Hacia un registro estadístico del territorio

Eduard Suñé Luis

Àrea de Població i Territori. Subdirecció General de Producció i Coordinació. Institut d'Estadística de Catalunya. Via Laietana 58. Barcelona. esl@idescat.

RESUMEN

Desde hace ya unas décadas, y sobre todo en los países nórdicos, están en funcionamiento infraestructuras de información basadas en registros administrativos con fines estadísticos.

La razón de todo ello es que la administración genera información suficiente sobre la población y las empresas, de manera que las operaciones masivas como los censos de población, podrían tener una alternativa mucho menos costosa si se utilizan este tipo de infraestructuras.

En este contexto, las variables territoriales (donde ocurren los eventos) tendrían que integrarse adecuadamente en el sistema, es decir, el registro estadístico de población o de empresas debería acompañarse de un registro estadístico de territorio para poder describir y validar las variables territoriales.

En la presente comunicación se propone un modelo de datos mínimo que podría conformar el registro estadístico de territorio y se analiza su relación con los modelos existentes, como la base de datos de direcciones de Catalunya (BDMAC), el modelo de direcciones de la Administración General del Estado, Cartociudad y Catastro.

Por último se propone una solución para la construcción de un registro estadístico de territorio en el ámbito de la Comunidad Autónoma de Catalunya.

Palabras clave: *Territorio, registros administrativos, estadística, georeferenciación*

INTRODUCCIÓN

El Institut d'Estadística de Catalunya (IDESCAT) ha ido realizando, mediante el servicio de geocodificación del Institut Cartogràfic i Geològic de Catalunya (ICGC), los trabajos de georeferenciación e imputación de coordenadas del Registro de Población desde el año 2012 en adelante. Estos trabajos permiten la obtención de una capa de puntos con las posiciones correspondientes al conjunto de direcciones que aparecen en el Registro de Población.

No obstante, IDESCAT, ha iniciado una nueva línea de producción basada en registros administrativos [1] con la finalidad de mejorar la eficiencia en la producción estadística. La idea básica es aprovechar la información que la Administración dispone de los habitantes y empresas y construir sistemas de información que permitan conocer sus características básicas sin necesidad de incurrir en operaciones estadísticas ad-hoc extremadamente costosas.

En primer lugar describiremos el proceso de geocodificación del registro de población 2014, los métodos de imputación empleados, para pasar seguidamente a la descripción del Registro Estadístico de Territorio (RET) y su relación los Registros Estadísticos de Población (REP) y de Empresas y Establecimientos (REEE)

GEOCODIFICACIÓN DEL REGISTRO DE POBLACIÓN 2014.

El Registro de Población es un producto derivado del Padrón de Habitantes, gestionado por el Instituto Nacional de Estadística (INE), que, periódicamente, facilita a Idescat obtiene copias referidas a su ámbito de actuación (Comunidad Autónoma de Catalunya).

En este registro, exhaustivo, encontramos una serie de variables demográficas (edad, sexo, nacionalidad etc.) y otras territoriales: provincia, municipio, distrito, sección censal y dirección postal.

La dirección postal, está descrita por una serie de campos como son el tipo de vía, código de vía, nombre de la vía y unos campos de numeración que describen la parte horizontal y vertical de la dirección postal. Un primer tratamiento, que agrupa los habitantes que residen en una misma dirección permite obtener la base de datos del tipo descrita en la figura 1.

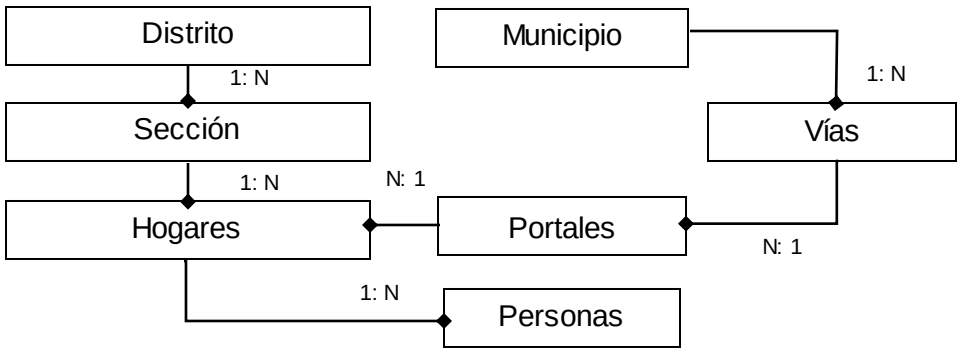


Figura 1: Base de datos asociada al Registro de Población.

La tabla portales representa el conjunto de direcciones postales a nivel horizontal en donde reside alguna persona y es el *target* de nuestro tratamiento.

El Institut Cartogràfic i Geològic de Catalunya (ICGC) dispone de un servicio web [2], publicado bajo el estándar SOAP, que permite la integración en nuestras aplicaciones del conjunto de operaciones que este servicio implementa. Entre ellas, disponemos del típico proceso de geocodificación: obtención de unas coordenadas a partir de una dirección postal.

Este servicio ha ido mejorando a lo largo del tiempo en cuanto al porcentaje de casos resueltos, como puede observarse en la tabla adjunta:

Tabla 1. Resultados de la geocodificación del Registro de población. 2012-2014.

