría la predicción para las coordenadas X, Y del número nuevo:

$$Y_i = f(X_i) + aI_i + e_i \quad (1)$$

donde

- *Y* coordenada geográfica UTM Y en el huso 29,
- *X* coordenada geográfica UTM X en el huso 29,
- *I* indicadora de la paridad del número (1 si es par; 0, impar),
- *e* variable aleatoria N(0, σ^2),
- *f* función suave,
- *i*=1, ..., nº total de portales disponibles de la vía.

Para su ajuste se emplearán los splines penalizados.

Una vez establecido este primero modelo se necesita un segundo modelo que permita calcular la coordenada geográfica X de un nuevo portal que se va a imputar. El planteamiento para cada una de las vías en este segundo modelo es:

$$X_i = g(N_i) + bI_i + e_i' \quad (2)$$

donde

- *X* coordenada geográfica UTM X en el huso 29,
- *N* nº del portal del inmueble,
- *I* indicadora de la paridad del número (1 si es par; 0, impar),
- *g* función suave,
- *e'* variable aleatoria N(0, σ'^2),
- *i*=1, ..., nº total de portales disponibles de la vía.

Para lo ajuste de este segundo modelo también se emplearán los splines penalizados.

En caso de que no haya suficientes datos para aplicar la regresión se optó por asignarle la coordenada del número más próximo.

Paso 2. Para los portales que quedan sin georreferenciar en el paso 1 se empleará Cartociudad que proporcionándole el nombre de la vía, el número y el ayuntamiento devuelve las coordenadas X, Y.

Paso 3. Para los portales que quedan sin georreferenciar en el paso 2 se empleará la información que proporciona la Dirección General de Catastro. Esta cartografía ofrece las coordenadas de las parcelas donde se ubican los inmuebles. El problema que ofrece estas dos bases de datos es que no existe un nexo de unión entre los ficheros del Padrón estadístico y los ficheros del Catastro. Como en los dos casos disponemos de direcciones de los inmuebles, lo que se hará será unir por direcciones tratando de vincular las vías que nos proporciona a Padrón estadístico con las vías que proporciona la Dirección General de Catastro en sus archivos. Para esto se empleará un procedimiento basado en unir tablas mediante variables que no tienen una coincidencia exacta. Se empleará el paquete de R *text2vec* (Selivanov et al., 2020). También se empleará el mapa de secciones censales, de tal manera que la vía del Catastro tiene que estar en la misma sección censal que la vía disponible en el Padrón estadístico.

Una vez hecha la correspondencia entre el código de vía del Padrón estadístico y el código de vía del catastro, es posible que en el catastro no estén todos los números de los portales. En este caso pueden ocurrir dos situaciones diferentes:

1- Algunos números que proporciona catastro son ceros

Si los números de portal que proporciona catastro tienen ceros no se puede saber si se trata de números pares o impares. Para averiguar esto emplearemos el análisis de componentes principales (ACP). Aplicaremos el ACP a las coordenadas X, Y que nos proporciona el Catastro. Se puede comprobar que la primera componente principal se corresponde con la dirección de la vía y la segunda, ortogonal a la primera, con los pares e impares. Una vez que se sabe cuáles son los puntos pares e impares, ya se puede calcular el centroide de las coordenadas de los pares y de los impares y asignárselo al número que se quiere georreferenciar, dependiendo de que sea par o impar.

2- Ningún número es nulo

En esta situación para determinar las coordenadas del número nuevo aplicaremos la regresión con splines, método empleado en el paso 1.

Paso 4. Para aquellas calles que no se georreferencian con el paso anterior se empleará la API de Google y los servicios de la API de Here. En concreto, en el caso de Google, se empleará la función de R *geocode* del paquete *ggmap* (Kahle and Wickham, 2013) que proporcionándole el nombre de la vía, el número del portal y el ayuntamiento devuelve las coordenadas X, Y. En el caso de Here, desde R se llamará a la API con el nombre de vía, el número de portal y el ayuntamiento, y se obtendrá de vuelta las coordenadas X, Y.

