Distribución espacial das características da poboación de Galicia por cuadrícula de 1km^2

METODOLOXÍA

Normativa e organización

En virtude do Convenio sobre cooperación estatística e intercambio de información entre o Instituto Nacional de Estadística (INE) e o Instituto Galego de Estatística (IGE), firmado en maio de 2020, o INE remite periodicamente ao IGE:

- Estadística del Padrón Continuo: ficheiro final da Estadística do Padrón Continuo con microdatos de poboación axustada a cifras oficiais a 1 de xaneiro de cada ano, incluíndo nome, apelidos, enderezo, código postal, código de vía e número de folla padronal para o ámbito da Comunidade Autónoma (petición de carácter continuo); neste documento referirémonos a el como Padrón estatístico
- Padrón Municipal Continuo: coordinación de padróns municipais (actividade non incluída no Plan Estadístico Nacional): microdatos procedentes das descargas nominais a 1 de xaneiro e 1 de xullo de cada ano para o ámbito da Comunidade Autónoma (petición de carácter continuo); no que segue referirémonos a el como Padrón continuo

O Real Decreto 1314/1984, do 20 de xuño, polo que se regula a estrutura e competencias da Tesourería Xeral da Seguridade Social, atribúe a este organismo a competencia relativa a inscrición de empresas e a afiliación, altas e baixas das persoas traballadoras.
En abril de 2011, o Instituto Galego de Estatística (IGE) firmou un convenio de colaboración en materia estatística coa Tesourería Xeral da Seguridade Social e co Instituto Social da Mariña (ISM) polo cal recibirá trimestralmente a seguinte información extraída do ficheiro xeral de afiliación:

- ficheiro de afiliacións en alta laboral
- ficheiro de contas de cotización en alta ou baixa producida nos tres meses

O Real Decreto 2583/1996, de 13 de decembro, de estrutura orgánica e funcións do Instituto Nacional da Seguridade Social (INSS) atribúelle a xestión e funcionamento do Rexistro de Prestacións Sociais Públicas ao INSS.
En xullo de 2011 asinouse un convenio de colaboración entre o INSS, o ISM e o IGE en materia estatística (BOE núm. 219 do 12 de setembro de 2011). En virtude deste convenio, o

IGE recibirá anualmente de forma telemática información das pensións xestionadas polo INSS e polo ISM que figuran no Rexistro de Prestacións Socias Públicas, do cal é titular o INSS.

Introdución

Na actual organización administrativa do Estado español, os municipios constitúen as unidades administrativas menores nas que se divide o territorio e que teñen asignados lindes precisos. Por esta razón, os estudos cuxo obxectivo é a localización da poboación adoitan descender ata o nivel municipal (Goerlich e Mas, 2009). De todos os xeitos, diversos autores (Reher, 1994; Rúa e outros, 2003) recomendaron que para o estudo do asentamento da poboación sobre o territorio esta división é claramente insuficiente e é necesario aumentar a resolución xeográfica de análise.

Ademais, o tratamento histórico dado aos enderezos postais conducía á asignación de divisións territoriais, como é o caso da poboación en municipios, entidades colectivas ou seccións censuais. O efecto final desta asignación é que os resultados estatísticos só poden obterse respecto destas zonas ou os seus agregados, producíndose unha perda de información non desexable. Desde o punto de vista xeográfico-estatístico, a división do territorio é relevante se os dominios da partición son homoxéneos, pero no caso dos municipios e entidades colectivas, estamos ante unha situación de falta de homoxeneidade de superficie, poboación ou concentración de unidades de produción.
Tendo en conta esta problemática, hai uns anos, algúns institutos de estatística da Unión Europea co apoio de Eurostat lanzaron o European Forum of Geography and Geostatistics (EFGS), co obxectivo de harmonizar estatísticas europeas sobre a base dunha cuadrícula de 1 km de lado e nun sistema xeodésico de referencia común. Este proxecto realiza tarefas de estimación da distribución de celas de poboación para unha ampla gama de países en todo o continente europeo.

Ademais, a crecente demanda de estatísticas cun elevado detalle territorial para unha análise espacial máis precisa concrétase normativamente no Regulamento (UE) 2017/2391 do Parlamento Europeo e do Consello do 12 de decembro de 2017 polo que se modifica o Regulamento (CE) num. 1059/2003 no que respecta ás tipoloxías territoriais (Tercet), publicado no Diario Oficial da Unión Europea do 29 de decembro de 2017. Este regulamento sinala que “debe aplicarse un sistema de mallas estatísticas para calcular e atribuír os tipos territoriais ás rexións e zonas en cuestión, xa que ditos tipos dependen da distribución e da densidade da poboación en celas de malla dun quilómetro cadrado”. Este novo sistema zonal pode permitir combinar o máximo detalle territorial na difusión coa preservación da confidencialidade estatística.

Tendo en conta todo o anterior, o obxectivo desta actividade é representar á poboación de Galicia nun mapa buscando a homoxeneidade espacial e unha ampla desagregación territorial. Concretamente realízase esta actividade para analizar as posibilidades de representación da información en unidades territoriais reducidas e cunha unidade de observación homoxénea. Empregarase unha malla regular formada por celas cadradas de 1km de lado, na que se pretende representar de xeito **aproximado** a poboación de Galicia obtida a partir da explotación de diversos rexistros administrativos dispoñibles no IGE.
Para xerar a malla regular, séguense as indicacións derivadas dos experimentos realizados polo proxecto Geostat (proxecto ESSnet Geostat) do EFGS, que desenvolve a xeración dunha malla formada por celas de 1 km a cada lado, empregando o mesmo sistema de referencia espacial para toda Europa. As celas están codificadas cun sistema estándar que segue as directrices da Directiva INSPIRE e foron xeradas polo Instituto de Estudos do Territorio (IET).