Año	Portales	Resueltos	% Resueltos	% no Resueltos
2012	992.660	918.192	92,5	7,5
2013	990.000	914.316	92,4	7,6
2014	**1.038.595**	**968.166**	**93,2**	**6,8**

En cada petición resuelta el servicio devuelve un indicador de exactitud asociado a las coordenadas. En ciertas partes del territorio, típicamente zonas diseminadas, la exactitud es menor que en zonas urbanas, donde mayoritariamente es a nivel de portal y a lo largo del tiempo ha ido mejorando, alcanzándose para el año 2014 un nivel de exactitud a nivel de portal del 31% para el conjunto de Cataluña y del 70% en Barcelona capital, como puede observarse en la tabla 2.

Tabla 2. Evolución de la exactitud del geocodificador ICGC

Exactitud	2013	2014	2014 Barcelona capital
Portal	-	31 %	70 %
Portal interpolado	~ 70 %	47 %	28 %

Para el año 2014 sólo el 6,8% de los portales tratados no fueron resueltos por el servicio y, como ya se hizo para los años anteriores, se imputaron las coordenadas. No obstante, los métodos de imputación empleados fueron radicalmente diferentes a los anteriores ya que disponíamos de otro tipo de información. En efecto, para los años 2012 y 2013 disponíamos del seccionado, el SIGPAC [3] y un conjunto de fotografías satélite nocturnas de la NOAA [4], empleándose una serie de métodos de imputación adecuados para tratar la información disponible [5].

Para el año 2014 (RP2014) disponíamos de la información del archivo CAT del Catastro y del SIGPAC pero no disponíamos del seccionado. Así pues, los donantes de las posiciones serían los centroides de las fincas del archivo CAT del Catastro, posiciones que hubieran podido ser asignadas directamente si no fuera porque Catastro e INE utilizan un conjunto de códigos de vía diferentes y no existe hoy por hoy una tabla de correspondencias. Este hecho impide la búsqueda por código de las vías Catastro correspondientes a una vía INE, primer paso necesario en la asignación.

La alternativa a esta búsqueda directa pasa por la utilización de los literales de tipo y nombre de vía. Sólo un 50% de esos literales son exactamente iguales por lo que o se intenta construir una tabla de equivalencia entre los dos conjuntos de códigos o, mediante las similitudes de los literales, se construye una tabla lo más completa posible aunque con la posibilidad natural de contener falsos positivos.

Como los costes de construcción de una tabla de correspondencia completa son muy elevados (se trata de unas 100.000 vías) y nuestro problema es la imputación de posiciones se decidió realizar búsquedas difusas (*fuzzy search*) mediante los literales de las vías. Para ello, se deben tomar dos decisiones importantes:

- qué métrica utilizar
- el umbral por encima del cual supondremos que, efectivamente, las dos vías son la misma.

Existen multitud de métricas: Levenshtein [6], Jaro-Winkler [7], Trigrams, etc., así como estudios comparativos entre ellas. Como puede verse en el diagrama de dispersión de los valores de similitud Levenshtein y Jaro-Winkler (figura 2) los resultados van a depender de qué métrica utilizamos ya que aun observándose cierta correlación, esta no es perfecta.

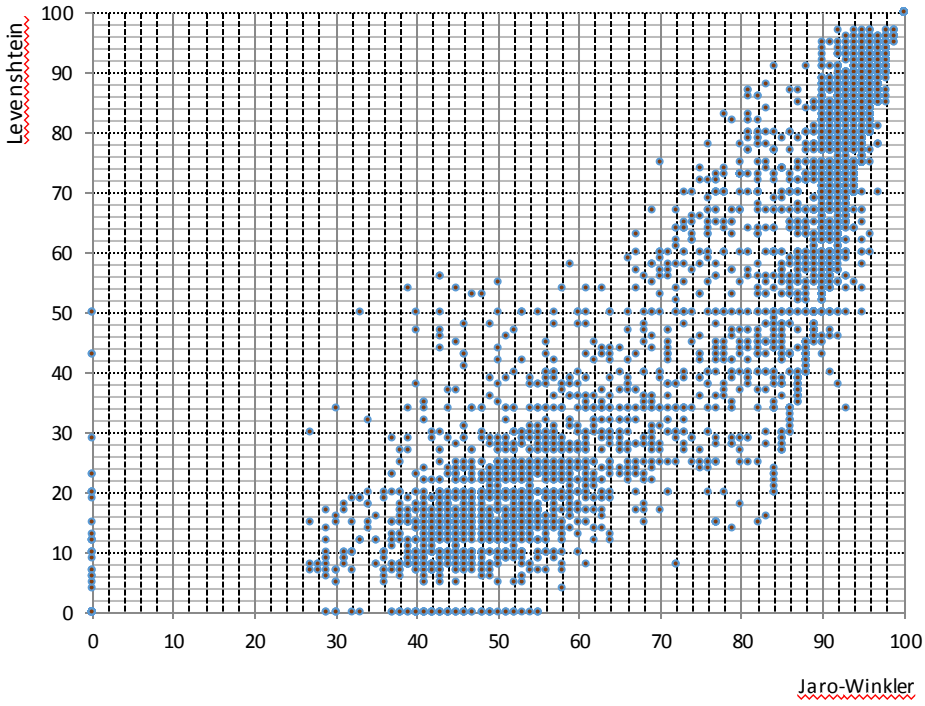


Figura 2: Similitud Jaro-Winkler vs Levenshtein

Otro parámetro a utilizar es el valor de la similitud a partir del cual supondremos que las dos vías son la misma. En la figura 3 puede observarse el número de casos que superan un umbral determinado de los valores de similitud, cuando se calcula para todas las vías INE y todas las vías DGC (municipio a municipio), para las métricas Levenshein y Jaro-Winkler.