Parte rural:

Paso 1.- Se georreferencian los portales que están en las entidades que no tienen disponible a clase de vía, código o nombre de vía. En este caso la georreferenciación se realiza a nivel de núcleo/diseminado y número (del portal) de la siguiente manera:

- Para aquellos portales que tienen un número en el Padrón estadístico y este número está disponible en el MDAGE, se le asigna la coordenada disponible en el MDAGE.
- Para aquellos portales que tienen un número en el Padrón Estadístico y este número no está disponible en el MDAGE se le asigna el portal más próximo de Catastro dentro de la entidad singular donde está ubicado. Se emplea para esto el mapa de entidades singulares del IET.

Paso 2. Para los portales que quedan sin georreferenciar en el paso 1 se sigue el siguiente procedimiento:

- Para los portales no georreferenciados en el punto anterior se buscó el más próximo de Catastro dentro de la parroquia a la que pertenecen. Se emplea para esto el mapa parroquias del IET y la función de R *knn* del paquete *class* (Venables and Ripley, 2002) que busca los vecinos más próximos.
- Para los restantes portales se buscó el lugar más próximo dentro de la cartografía de Lugares del IET de los LugaresNomenclatorGalicia, y se le asignó el centroide del Lugar.

Paso 4º: cruce con el fichero de afiliaciones en alta laboral a la Seguridad Social

En el 4º paso del procedimiento de construcción de la tabla de datos que servirá de base para la extracción de las características socioeconómicas de la población de Galicia en celdas de un 1km2 se tratará de anexar la información de los ficheros de afiliaciones en alta laboral de la Seguridad Social. Para ello, se aprovechan las tareas internas realizadas en el marco de la operación estadística *Afiliaciones a la Seguridad Social por ayuntamiento de residencia de la persona afiliada*, que publica cada 3 meses el IGE con información desde el año 2006. El objetivo de esta operación es ofrecer información de la evolución de la afiliación al Sistema de la Seguridad Social en Galicia, con un nivel de desglose territorial que llega al ayuntamiento de residencia de la persona afiliada. En los ficheros que suministra la Tesorería de la Seguridad Social figura esta variable, el ayuntamiento de residencia del afiliado o de la afiliada pero, en una parte importante de los registros, está desactualizada, reflejando el ayuntamiento en el que residía la persona cuando se dio de alta en el Sistema y no la residencia actual. Por este motivo, se procede a cruzar estos ficheros de afiliación con el Padrón continuo de habitantes, a 01/01 del año de referencia en las tablas de la Seguridad Social suministradas a 31/03 y 30/06 del año en cuestión, y con el Padrón continuo a 01/07 para los ficheros de afiliación relativos a 31/09 y 31/12. Para analizar el procedimiento de cruce entre el Padrón continuo y los registros

trimestrales de afiliaciones a la Seguridad Social, se puede consultar el proyecto técnico de la operación, en el siguiente enlace:
https://www.ige.eu/estatico/pdfs/s3/proxectosTecnicos/24-106_AfiliacionsASeguridadeSocialporConcelloResidenciaAfiliado.pdf

Hecha esta unión en el marco de la operación reseñada y, puesto que en el 2º paso se cruzaron los Padrones estadístico y continuo, la unión entre el Padrón estadístico y los registros de afiliación a la Seguridad Social es automática. No obstante, queda una parte muy pequeña del Padrón estadístico, aquella que no se pudo vincular con el continuo, para la que no se tiene correspondencia con el registro de Afiliaciones a la Seguridad Social. Para estos registros se aplica un procedimiento de unión específico entre el Padrón estadístico y los ficheros de afiliación, muy similar al empleado en el caso de la unión entre Padrones: solo se permiten cruces biunívocos, se van relajando criterios de similitud/semejanza entre variables en cada paso del procedimiento de unión (en el primer paso se emplean todas las variables comunes a ambos ficheros, en el segundo se elimina una de ellas, en el tercero dos, etc.) y se crea una variable calidad que indica el grado de precisión de la unión. Finalmente, se contrasta que se trata de la misma persona en ambos ficheros, en el Padrón estadístico y en los registros de afiliaciones a la Seguridad Social.