Por outra banda, os rexistros administrativos constitúen unha fonte vital de información para os institutos de estatística pública. O aproveitamento destas bases de datos reduce a necesidade de recorrer a outras fontes, coma os censos e as enquisas; este feito trae aparellado a diminución, por unha banda, da carga sobre a poboación enquisada e, por outra, os custos que a elaboración destas operacións supoñen para os institutos de estatística. Ademais, se os datos polos que se interroga aos cidadáns xa están dispoñibles en rexistros administrativos, non ten sentido tratar de volver obtelos por medio dun censo ou dunha enquisa.

Non obstante, non todo son vantaxes á hora de traballar con rexistros administrativos. Hai que ter presente que a finalidade para a que se crean non é a labor estatística, polo que os conceptos e definicións incluídos neles rara vez coincidirán cos das estatísticas oficiais. Á hora de explotalos estatisticamente haberá que ter en conta tamén a normativa que está detrás da creación e xestión dos mesmos, e os posibles cambios que afecten a dita normativa.

No IGE comezamos coa tarefa de fusionar rexistros administrativos dos que dispomos, coa finalidade de crear un sistema de información que conteña datos socioeconómicos da poboación de Galicia. Estes traballos comezaron no ano 2019 coa execución da actividade de interese estatístico “Desenvolvemento dunha base de datos sociodemográfica empregando diversas fontes administrativas e Estatísticas”, dispoñible no Decreto 165/2018, do 27 de decembro, polo que se aproba o Programa estatístico anual da Comunidade Autónoma de Galicia para o ano 2019.

O rexistro de partida na fusión é o Padrón estatístico do INE, que contén información demográfica sobre toda a poboación que reside nalgún concello galego a unha data determinada. A priori, é lóxico pensar que a variable de fusión por antonomasia no caso dos rexistros administrativos é o DNI, xa que se trata dun documento que permite identificar

“univocamente” a cada persoa. Pero, non sempre se pode recorrer a esta variable no procedemento de unión por dous motivos:

- Nalgunhas bases de datos e rexistros administrativos non se dispón desta variable, ben porque o usuario non está obrigado a subministrala cando se inscribe no rexistro, ou ben porque o organismo xestor non a ofrece acolléndose á Lei de Protección de datos de Carácter Persoal
- Constatouse que o DNI non permite identificar univocamente a cada persoa, xa que, aínda que en raras ocasións, hai persoas que teñen asignado o mesmo número e tamén hai persoas que non dispoñen deste identificador

Posto que non sempre se pode recorrer ao DNI no enlace de bases de datos, cruzaremos os distintos rexistros administrativos empregando un conxunto de variables, coma o nome e os apelidos da persoa, a súa data de nacemento, o concello de residencia, etc; isto é, variables comúns a dous rexistros. Nalgunhas ocasións deberemos someter as variables de cruce a un procedemento previo de depuración.

Obxectivo

O obxectivo desta actividade estatística é representar a distribución da poboación de Galicia, segundo diversas características sociodemográficas, nun mapa cunha ampla desagregación territorial. Emprégase unha malla regular formada por celas de 1km de lado, na que se difundirá unha **aproximación** da poboación galega a partir da fusión e explotación de varios rexistros administrativos.

Coñecer o emprazamento concreto de residencia da persoa permitirá analizar relacións entre localizacións xeográficas concretas, áreas ou zonas de influencia, que vaian máis alá das delimitacións que se soen empregar na estatística pública (concellos ou, baixando un nivel, entidades de poboación). Esta información tamén pode axudar aos poderes públicos a trazar políticas de actuación focalizadas en territorios concretos.

Unidades estatísticas

A poboación obxecto de estudo é a poboación residente en Galicia a 01/01 do ano en curso.

Variables

A variable obxecto de estudo é a poboación residente en Galicia, tanto total como clasificada segundo grupos de idade, sexo, lugar de nacemento en relación ao lugar de residencia, afiliación á Seguridade Social, pensións contributivas da Seguridade Social e a súa percepción. As categorías contempladas son as seguintes:

Sexo

- Home
- Muller

Grupos de idade

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

Lugar de nacemento en relación ao lugar de residencia

- Mesmo municipio
- Distinto municipio dentro de Galicia
- Resto de España
- Estranxeiro

Afiliación á Seguridade Social a 31/12

- Afiliacións á Seguridade Social
- Persoas afiliadas á Seguridade Social
- Porcentaxe de homes afiliados á Seguridade Social
- Porcentaxe de mulleres afiliadas á Seguridade Social
- Ratio de feminidade da poboación afiliada
- Porcentaxe de poboación entre 16 e 34 anos afiliada á Seguridade Social
- Porcentaxe de poboación entre 35 e 54 anos afiliada á Seguridade Social
- Porcentaxe de poboación de 55 ou máis anos afiliada á Seguridade Social
- Porcentaxe de poboación de nacionalidade estranxeira afiliada á Seguridade Social
- Porcentaxe de afiliacións no Réxime Xeral e minería do carbón
- Porcentaxe de afiliacións no Réxime Especial de Autónomos
- Porcentaxe de afiliacións á Seguridade Social na agricultura e a pesca
- Porcentaxe de afiliacións á Seguridade Social na industria
- Porcentaxe de afiliacións á Seguridade Social na construción
- Porcentaxe de afiliacións á Seguridade Social nos servizos

Pensións contributivas da Seguridade Social

- Perceptores de pensións contributivas
- Perceptores de pensión contributivas homes
- Perceptores de pensións contributivas mulleres
- Ingresos por pensións contributivas

As fontes empregadas para a elaboración desta actividade son as seguintes:

- Padrón estatístico: ficheiro final da Estatística do Padrón Continuo con microdatos de poboación axustada a cifras oficiais ao 1 de xaneiro de cada ano. Organismo responsable: INE.