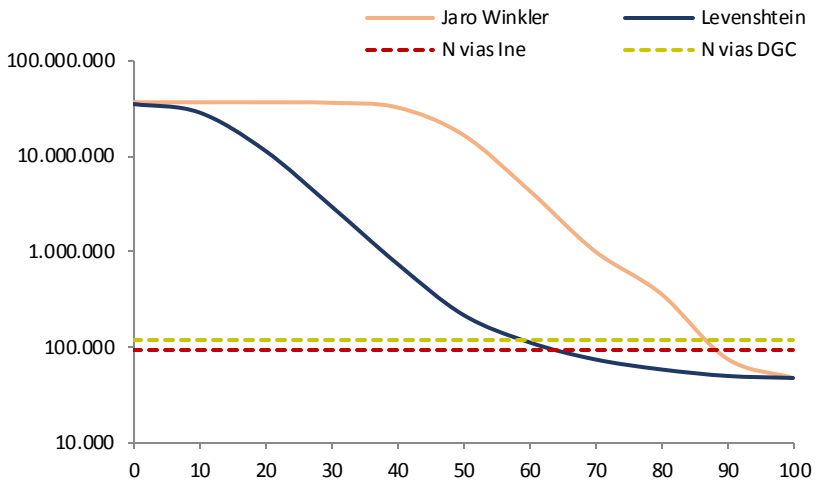


Figura 3: Casos según umbrales de similitud

Un valor de 90 (en un rango de cero a cien) para Jaro-Winkler y de 60 para Levenshein parecen los más adecuados. Finalmente se tomó la decisión de utilizar la métrica de Jaro-Winkler y valor umbral de 90; es decir, si:

$$Sjw(literal\ via\ INE, literal\ via\ DGC) \geq 90 \ \rightarrow via\ INE \equiv via\ DGC \quad (1)$$

No obstante existe otro criterio que puede utilizarse para la construcción de la tabla de equivalencias. Por un lado disponemos, gracias al proceso de geocodificación del RP2014, de una capa de puntos con las posiciones de los portales y por otro de las posiciones de los centroides de las fincas. Puede establecerse una relación entre las dos capas: cada punto geocodificado RP2014 tiene al menos un punto más cercano de la capa fincas. Esta relación *'más cercano'* establece implícitamente un par (vía INE, vía DGC). Si se calcula para todos los puntos cual es la frecuencia relativa de esos pares (vía INE, vía DGC) llegaríamos rápidamente a la conclusión de que si ese valor es muy cercano a uno podríamos suponer que efectivamente se trata de la misma vía.

En la figura 4 pueden observarse los puntos geocodificados (en rojo) y los centroides de las fincas (negro), para el caso de la calle Av. Número 342 del municipio de Castelldefells.

Figura 4: Puntos geocodificados (rojo) y centroides de las fincas (negro)

Se da el caso que para esta calle, la frecuencia relativa de los pares (vía INE, vía DGC) es 1 mientras que la similitud Jaro-Winkler es cero, ya que mientras INE utiliza letras en el literal, Catastro utiliza números. Este es un caso extremo de falso negativo si sólo hubiéramos utilizado el criterio de las similitudes de literales con cualquier métrica.

La tabla que finalmente se ha utilizado tiene en cuenta los dos criterios expuestos: similitud Jaro-Winkler y relación espacial '*más cercano*' entre los puntos geocodificados RP2014 y los centroides de las fincas, y se corresponde con el diagrama de dispersión de la figura 5:

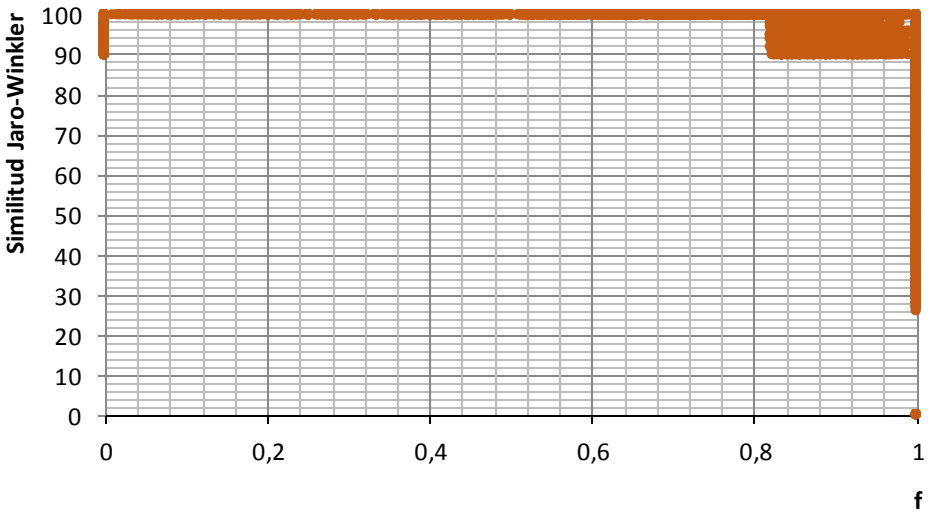


Figura 5: Casos incluidos en la tabla de correspondencias vías INE-DGC

Tiene que tenerse en cuenta que esta tabla de correspondencias lo es en un sentido probabilístico, dados los criterios que se han utilizado en su confección. No está exenta, seguramente, de falsos positivos y no cubre el 100% de las vías INE, pero ha de recordarse que su utilización se limita a la imputación de los casos no geocodficados del RP2014.