Paso 5º: cruce con el fichero de prestaciones de la Seguridad Social

En este punto el objetivo es cruzar el Padrón estadístico con el fichero de Pensiones contributivas gestionadas por el INSS y por el ISM y proporcionadas al IGE con fecha de referencia el 31 de diciembre de cada año.
El procedimiento empleado para hacer el cruce es el siguiente:
Paso 1.- Si coincide el identificador (DNI, pasaporte y DNI europeo o documento de extranjero) de los dos ficheros se considera que la persona es la misma.
Paso 2.- Si no se puede efectuar lo cruce en el paso 1, se procede a hacer el enlace por nombre, apellidos, sexo y fecha de nacimiento. Si coinciden estas variables en los dos ficheros se considera que la persona es la misma.
Paso 3.- Si no se puede efectuar el cruce en los pasos 1 y 2 , se procede a hacer el enlace por nombre, apellidos y sexo. Si coinciden estas variables en los dos ficheros se considera que la persona es la misma.
Paso 4.- Si no se puede efectuar lo cruce en los pasos 1, 2 y 3 , se procede a hacer el enlace por nombre, primer apellido y fecha de nacimiento. Si coinciden estas variables en los dos ficheros se considera que la persona es la misma.
Paso 5.- Si no se puede efectuar los cruce en los pasos 1, 2, 3 y 4 , se procede a hacer el enlace por nombre, segundo apellido y fecha de nacimiento. Si coinciden estas variables en los dos ficheros se considera que la persona es la misma.

Paso 6.- Si no se puede efectuar el cruce en los pasos 1, 2, 3, 4 y 5 , se procede a hacer el enlace por apellidos, sexo y fecha de nacimiento. Si coinciden estas dos variables en los dos ficheros se considera que la persona es la misma.
Paso 7.- Si no se puede efectuar el cruce en los pasos 1, 2, 3, 4 ,5 y 6, se procede a hacer el enlace por nombre, apellidos y por fechas de nacimiento donde la distancia sea menor de 3. Si se cumplen las condiciones se considera que la persona es la misma.

En este punto hay que destacar que en el 1º paso cruzan aproximadamente el 98% de los registros de pensiones y hay en torno a 10.000 pensiones que no se consigue cruzar con el Padrón estadístico y con el procedimiento establecido.

Diseño de la explotación de los datos de la fusión de los registros

A la información fusionada y georreferenciada se le asignó las celdas de la malla regular de 1 km2 generada. El sentido de esta operación se encuentra en la necesidad, por una parte, de proteger la confidencialidad de la información, y por la otra, de facilitar la representación cartográfica de la información.
En la operación de geocodificación se obtuvieron resultados que no llegan al 100% de la totalidad de la población, en concreto para el año 2018 se geocodificó el 99,9% de la población de Galicia.
El resultado final es una capa vectorial compuesta por una malla regular de celdas cuadradas de 1km de lado. Para cada una de estas celdas se presenta la información estadística de carácter sociodemográfica que le corresponde, desglosada de la siguiente forma (entre paréntesis el código de la variable):

- Población de hombres (hombres)
- Población de mujeres (mujeres)
- Índice de feminidad: Mujeres/ Hombres 100 (ratio_ femi)
- Porcentaje de población menor de 16 años (porc_menor_16)
- Porcentaje de población entre 16 y 64 años (porc_mayor15_me)
- Porcentaje de población con 65 años o más (porc_mayor_64)
- Edad media de la población: promedio de las edades de la población (idadeMedia)
- Índice de envejecimiento: población >65 años/población <20 años 100 (ind_ envelle)
- Población con nacionalidad extranjera (porc_ estranx)
- Población residente en el mismo ayuntamiento de nacimiento (porc_mismo_ conc)
- Población residente nacida en Galicia en un ayuntamiento diferente al de nacimiento (porc_ outroconce)
- Población nacida en otra Comunidad Autónoma (porc_ OutraCCAA)
- Población nacida en otro país (porcNacExtr)
- Afiliaciones a la Seguridad Social (afiliacions)