- Padrón continuo: microdatos procedentes das descargas nominais a 1 de xaneiro e 1 de xullo de cada ano do Padrón Municipal Continuo (coordinación dos padróns municipais) para o ámbito de Galicia. Organismo responsable: INE.
- Afiliacións á Seguridade Social en alta laboral: traballadores en alta laboral na SS a 31/12 do ano en curso. Organismo responsable: Tesorería General de la Seguridad Social.
- Pensións da Seguridade Social: pensións xestionadas polo INSS e polo ISM que figuran no Rexistro de Prestacións Socias Públicas, do cal é titular o INSS. Organismo responsable: INSS e ISM
- Cartociudad: proxecto colaborativo liderado polo Instituto Geográfico Nacional (IGN) de produción e publicación mediante servizos web de datos espaciais de cobertura nacional. Contén información da rede viaria continua coas rúas, con portais, e as estradas, con puntos quilométricos. Dispón dun servizo de procesamento web baseado en cálculos programados que operan sobre a información xeorreferenciada. Organismo responsable: IGN
- Planimetría das seccións censuais de Galicia, con data 1 de xaneiro do ano de referencia dos datos. Organismo responsable: INE.
- Modelo de Direcciones de la Administración General del Estado (MDAGE). Organismo responsable: INE.
- Catastro: información alfanumérica e cartografía catastral de todos os concellos de Galicia. Organismo responsable: Dirección General del Catastro.
- Rueiro do censo electoral. Organismo responsable: INE
- Mapa das parroquias de Galicia. Organismo responsable: IET
- Mapa de entidades singulares de Galicia: elaborado a partir dos datos do Censo de Vivendas do ano 2011 e doutros datos dispoñibles no IET. Organismo responsable: IET
- Cartografía dos Lugares do Nomenclator Galicia. Organismo responsable: IET

Definicións e conceptos

Directiva INSPIRE. A Directiva INSPIRE (Infrastructure for Spatial Information in Europe) determina as regras xerais para o establecemento dunha Infraestrutura de Información Espacial na Unión Europea. Iníciase ante a necesidade de organizar e poñer en común a información espacial das diferentes Infraestruturas de datos Espaciais dos Estados Membros e co obxectivo de superar os problemas de dispoñibilidade, calidade, xestión, accesibilidade e posta en común de toda a xeo-información.

Sistema de referencia espacial. Un sistema de referencia espacial permite asignar coordenadas a puntos sobre a superficie terrestre. Son utilizados en xeodesia, navegación, cartografía e sistemas globais de navegación por satélite para a correcta xeorreferenciación de elementos na superficie terrestre. Estes sistemas son necesarios dado que a Terra non é unha esfera perfecta.

Sistema de referencia ETRS89- LAEA. Sistema de referencia espacial que a Directiva INSPIRE recomenda para a xeración dunha capa vectorial de celas uniformes de 1 km de lado que sexa homoxénea para toda Europa. Usa o sistema de coordenadas Lambert Azimutal Equal Area (LAEA).

Residente. Considérase residente a toda persoa física que ten a súa residencia habitual nun dos concellos da Comunidade Autónoma de Galicia.

Persoas afiliadas á Seguridade Social. A afiliación ao Sistema da Seguridade Social é obrigatoria para todas as persoas incluídas no campo de aplicación da Seguridade Social e única para toda a vida do traballador e da traballadora e para todo o sistema, sen prexuízo das baixas, altas e demais variacións que con posterioridade á afiliación poidan producirse. É dicir, o traballador ou a traballadora afíliase cando comeza a súa vida laboral e se dá de alta nalgún dos réximes do Sistema da Seguridade Social. Esta situación denomínase alta inicial. Se cesa na súa actividade será dado/a de baixa pero seguirá afiliado/a en situación de baixa laboral. Se retoma a actividade producirase unha alta denominada alta sucesiva a efectos estatísticos, pero non terá que afiliarse novamente, dado que, como xa se indicou, a afiliación é única para toda a vida do traballador ou da traballadora. Nesta actividade estatística difúndese información das persoas en alta laboral na Seguridade Social o 31/12 do ano en curso.

Pensións contributivas da Seguridade Social. Son prestacións económicas e de duración indefinida, aínda que non sempre, nas que a concesión está xeralmente supeditada a unha previa relación xurídica coa Seguridade Social (acreditar un período mínimo de cotización en determinados casos, ...), sempre que se cumpran os demais requisitos esixidos. A súa contía determínase en función das achegas efectuadas polo traballador e o empresario, se se trata de traballadores por conta allea, durante o período considerado para os efectos da base reguladora da pensión de que se trate. As clases de pensións son: incapacidade permanente, xubilación, viuvez, orfandade e a favor de familiares.

Procesamento de datos

Para a realización desta actividade estatística realizáronse os seguintes pasos:

1. Depuración dos códigos que identifican o distrito e a sección censuais de residencia no Padrón continuo
2. Ligazón do Padrón estatístico e o Padrón continuo
3. Xeorreferenciación dos portais do Padrón estatístico
4. Cruce co ficheiro de afiliacións en alta laboral á Seguridade Social
5. Cruce co ficheiro de prestacións da Seguridade Social.

1º Paso: depuración das seccións e dos distritos censuais no padrón continuo

A primeira tarefa que hai que realizar para a fusión dos rexistros é a depuración dos códigos que identifican o distrito e a sección censuais de residencia no Padrón continuo de habitantes.