Así pues, un caso no geocodificado se corresponde con alguno de estos tres casos:

1. Su vía existe en la tabla de correspondencias y existe una finca con igual numeración.
2. Su vía existe en la tabla de correspondencias pero no existe una finca con igual numeración
3. Su vía no existe en la tabla de correspondencias

En el primer caso la asignación directa es posible. En el resto debería realizarse una imputación pero respetando la sección censal dado que conocemos a priori esa información. En otras palabras, debemos delimitar el espacio para que las fincas donantes estén situadas en la sección censal asignada. Se dio el caso que en el momento en que se realizó todo el tratamiento, la cartografía de las secciones censales no estaba disponible, por lo que fue necesario construir una aproximación (para municipios de más de una sección censal).

La idea básica de esta aproximación consiste en calcular el *convex hull* de los puntos geocodificados agrupados por sección censal. El resultado de esta operación da la sorprendente imagen de la figura (6):

Figura 6: Convex-hull de puntos RP2014 por sección censal

Este resultado se debe a que en relativamente pocos casos la sección censal asignada inicialmente es incorrecta. En la figura 6, el *convex* seleccionado (amarillo) tiene esa forma porque hay un portal que tiene asignada incorrectamente esa sección.

Así pues, debería generarse el *convex* sin utilizar puntos que estuvieran anormalmente lejos del resto de puntos de la sección censal. Estos puntos problemáticos son *outliers* en la distribución de las medias de las distancias respecto al resto de puntos (figura 7).

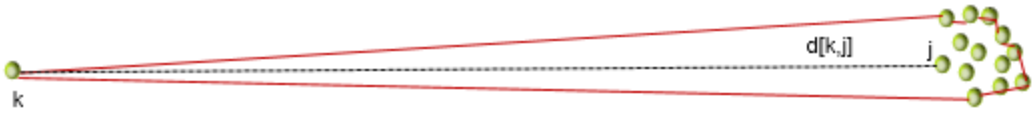


Figura 7: Un conjunto de puntos análogo al seleccionado en la figura 6

Calculando para cada punto la distancia media al resto de puntos de la sección y el z-score asociado, pueden descartarse convenientemente puntos anómalos en la confección del *convex*. El valor umbral que hemos tomado para el z-score ha sido de 0.9; todo punto cuyo valor fuera superior no ha participado en la generación del *convex*.

El resultado para una zona de Barcelona puede observarse en la figura 8.

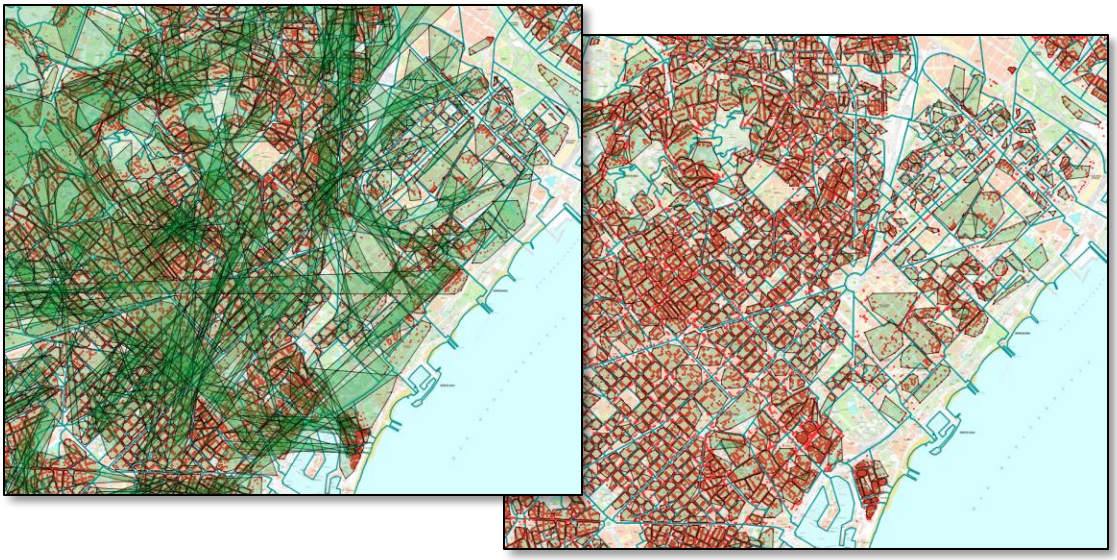
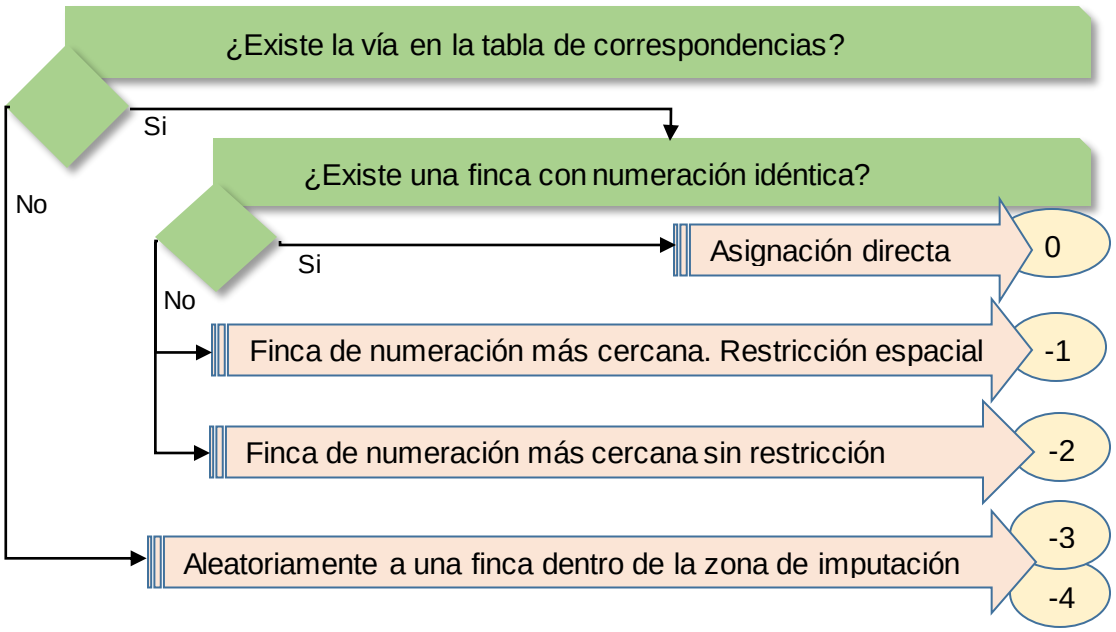