- Personas afiliadas a la Seguridad Social (afiliados)
- Porcentaje de hombres afiliados a la Seguridad Social (porc_ hom_ afi)
- Porcentaje de mujeres afiliadas a la Seguridad Social (porc_ mull_ afi)
- Ratio de feminidad de la población afiliada (ratio_ fem_ afi): Mujeres afiliadas/Hombres afiliados 100
- Porcentaje de población entre 16 y 34 años afiliada a la Seguridad Social (porc_16_34_ afi)
- Porcentaje de población entre 35 y 54 años afiliada a la Seguridad Social (porc_35_54_ afi)
- Porcentaje de población de 55 o más años afiliada a la Seguridad Social (porc_55_me las_ afi)
- Porcentaje de población de nacionalidad extranjera afiliada a la Seguridad Social (porc_ estran_ afi)
- Porcentaje de afiliaciones en el Régimen General y minería del carbón (porc_general_ afi)
- Porcentaje de afiliaciones en el Régimen Especial de Autónomos (porc_auto_ afi)
- Porcentaje de afiliaciones a la Seguridad Social en la agricultura y la pesca (porc_prima_ afi)
- Porcentaje de afiliaciones a la Seguridad Social en la industria (porc_ indus_ afi/ porc_ cons_ afi)
- Porcentaje de afiliaciones a la Seguridad Social en la construcción (porc_ cons_ afi)
- Porcentaje de afiliaciones a la Seguridad Social en los servicios (porc_ serv_ afi)
- Población perceptora de pensiones contributivas de la Seguridad Social (Pensionistas)
- Hombres perceptores de pensiones contributivas de la Seguridad Social (Pensionistas_ ho)
- Mujeres perceptoras de pensiones contributivas de la Seguridad Social (Pensionistas_ mu)
- Índice de feminidad de las personas perceptoras de pensiones contributivas de la Seguridad Social (ratio_ fem_ pens): Mujeres perceptoras de pensiones/Hombres perceptores de pensiones 100
- Pensión media de los pensionistas de las Seguridad Social: importe bruto mensual de la prestación en € (pensionMedia)

Secreto estadístico

Se garantizó la protección y la confidencialidad de la información mediante el método de eliminación de datos. Inicialmente se realizó una eliminación de los datos considerados sensibles y, posteriormente, cuando fue necesario, una supresión secundaria que impida la identificación exacta por deducción de los datos sensibles censurados, por diferencia con respeto al total. En concreto, los criterios empleados fueron los siguientes:

- Se eliminaron las celdas que tienen una población total menor de 21 habitantes.
- Se censuraron las celdas que tienen menos de 5 habitantes (>0) en alguna de las características sociodemográficas por las que se clasifica.
- Se censuraron celdas complementarias para evitar la deducción de datos sensibles censurados en el paso anterior por diferencia respeto al total. En este caso se eliminaron los datos teniendo en cuenta su valor, con la intención de minimizar el coste en términos de pérdida de información

Por lo tanto, los criterios aplicados impiden la identificación exacta de los datos considerados sensibles, no su aproximación. El proceso de eliminación de datos se realizó aplicando un algoritmo desarrollado ad-hoc que cumple los criterios de secreto estadístico establecidos.

El territorio de Galicia se cubre con un total de 30.776 celdas de 1 km2. En el año 2018, 19.021 celdas tenían población. De estas, se difunden 10.775 celdas que tienen una población superior a 20 personas. Estas celdas concentran el 97,2% de la población de Galicia, por lo tanto, las 8.246 celdas que no se puede difundir por problemas de secreto estadístico, concentrarían el 2,8% de la población de Galicia. Este sería el coste en términos de pérdida de información que hay que asumir por el cumplimiento del secreto estadístico.