O Padrón continuo constitúe unha primeira aproximación á poboación de Galicia. Este rexistro, que nos remite o INE dúas veces ao ano, con referencia dos datos o 1 de xaneiro e o 1 de xullo de cada ano, contén o DNI da persoa. Non obstante, trátase dunha primeira aproximación que pode conter erros. O INE somete esta base de datos a un proceso de depuración e remítenos o resultado desta depuración, con data de referencia o 1 de xaneiro. Este segundo "Padrón" é o que se denomina como Padrón estatístico. Esta base non contén o DNI da persoa, que é unha variable moi relevante á hora de cruzar rexistros administrativos; polo que tratamos de vincular o Padrón estatístico co continuo a 1 de xaneiro para poder contar co DNI da persoa.

O obxectivo neste punto é depurar as variables que identifican o distrito e a sección de residencia de cada persoa no Padrón continuo, co obxecto de maximizar, a posteriori, o número de rexistros que ofrece o cruce dos Padróns continuo e estatístico. Para elo, recorremos ao Rueiro do censo electoral, que ofrece unha relación sistemática e actualizada a 1 de xaneiro do ano en curso das vías e tramos de vía que pertencen a cada sección censual. En particular, detectouse que o Padrón continuo contén erratas nestas dúas variables:

- Nas zonas urbanas, onde as vías soen dispor dun código identificativo (*cvia*) e as vivendas soen estar numeradas (*numer*), a meirande parte das erratas advertidas teñen que ver con...
    - Rexistros asignados a unha vía (*cvia*) e concello concretos, cando no Rueiro non existe tal código de vía nese concello
    - Rexistros asignados a unha vía (*cvia*) cunha numeración de vivenda (*numer*) que non se atopa entre os extremos inferior e superior de numeración desa vía no Rueiro (*ein* e *esn*)
- Nas zonas rurais, onde as vías non soen dispor de código identificativo (*cvia=0*) e as vivendas, moitas veces, non se atopan numeradas (*numer* en branco), a meirande parte das erratas débense a que o código que identifica a entidade colectiva (*cun*) figura asociado con códigos distintos para o distrito e a sección (*dist+secc*) en ambas bases de datos (Padrón continuo e Rueiro).

As variables fundamentais coas que se traballará para intentar depurar o distrito e a sección do Padrón continuo a partir do Rueiro do censo electoral son as seguintes:

- *cpro*: código da provincia de residencia (chámase igual no Padrón e máis no Rueiro)
- *cmun*: código do concello de residencia (chámase igual no Padrón e máis no Rueiro)
- *cun*: código que identifica a entidade de residencia (chámase igual no Padrón e máis no Rueiro); trátase dunha variable creada polo INE concatenando...
    - O código da entidade colectiva (2 díxitos)
    - O código da entidade singular (2 díxitos)
    - Un díxito de control
    - O código do núcleo (99 se é diseminado)
- *cvia*: código que identifica a vía (chámase igual no Padrón e máis no Rueiro)

- *tinum*: variable que toma valor 0 se a vía non se atopa numerada, 1 para o tramo de numeración impar e 2 para o tramo de numeración par (chámase igual no Padrón e máis no Rueiro)
- *numer*: número da vivenda de residencia (só se atopa no Padrón continuo)
- *ein*: extremo inferior de numeración da vía (só se atopa no Rueiro)
- *esn*: extremo superior de numeración da vía (só se atopa no Rueiro)
- *cein* (Rueiro)/*cnumer* (Padrón): cualificador do extremo inferior de numeración
- *cesn* (Rueiro)/*cnumers* (Padrón): clarificador do extremo superior de numeración
- *nviac*: nome corto da vía (atópase en ambas bases de datos, pero só imos empregar os valores do Padrón continuo)
- *nentsic*: nome curto da entidade singular de poboación (atópase en ambas bases de datos, pero só imos empregar os valores do Rueiro)

2º Paso: ligazón do Padrón estatístico e o Padrón continuo

O obxectivo neste punto é tomar como base da fusión o Padron estatístico e cruzar co Padrón continuo (co distrito e sección depurados), coa finalidade de crear unha táboa auxiliar que sirva de nexo entre ambos os dous rexistros. Desta forma, poderase recuperar de forma rápida e sinxela o documento identificativo da persoa (xa se trate do DNI, do pasaporte/DNI UE ou do NIE) do Padrón continuo, cando sexa útil empregar esta variable no cruce entre o Padrón estatístico e outros rexistros administrativos.

As variables fundamentais coas que traballamos para cruzar son as seguintes:

- *cpro*: código da provincia de residencia
- *cmun*: código do concello de residencia
- *cvia*: código que identifica a vía de residencia
- *dist* (Padrón estatístico) e *dist_dep* (Padrón continuo): código do distrito censual de residencia
- *secc* (Padrón estatístico) e *secc_dep* (Padrón continuo): código da sección censual de residencia
- *sexo*: código identificativo do sexo da persoa (1 se é home, 6 se é muller)
- *anno+mes+dia* (Padrón estatístico) e *fnac* (Padrón continuo): data de nacemento
- *nomb*: nome da persoa
- *ape1*: primeiro apelido da persoa
- *ape2*: segundo apelido da persoa
- *cpron*: código da provincia de nacemento
- *cmunn*: código do concello de nacemento

Nalgunhas, como é o caso da variable *sexo*, a unión é automática, posto que se trata dunha variable categórica que só pode tomar dous valores, 1 ou 6. Neste caso, o cruce realízase

directamente co programa SQL Server, na contorna da propia base de datos na que se grava a información. Non obstante, para cruzar por medio de variables de tipo carácter, coma o nome, que poden diferir para unha mesma persoa dunha táboa a outra (por exemplo, nun rexistro pode figurar o nome simple da persoa, “MARÍA”, e no outro o composto, “MARÍA DO CARME”), é preciso recorrer a cruces indirectos, que non busquen similitude exacta, senón grao de semellanza. Realizar este tipo de cruces coas ferramentas de consulta de SQL é complicado, polo que se empregan as librarías *stringdist* e *fuzzyjoin* (Robinson, 2016) do software libre de programación R. Ambas permiten comparar o grao de similitude que existe entre cadeas de texto ou entre variables numéricas dun xeito rápido e sinxelo, vinculando cada rexistro dunha base co rexistro da outra base de datos co que garde maior grao de semellanza. Polo tanto, unha parte dos cruces realizarase en SQL e a outra en R.