Figura 8: A la izquierda el convex sección a sección sin restricciones. A la derecha convex calculado con aquellos puntos con Z-Score < 0.9.

Así pues, las zonas de imputación que hemos utilizado son:

1. Municipios de sección única: límites municipales con intersección zona urbana SIGPAC.
2. Municipios con más de una sección: *convex-hull* (Z < 0.9) con intersección zona urbana SIGPAC.

Con todo esto el esquema final de imputación es:



Los valores asociados a cada acción (0, -1, -2, -3 y -4) representan la calidad de imputación y se almacenan en la base de datos para cada punto tratado, de igual forma que lo habían sido con el índice de calidad que devuelve el geocodificador del ICGC.

El resultado final de los procesos de geocodificación e imputación es una capa de puntos (figura 9) relacionada con el Registro de Población. Es una información valiosa ya que permite a la Administración tomar decisiones adecuadas en multitud de ámbitos.

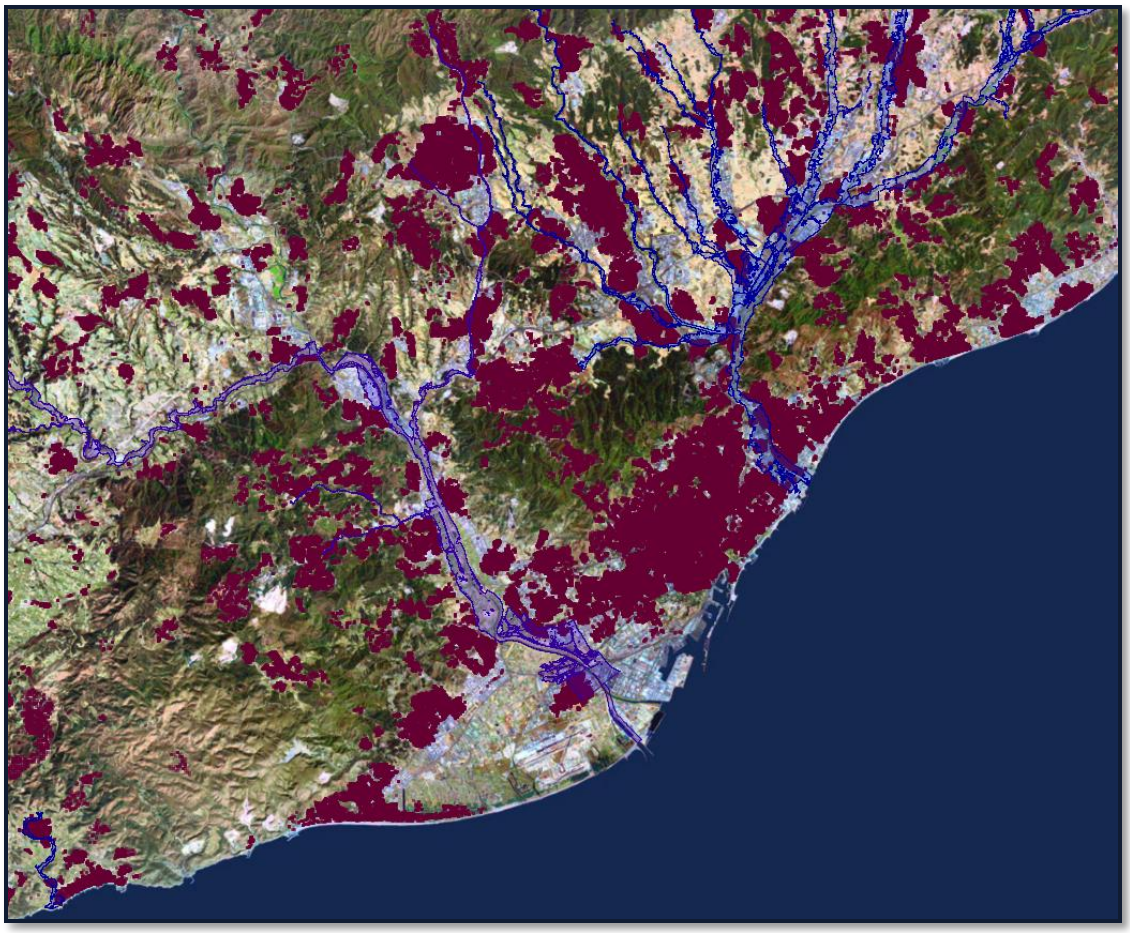


Figura 9: Capa de puntos asociada al Registro de Población 2014.
Detalle de zonas metropolitanas y de inundación de los ríos Llobregat y Besós.

HACIA UN SISTEMA BASADO EN REGISTROS ESTADISTICOS.

Como decíamos en la introducción, IDESCAT ha iniciado una nueva forma de producir estadísticas basadas en la gestión de la información que va generando la interacción del administrado con la Administración, tanto a nivel de personas (Registro Estadístico de Población, REP) como de empresas (Registro Estadístico de Empresas y Establecimientos, REEE).

Quizás la mejor forma de describir el Registro Estadístico de Población (REP) consista en reflexionar acerca de los eventos vitales que pueden ocurrirle a una persona a la largo de su vida.

Desde el nacimiento van ocurriendo o pueden ocurrir una serie de eventos que modifican nuestras características: nos instruimos, trabajamos, migramos, formamos una familia, nos reproducimos, nos jubilamos y más tarde o más temprano acabamos formando parte de las estadísticas de mortalidad.

Pues bien, cada uno de estos eventos produce una interacción con la Administración, de tal forma que, en teoría, con una gestión adecuada, podríamos conocer las características básicas de la población en un instante determinado, partiendo de una situación inicial conocida.

Esta forma de proceder tiene la ventaja de que no haría falta preguntar a la población sobre unas características (edad, sexo, estado civil, nivel de instrucción, relación con la actividad, etc.) de las que la Administración ya tiene conocimiento previo. Ahorramos en costes de producción y generamos menos molestias al administrado.