Control de calidad

Se realizó una revisión de la información espacial de los portales utilizados para la generación de la malla de población. En concreto se realizaron los siguientes contrastes:

- Los portales georreferenciados deben estar geográficamente localizados en el ayuntamiento del Padrón estadístico.
- Los portales georreferenciados deben estar geográficamente localizados en la sección del Padrón estadístico o en la misma entidad singular o, en su defecto, en la entidad singular vecina o en la misma parroquia.

Difusión de la información

Se desarrolló un visualizador cartográfico de fácil manejo y muy intuitivo. En el mismo se puede consultar la información en forma de mapas. A cada celda se le asigna un color según el intervalo en el que se encuentre la variable sociodemográfica que se va a representar. Los datos de celdas con población menor de 21 habitantes no se representan y los restantes datos censurados aparecerán representados con una celda blanca.

El visualizador permite realizar desplazamientos por todo el territorio, así como todos aquellos zooms que muestren al usuario a información en el nivel de detalle que desee. Estos niveles van desde lo más general que permite la visión global de toda Galicia, hasta escalas urbanas de gran detalle. El visualizador permite también cambiar de un mapa a otro conservando la vista del mapa anterior.

Existe también la posibilidad de consultar los datos de cada celda, mediante el recurso de picar para que aparezca un cuadro emergente, en el que se muestra la información del ayuntamiento donde se localiza la celda, el número de personas residentes en la misma y la característica elegida para representar en el mapa: hombres, mujeres, número de personas perceptoras de pensiones contributivas, ...

También se puede consultar la información en formato tabla, descargable en formato hoja de cálculo.

Por último, se incluyó la consulta de la información de las celdas mediante una API interoperable que permite obtener información para las celdas de:

- Galicia: http://www.ige.eu/igebdt/igeapi/csv/grid1km/<añoo>/0/0/<código de las variables separados por />
- Provincias: http://www.ige.eu/igebdt/igeapi/csv/grid1km/<ano>/<código de la provincia>/0/<código de las variables separadas por />
- Un ayuntamiento: http://www.ige.eu/igebdt/igeapi/csv/grid1km/<ano>/<código de la provincia>/<código del municipio>/<código de las variables separadas por />

La puesta a disposición de la información en formatos interoperables contribuye 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, así como, de la ciudadanía en general.

Bibliografía

GOERLICH, F. J. and MAS, M (2009): "Drivers of agglomeration: Geography versus History". The Open Urban Studies Journal 2. Recuperado de: htttp://www.bentham.org/open/tousj/openaccess2.htm

KAHLE, D. AND WICKHAM, H.. (2013) ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

R CORE TEAM (2016). "R: Language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. Recuperado de: https://www.r-project.org/.

REHER, D. (1994): "Ciudades, procesos de urbanización y sistemas urbanos en la Península Ibérica 1550-1991". En M. Guardia, F.J. Monclús, J. Oyón (dirs.). *Atlas histórico de ciudades europeas*. Barcelona: Centre de Cultura Contemporánea de Barcelona y Salvat, 1.29.

ROBINSON, D. (2016): "fuzzyjoin: Join Tables Together on Inexact Matching. R package version 0.1.2". Recuperado de: https://CRAN.R-project.org/package=fuzzyjoin

SELIVANOV,D. , BICKEL, M., y WANG,Q. (2020). text2vec: Modern Text Mining Framework for R. R package version 0.6. https://CRAN.R-project.org/package=text2vec

RÚA, A., REDONDO, R. y DEL CAMPO C. (2003): “Distribución municipal de la realidad socioeconómica gallega”. *Revista Galega de Economía*, vol 12, núm 2, pp 243-262

VENABLES, W. N. AND RIPLEY, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0