O procedemento de cruce consta de varios pasos; no primeiro, procédese á unión de rexistros por medio de todas aquelas variables comúns nas dúas táboas, as mencionadas no parágrafo precedente. Con este primeiro paso lógrase vincular a máis do 80% dos rexistros de ambos Padróns, continuo e estatístico. Para os restantes, procédese a realizar unha nova unión, prescindindo neste segundo paso dunha das variables de cruce: o sexo da persoa. O procedemento continúa, relaxando criterios de semellanza en cada novo paso (o lugar de residencia, o lugar de nacemento, etc.), ata que se consegue completar a unión de máis do 98% dos rexistros do Padrón estatístico. En todo paso, esíxese que o cruce sexa biunívoco, isto é: se dous ou máis rexistros do Padrón continuo cumpren os requisitos de semellanza cun dos rexistros do estatístico, prescíndese do cruce. Ademais, a cada paso do procedemento de unión asígnaselle unha calidade de precisión, que vai de 1 (máxima precisión) a 12 (mínima precisión aceptable). Logo de completado, compróbase que as unións son correctas, en particular nos cruces con precisión baixa (valores elevados na variable que amosa a calidade), para corroborar que se trata da mesma persoa no Padrón estatístico e no continuo. Para realizar estas comprobacións, recórrese tamén á busca da información da persoa nos rexistros de anos precedentes, en particular cando se trata de unións onde a coincidencia en variables clave coma o nome e os apelidos non é exacta ou cando cambia a data de nacemento dun Padrón a outro. Este tipo de contraste do historial da persoa nos rexistros administrativos resultou ser moi útil para acadar un alto grao de fiabilidade da unión realizada entre as bases de datos de poboación.

3º Paso: xeorreferenciación dos portais

Neste punto exponse a metodoloxía empregada para xeorreferenciar os portais onde reside a poboación de Galicia. A cada un dos portais asignaráselle unha coordenada X-Y nun sistema de Referencia estándar.

Neste procedemento de xeorreferenciación botouse man tamén do software libre *R* e dos diversos paquetes cartográficos que ofrece. O obxectivo neste punto é xeorreferenciar a poboación dispoñible nos portais do Padrón estatístico.

A fonte principal que se emprega para xeorreferenciar é a base de datos proporcionada polo INE MDAGE. Nesta base de datos están dispoñibles as aproximacións postais (portais) en Galicia coas súas coordenadas UTM X, Y no FUSO 30.

As variables que se empregarán para xeorreferenciar son:

- *cvia*: código de vía
- *tvia*: tipo de vía
- *nvia*: nome da vía
- *num*: número do portal
- *cualificador*: cualificador do portal
- *cec*: código de entidade colectiva
- *ces*: código de entidade singular
- *cnuc*: código de núcleo ou diseminado
- *dis*: código de distrito censual
- *secc*: código de sección censual

Estas variables están dispoñibles nas dúas bases de datos.

Hai que resaltar que a base de datos MDAGE non é completa, existen portais que non teñen unhas coordenadas X, Y válidas ou que non están xeorreferenciados. De todos os xeitos, esta será a fonte de información principal, aínda que despois se complete con outras fontes auxiliares.

Polo tanto, a estratexia que se empregará para xeorreferenciar a poboación será realizar un procedemento concello a concello. Primeiro xeorreferenciarase as vías que teñen *cvia*>0 e que se corresponde (aproximadamente) coa parte máis urbana de Galicia e a continuación xeorreferenciaranse as vías con *cvia*=0, que se corresponde (aproximadamente) coa parte máis rural de Galicia. Dentro de cada concello os pasos a seguir descríbense a continuación:

Parte urbana:

Paso 1.- Xeorreféncianse as rúas que teñen dispoñible o **código de vía**, **tipo de vía**, **nome de vía** e **número**. Neste caso crúzase a base de datos MDAGE co Padrón estatístico empregando as variables anteriores e asígnaselle as coordenadas correspondentes:

- Neste proceso xorde o problema de que na MDAGE existen casos onde dous portais teñen os mesmos códigos de vía e número, mais distintas coordenadas. Para resolver este problema calculouse a media das coordenadas e asignóuselle a ambos os dous portais este valor medio.
- Para as rúas que teñen o mesmo código de vía pero o número do Padrón estatístico non está dispoñible no MDAGE o que se fixo foi aplicar a regresión con splines. Terase

en conta se os números do portal son pares ou impares. Aplicando o seguinte modelo con splines obteríamos a predición para as coordenadas X, Y do número novo:

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

onde

- *Y* coordenada xeográfica UTM Y no fuso 29,
- *X* coordenada xeográfica UTM X no fuso 29,
- *I* indicadora da paridade do número (1 se é par; 0, impar),
- *e* variable aleatoria N(0, σ^2),
- *f* función suave,
- *i*=1, ..., nº total de portais dispoñibles da vía.

Para o seu axuste empréganse os splines penalizados.

Unha vez establecido este primeiro modelo necesítase un segundo modelo que permita calcular a coordenada xeográfica X dun novo portal que se vai imputar. O planteamento para cada unha das vías neste segundo modelo é:

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

onde

- *X* coordenada xeográfica UTM X no fuso 29,
- *N* nº do portal do inmoble,
- *I* indicadora da paridade do número (1 se é par; 0, impar),
- *g* función suave,
- *e'* variable aleatoria N(0, σ'^2),
- *i*=1, ..., nº total de portais dispoñibles da vía.