Tanto en la situación inicial de partida, los datos del Registro de Población a una fecha determinada, como en los eventos anteriormente descritos (nacimientos, matrimonios, defunciones, etc.) aparecen posiciones, en forma de dirección postal, y es necesario crear un sistema de información territorial, que denominaremos Registro Estadístico de Territorio (RET), que le de soporte.

No se trata sólo de conocer las posiciones relativas a estas direcciones postales, sino de validarlas, asignar una clave primaria que se propague convenientemente en el REP (y en el REEE) y relacionar esa dirección con otros atributos. Así, si un hogar reside en una dirección determinada, el RET debería:

- Asignar una clave primaria que se propague a lo largo del REP
- Validar esa dirección
- Proporcionar las coordenadas de esa dirección
- Obtener otras características derivadas de la dirección: superficie del inmueble, año de construcción, características de edificio, etc.

Para este fin, sería necesario disponer como mínimo del conjunto de direcciones en nuestro ámbito de actuación (Comunidad Autónoma de Catalunya), porque en el fondo las dos primeras finalidades, las más importantes, son análogas al problema de validar un código de ocupación (o un literal) contra una clasificación oficial de ocupaciones.

La diferencia estriba en que mientras existe una clasificación oficial para las ocupaciones, no existe una única fuente contra la que validar una dirección postal, tanto a nivel horizontal (portal) como a nivel vertical (dirección postal completa) a día de hoy.

En todo caso, la fuente o fuentes que se vayan a utilizar deben ser de calidad y exhaustivas, ya que su fin primordial será la validación de una dirección.

La asignación y propagación en el REP de una clave primaria para cada dirección horizontal y vertical (*address code*) es siempre posible. Se trata simplemente de extraer del REP el conjunto de direcciones horizontales y verticales y asignar un número de orden, manteniendo, eso sí, la relación horizontal-vertical existente. Con esto obtendríamos una buena clave primaria ya que nunca será nula y siempre será única. Es simplemente una normalización de nuestra base de datos.

La validación contra una fuente determinada (o contra varias) podría entonces obtenerse construyendo una tabla de correspondencias entre nuestra clave primaria (*address code*) y la clave primaria de la fuente (o fuentes) contra la que validamos.

Estas ideas se muestran esquemáticamente en la figura 10.