Para o axuste deste segundo modelo tamén se empregarán os splines penalizados.

No caso de que non haxa suficientes datos para aplicar a regresión optouse por asignarlle a coordenada do número máis próximo.

Paso 2. Para os portais que quedan sen xeorreferenciar no paso 1 empregarase Cartociudad que achegándolle o nome da vía, o número e o concello devolve as coordenadas X, Y.

Paso 3. Para os portais que quedan sen xeorreferenciar no paso 2 empregarase a información que proporciona a Dirección General de Catastro. Esta cartografía ofrece as coordenadas das parcelas onde se ubican os inmobles. O problema que ofrecen estas bases de datos é que non existe un nexo de unión entre os ficheiros do Padrón estatístico e os ficheiros do Catastro. Como nos dous casos dispoñemos de enderezos dos inmobles, o que se fará será unir por enderezos tratando de vincular as **vías** que nos proporciona a Padrón estatístico coas **vías** que proporciona a Dirección General de Catastro nos seus arquivos. Para isto empregarase un procedemento baseado en unir táboas mediante variables que non teñen unha coincidencia exacta. Empregarase o paquete de *R text2vec* (Selivanov et *al*., 2020). Tamén se empregará o mapa de seccións censuais, de tal maneira que a vía do Catastro ten que estar na mesma sección censual que a vía dispoñible no Padrón estatístico.

Unha vez feita a correspondencia entre o código de vía do Padrón estatístico e o código de vía do catastro, é posible que no catastro non estean todos os números dos portais. Neste caso poden ocorrer dúas situacións diferentes:

1- Algúns números que achega catastro son ceros

   Se os números de portal que achega catastro teñen ceros non se pode saber se se trata de números pares ou impares. Para averiguar isto empregaremos a análise de compoñentes principais (ACP). Aplicaremos a ACP ás coordenadas X, Y que nos achega o Catastro. Pódese comprobar que a primeira compoñente principal correspóndese coa dirección da vía e a segunda, ortogonal á primeira, cos pares e impares. Unha vez que se sabe cales son os puntos pares e impares, xa se pode calcular o centroide das coordenadas dos pares e dos impares e asignarllo ao número que se quere xeorreferenciar, dependendo de que sexa par ou impar.

2- Ningún número é nulo

   Nesta situación para determinar as coordenadas do número novo aplicaremos a regresión con splines, método empregado no paso 1.

Paso 4 Para aquelas rúas que non se xeorreferencian co paso anterior empregarase a API de Google e os servizos da API de Here. En concreto, no caso de Google, empregarase a función de *R geocode do paquete ggmap* (Kahle and Wickham, 2013) que achegándolle o nome da vía, o número do portal e o concello devolve as coordenadas X, Y. No caso de Here, dende *R* chamarase a API co nome de vía, o número de portal e o concello, e obterase de volta as coordenadas X,Y.

Parte rural:

Paso 1.- Xeorreféncianse os portais que están nas entidades que non teñen dispoñible a clase de vía, código ou nome de vía. Neste caso a xeorreferenciación realízase a nivel de **núcleo/diseminado** e número (do portal) do seguinte xeito:

- Para aqueles portais que teñen un número no Padrón estatístico e este número está dispoñible no MDAGE, asígnaselle a coordenada dispoñible no MDAGE.
- Para aqueles portais que teñen un número no Padrón Estatístico e este número non está dispoñible no MDAGE asígnaselle o portal máis próximo de Catastro dentro da entidade singular onde está ubicado. Emprégase para isto o mapa de entidades singulares do IET.

Paso 2. Para os portais que quedan sen xeorreferenciar no paso 1 síguese o seguinte procedemento:

- Para os portais non xeorreferenciados no punto anterior buscouse o máis próximo de Catastro dentro da parroquia á que pertencen. Emprégase para isto o mapa parroquias do IET e a función de R *knn* do paquete *class* (Venables and Ripley, 2002) que busca os veciños máis próximos.
- Para os restantes portais buscouse o lugar máis próximo dentro da cartografía de Lugares do IET dos LugaresNomenclatorGalicia, e asignóuselle o centroide do Lugar.

Paso 4º: cruce co ficheiro de afiliacións en alta laboral á Seguridade Social

No 4º paso do procedemento de construción da táboa de datos que servirá de base para a extracción das características socioeconómicas da poboación de Galicia en celas dun 1km2 tratarase de anexar a información dos ficheiros de afiliacións en alta laboral da Seguridade Social. Para elo, aprovéitanse as tarefas internas realizadas no marco da operación estatística *Afiliacións á Seguridade Social por concello de residencia da persoa afiliada*, que publica cada 3 meses o IGE con información dende o ano 2006. O obxectivo desta operación é ofrecer información da evolución da afiliación ao Sistema da Seguridade Social en Galicia, cun nivel de desagregación territorial que chega ao concello de residencia da persoa afiliada. Nos ficheiros que subministra a Tesourería da Seguridade Social figura esta variable, o concello de residencia do afiliado ou da afiliada pero, nunha parte importante dos rexistros, está desactualizada, reflectindo o concello no que residía a persoa cando se deu de alta no Sistema e non a residencia actual. Por este motivo, procédese a cruzar estes ficheiros de afiliación co Padrón continuo de habitantes, a 01/01 do ano de referencia nas táboas da Seguridade Social subministradas a 31/03 e 30/06 do ano en cuestión, e co Padrón continuo a 01/07 para os ficheiros de afiliación relativos a 31/09 e 31/12. Para analizar o procedemento de cruce entre o Padrón continuo e os rexistros trimestrais de afiliacións á Seguridade Social, pódese consultar o proxecto técnico da operación, na seguinte ligazón:
https://www.ige.eu/estatico/pdfs/s3/proxectosTecnicos/24-106_AfiliacionsASeguridadeSocialporConcelloResidenciaAfiliado.pdf

Feita esta unión no marco da operación reseñada e, posto que no 2º paso se cruzaron os Padróns estatístico e continuo, a unión entre o Padrón estatístico e os rexistros de afiliación á Seguridade Social é automática. Non obstante, queda unha parte moi pequena do Padrón estatístico, aquela que non se puido vincular co continuo, para a que non se ten correspondencia co rexistro de Afiliacións á Seguridade Social. Para estes rexistros aplícase un procedemento de unión específico entre o Padrón estatístico e os ficheiros de afiliación, moi similar ao empregado no caso da unión entre Padróns: só se permiten cruces biunívocos, vanse relaxando criterios de similitude/semellanza entre variables en cada paso do procedemento de unión (no primeiro paso empréganse todas as variables comúns a ambos os dous ficheiros, no segundo elimínase unha delas, no terceiro dúas, etc.) e créase unha variable *calidade* que indica o grao de precisión da unión. Finalmente, contrástase que se trata da mesma persoa en ambos os dous ficheiros, no Padrón estatístico e nos rexistros de afiliacións á Seguridade Social.

Paso 5º: cruce co ficheiro de prestacións da Seguridade Social

Neste punto o obxectivo é cruzar o Padrón estatístico co ficheiro de Pensións contributivas xestionadas polo INSS e polo ISM e proporcionadas ao IGE con data de referencia o 31 de decembro de cada ano.

O procedemento empregado para facer o cruce é o seguinte:

Paso 1.- Se coincide o identificador (DNI, pasaporte e DNI europeo ou documento de estranxeiro) dos dous ficheiros considérase que a persoa é a mesma.

Paso 2.- Se non se pode efectuar o cruce no paso 1, procédese a facer a ligazón por nome, apelidos, sexo e data de nacemento. Se coinciden estas variables nos dous ficheiros considérase que a persoa é a mesma.

Paso 3.- Se non se pode efectuar o cruce nos pasos 1 e 2 , procédese a facer a ligazón por nome, apelidos e sexo. Se coinciden estas variables nos dous ficheiros considérase que a persoa é a mesma.

Paso 4.- Se non se pode efectuar o cruce nos pasos 1, 2 e 3 , procédese a facer a ligazón por nome, primeiro apelido e data de nacemento. Se coinciden estas variables nos dous ficheiros considérase que a persoa é a mesma.

Paso 5.- Se non se pode efectuar o cruce nos pasos 1, 2, 3 e 4 , procédese a facer a ligazón por nome, segundo apelido e data de nacemento. Se coinciden estas variables nos dous ficheiros considérase que a persoa é a mesma.

Paso 6.- Se non se pode efectuar o cruce nos pasos 1, 2, 3, 4 e 5 , procédese a facer a ligazón por apelidos, sexo e data de nacemento. Se coinciden estas dúas variables nos dous ficheiros considérase que a persoa é a mesma.

Paso 7.- Se non se pode efectuar o cruce nos pasos 1, 2, 3, 4 ,5 e 6, procédese a facer a ligazón por nome, apelidos e por datas de nacemento onde a distancia sexa menor de 3. Se se compren as condicións considérase que a persoa é a mesma.

Neste punto hai que destacar que no 1º paso cruzan aproximadamente o 98% dos rexistros de pensións e hai entorno a 10.000 pensións que non se consigue cruzar co Padrón estatístico e co procedemento establecido.

Deseño da explotación dos datos da fusión dos rexistros

Á información fusionada e xeorreferenciada asignóuselle as celas da malla regular de 1 km^2 xerada. O sentido desta operación atópase na necesidade, por unha parte, de protexer a confidencialidade da información, e pola outra, de facilitar a representación cartográfica da información.

Na operación de xeocodificación obtivéronse resultados que non chegan ao 100% da totalidade da poboación, en concreto para o ano 2018 xeocodificouse o 99,9% da poboación de Galicia.

O resultado final é unha capa vectorial composta por unha malla regular de celas cadradas de 1km de lado. Para cada unha destas celas preséntase a información estatística de carácter

sociodemográfica que lle corresponde, desagregada da seguinte forma (entre parénteses o código da variable):

- Poboación de homes (homes)
- Poboación de mulleres (mulleres)
- Índice de feminidade: Mulleres/ Homes *100 (ratio_femi)
- Porcentaxe de poboación menor de 16 anos (porc_menor_16)
- Porcentaxe de poboación entre 16 e 64 anos (porc_maior15_me)
- Porcentaxe de poboación con 65 anos ou máis (porc_maior_64)
- Idade media da poboación: media das idades da poboación (idadeMedia)
- Índice de envellecemento: poboación >65 anos/poboación <20 anos*100 (ind_envelle)
- Poboación con nacionalidade estranxeira (porc_estranx)
- Poboación residente no mesmo concello de nacemento (porc_mesmo_conc)
- Poboación residente nada en Galicia nun concello diferente ao de nacemento (porc_outroconce)
- Poboación nada noutra Comunidade Autónoma (porc_OutraCCAA)
- Poboación nada noutro país (porcNacExtr)
- Afiliacións á Seguridade Social (afiliacions)
- Persoas afiliadas á Seguridade Social (afiliados)
- Porcentaxe de homes afiliados á Seguridade Social (porc_hom_afi)
- Porcentaxe de mulleres afiliadas á Seguridade Social (porc_mull_afi)
- Ratio de feminidade da poboación afiliada (ratio_fem_afi): Mulleres afiliadas/Homes afiliados*100
- Porcentaxe de poboación entre 16 e 34 anos afiliada á Seguridade Social (porc_16_34_afi)
- Porcentaxe de poboación entre 35 e 54 anos afiliada á Seguridade Social (porc_35_54_afi)
- Porcentaxe de poboación de 55 ou máis anos afiliada á Seguridade Social (porc_55_mas_afi)
- Porcentaxe de poboación de nacionalidade estranxeira afiliada á Seguridade Social (porc_estran_afi)
- Porcentaxe de afiliacións no Réxime Xeral e minería do carbón (porc_xeral_afi)
- Porcentaxe de afiliacións no Réxime Especial de Autónomos (porc_auto_afi)
- Porcentaxe de afiliacións á Seguridade Social na agricultura e a pesca (porc_prima_afi)
- Porcentaxe de afiliacións á Seguridade Social na industria (porc_indus_afi/porc_cons_afi)
- Porcentaxe de afiliacións á Seguridade Social na construción (porc_cons_afi)
- Porcentaxe de afiliacións á Seguridade Social nos servizos (porc_serv_afi)
- Poboación perceptora de pensións contributivas da Seguridade Social (Pensionistas)