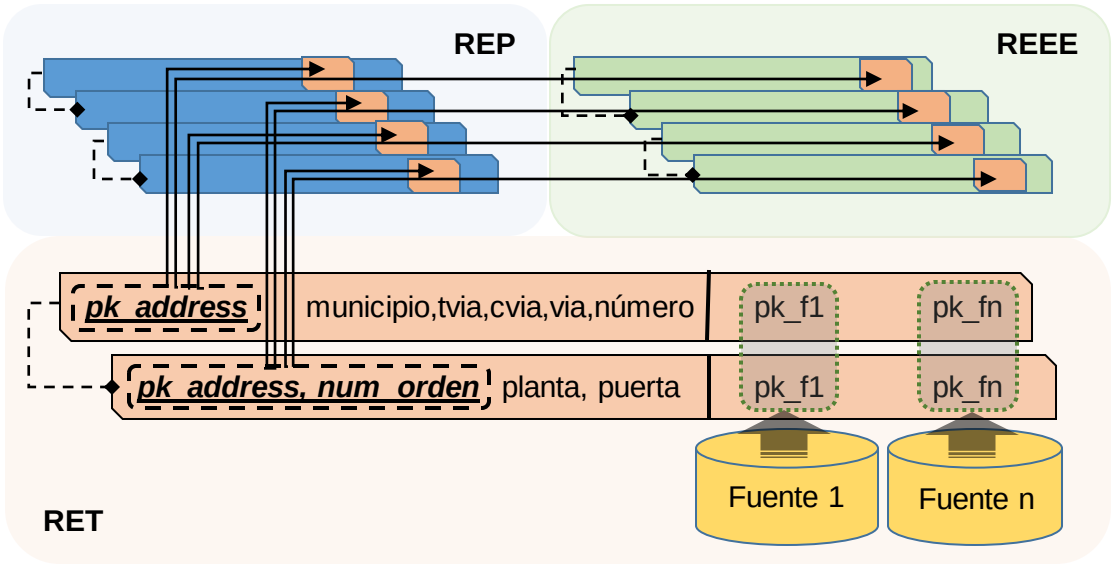


Figura 10: Esquema general del RET y su relación con el REP y REEE.

Es obvio que estas fuentes deben modelar como mínimo direcciones y lo deben hacer tanto en relación a la parte horizontal como a la vertical y, además, deben ser capaces de hacerlo a pesar de la variada casuística que podemos encontrar en el mundo de las direcciones postales. Pero en todo caso, insistimos, han de ser exhaustivas y de calidad ya que uno de los objetivos del sistema, el más importante, es la validación de direcciones postales.

FUENTES DEL RET

Un análisis de las posibles fuentes de información que podrían conformar el RET da como resultado las diferentes alternativas que aparecen en la tabla 3

Tabla 3. Posibles fuentes del RET

Fuente	Organismos	Observaciones
Catastro	DGC	Modeliza fincas, bienes inmuebles, etc. Códigos propios.
Cartociudad	IGN, INE, DGC, Correos	Modeliza vías, tramos de vías, portales, etc.
Censo de Edificios 2011	INE	Aproximaciones postales de los edificios incluidos en el Censo de 2011.
Direcciones AGE	INE, IGN,DGC, Correos	Modeliza direcciones. En fase diseño.
BDMAC	ICGC, Municipios, IDESCAT	Modeliza direcciones. Modelo aprobado por la Comisión Cartográfica de Catalunya.

En esta tabla se han marcado en gris aquellas fuentes que se han considerado poco apropiadas para su uso en el RET, ya sea porque no son exhaustivas, caso de las aproximaciones postales del Censo de edificios de 2011 en el que sólo se recogía información de edificios en los que habían residentes, ya sea porque están en una fase de diseño como el modelo de direcciones AGE [8] o porque presentan dificultades, como es el caso de Cartociudad [9], para satisfacer las finalidades del RET.

En efecto, Cartociudad modeliza portales, tramos de vía y vías y aunque incorpora una tabla de correspondencias entre los identificadores de vía propios de su base, las del INE y las del Catastro, esta es incompleta. Además, no modeliza las direcciones completas (verticales) y es natural que así sea ya que sus finalidades son ciertamente otras.

Las alternativas son por un lado la BDMAC, Base de datos municipal de direcciones de Catalunya [10] y, por otro, el Catastro. La BDMAC es un modelo aprobado por la Comisión Cartográfica de Catalunya y trata tanto direcciones horizontales (portales) como verticales (portales más planta y puerta).

No obstante la parte del modelo referente a direcciones verticales es optativa en cuanto a su implementación y, en este momento, es un modelo del que no existe una única implementación (los propietarios de los datos son los Municipios).

Por suerte, la evolución de los datos que dan soporte al geocodificador del ICGC, que en un principio se basaban en un grafo de calles sobre los que se procedía a la geocodificación por interpolación, han ido evolucionando hacia un modelo que contiene información relativa a portales. En la actualidad un 50% de la base de datos contiene información de portales y el modelo en sí se ha migrado a la BDMAC.

Si bien es cierto que en la implementación no se incluyen direcciones verticales, debido a que muchos ayuntamientos no disponen de esa información, el modelo BDMAC permite la incorporación de correspondencias entre el identificador de un objeto, un portal, con otros externos, como la referencia catastral de la finca correspondiente, permitiéndose así, la búsqueda de la referencia catastral de un bien inmueble correspondiente a una dirección completa (portal más planta y puerta).

Así pues la solución que proponemos, pasa por la obtención de la clave primaria BDMAC de un portal y la referencia catastral de la finca correspondiente. Mediante una segunda búsqueda sobre el archivo CAT del Catastro puede obtenerse la referencia catastral del bien inmueble, validando así la dirección postal completa (véase figura 11).

En los casos en que los datos a validar contengan los correspondientes códigos de vía INE, la búsqueda del portal BDMAC se realizará por el código, aumentando considerablemente la fiabilidad. En los casos en que esto no sea posible, como es el caso del DIRCE, se realizarán búsquedas a través de diccionarios de nombres de vía, tal y como el modelo BDMAC permite.

Aquellas direcciones que tengan un valor nulo de la clave primaria de la fuente BDMAC, es decir la clave foránea en la tabla de correspondencias, serán sospechosas de ser inválidas. Para estas direcciones se tomarán las medidas más adecuadas ya sea la edición de las direcciones o su imputación (con los criterios descritos en la primera parte de la comunicación).

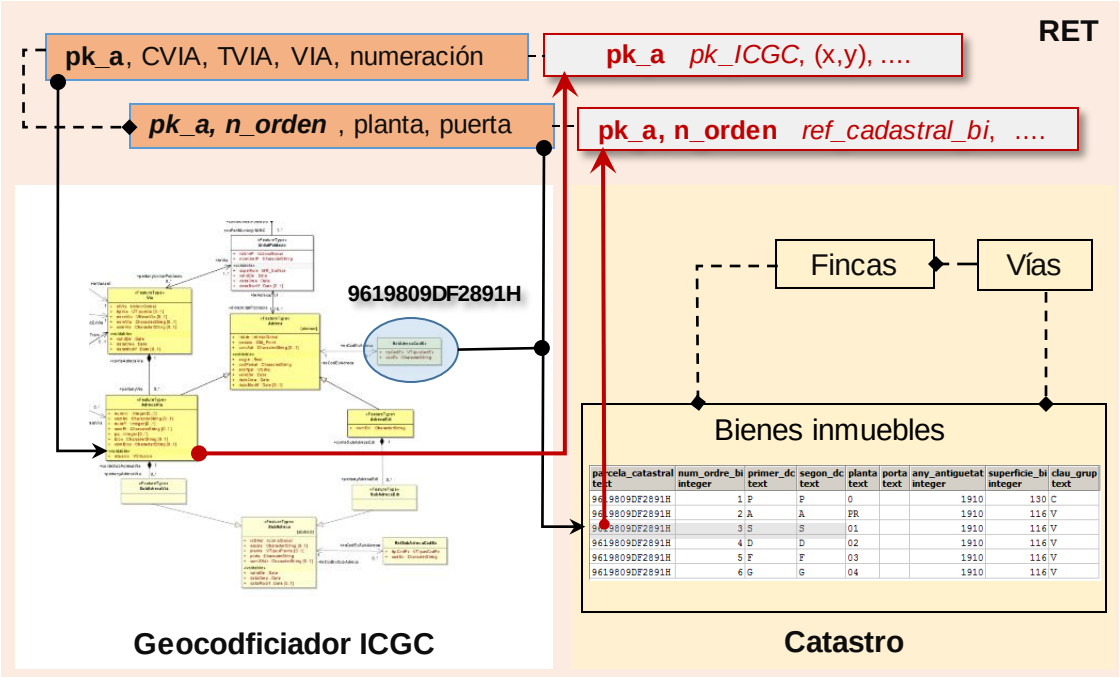


Figura 11: Esquema general del RET y procesos de validación de direcciones.

Es interesante destacar que mediante la tabla de correspondencia entre portales BDMAC y referencias catastrales, se está construyendo implícitamente una tabla de correspondencias vías INE vías DGC. Además, la búsqueda de los bienes inmuebles relativos a direcciones completas (portal más planta y puerta) permitiría la realimentación de la BDMAC, completando la información relativa a este tipo de direcciones.

CONCLUSIONES

La existencia del modelo de direcciones BDMAC y su correspondiente implementación permitirá la creación del Registro Estadístico de Territorio (RET), una infraestructura de datos que dará soporte a los Registros Estadísticos de Población (REP) y de Empresas y Establecimientos (REEE), mediante:

- La asignación de una clave primaria que se propague a lo largo del REP y REEE
- La validación de direcciones postales
- La obtención de las coordenadas de las direcciones postales
- La obtención de otras características derivadas de la dirección: superficie del inmueble, año de construcción, características de edificio, etc.

Para que todo esto sea una realidad, el ICGC y el IDESCAT colaboran intensamente tanto en la definición de una nueva interface SOAP de su servicio de geocodificación como en futuros procesos de realimentación derivada de los resultados obtenidos mediante el RET.

AGRADECIMIENTOS

Queremos agradecer especialmente el soporte y la colaboración del Institut Cartogràfic i Geològic de Catalunya.

REFERENCIAS

- Anders Wallgren, Britt Wallgren. Estadísticas basadas en registros. INEGI.
- Servei web de geocodificació de l'Institut Cartogràfic de Catalunya. http://www.gencat.cat/ptop/butlleti_innovacio/01/ICC_01.pdf.
- SIGPAC. Generalitat de Catalunya. *http://www.gencat.cat*
- NOAA. http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html
- E. Suñé. La georreferenciación de la población de Catalunya. VIII Jornadas SIG libre. Girona
- Levenshtein, Vladimir I. (February 1966). "Binary codes capable of correcting deletions, insertions, and reversals". *Soviet Physics Doklady* 10 (8)
- Jaro, M. A. (1989). "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida". *Journal of the American Statistical Association* 84
- Modelo de Direcciones de la Administración General del Estado v.2
- Proyecto Cartociudad
- Especificacions tècniques de la Base de dades municipal d'adreces de Catalunya v1.0