- Homes perceptores de pensións contributivas da Seguridade Social (Pensionistas_ho)
- Mulleres perceptoras de pensións contributivas da Seguridade Social (Pensionistas_mu)
- Índice de feminidade das persoas perceptoras de pensións contributivas da Seguridade Social (ratio_fem_pens): Mulleres perceptoras de pensións/Homes perceptores de pensións*100
- Pensión media dos pensionistas das Seguridade Social: importe bruto mensual da prestación en € (pensionMedia)

Segredo estatístico

Garantiuse a protección e a confidencialidade da información mediante o método de eliminación de datos. Inicialmente realizouse unha eliminación dos datos considerados sensibles e, posteriormente, cando foi necesario, unha supresión secundaria que impida a identificación exacta por dedución dos datos sensibles censurados, por diferenza con respecto ao total. En concreto, os criterios empregados foron os seguintes:

- Eliminároнse as celas que teñen unha poboación total menor de 21 habitantes.
- Censuráronse as celas que teñen menos de 5 habitantes (>0) nalgunha das características sociodemográficas polas que se clasifica.
- Censuráronse celas complementarias para evitar a dedución de datos sensibles censurados no paso anterior por diferenza respecto ao total. Neste caso eliminároнse os datos tendo en conta o seu valor, coa intención de minimizar o custo en termos de perda de información

Polo tanto, os criterios aplicados impiden a identificación exacta dos datos considerados sensibles, non a súa aproximación. O proceso de eliminación de datos realizouse aplicando un algoritmo desenvolvido ad-hoc que cumpre os criterios de segredo estatístico establecidos.

O territorio de Galicia cóbrese cun total de 30.776 celas de 1 km^2. No ano 2018, 19.021 celas tiñan poboación. Destas, difúndense 10.775 celas que teñen unha poboación superior a 20 persoas. Estas celas concentran o 97,2% da poboación de Galicia, polo tanto, as 8.246 celas que non se pode difundir por problemas de segredo estatístico, concentrarían o 2,8% da poboación de Galicia. Este sería o custo en termos de perda de información que hai que asumir polo cumprimento do segredo estatístico.

Control de calidade

Realizouse unha revisión da información espacial dos portais utilizados para a xeración da malla de poboación. En concreto realizáronse os seguintes contrastes:

- Os portais xeorreferenciados deben estar xeograficamente localizados no concello do Padrón estatístico.

- Os portais xeorreferenciados deben estar xeograficamente localizados na sección do Padrón estatístico ou na mesma entidade singular ou, no seu defecto, na entidade singular veciña ou na mesma parroquia.

Difusión da información

Desenvolveuse un visor cartográfico de fácil manexo e moi intuitivo. No mesmo pódese consultar a información en forma de mapas. A cada cela asígnaselle unha cor segundo o intervalo no que se encontre a variable sociodemográfica que se vai representar. Os datos de celas con poboación menor de 21 habitantes non se representan e os restantes datos censurados aparecerán representados cunha cela branca.

O visor permite realizar desprazamentos por todo o territorio, así como todos aqueles zooms que mostren ao usuario a información no nivel de detalle que desexe. Estes niveis van desde o máis xeral que permite a visión global de toda Galicia, ata escalas urbanas de gran detalle. O visor permite tamén cambiar dun mapa a outro conservando a vista do mapa anterior.

Existe tamén a posibilidade de consultar os datos de cada cela, mediante o recurso de picar para que apareza un cadro emerxente, no que se mostra a información do concello onde se localiza a cela, o número de persoas residentes na mesma e a característica elixida para representar no mapa: homes, mulleres, número de persoas perceptoras de pensións contributivas, ...

Tamén se pode consultar a información en formato táboa, descargable en formato folla de cálculo.

Por último, incluíuse a consulta da información das celas mediante unha API interoperable que permite obter información para as celas de:

- Galicia: http://www.ige.eu/igebdt/igeapi/csv/grid1km/<ano>/0/0/<código das variables separados por />
- Provincias: http://www.ige.eu/igebdt/igeapi/csv/grid1km/<ano>/<código da provincia>/0/<código das variables separadas por />
- Un concello: http://www.ige.eu/igebdt/igeapi/csv/grid1km/<ano>/<código da provincia>/<código do concello>/<código das variables separadas por />

A posta a disposición da información en formatos interoperables contribúe ao desenvolvemento de procesos de xeración de valor engadido baseados na reutilización da información por parte da Administración Pública, os axentes económicos, sociais, así como, da cidadanía en xeral.

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