El registro estadístico de territorio: finalidades y calidad de sus fuentes

Eduard Suñé, Daniel Ibáñez, Anna Bernaus y Mireia Farré
Institut d'Estadística de Cataluña (Idescat)

El Institut d'Estadística de Catalunya (Idescat) está implementando un nuevo sistema de producción basado en registros administrativos. Este sistema está formado por una serie de subsistemas: el registro estadístico de población (REP), el registro estadístico de empresas y entidades (REE) y el registro estadístico de territorio (RET).

La idea es simple: si se gestiona adecuadamente la información que la administración conoce de la población y de las empresas, la obtención de resultados estadísticos resulta mucho más simple y económica que con el uso de métodos tradicionales que, generalmente, implican operaciones de campo.

Cuando hablamos de gestionar adecuadamente la información nos referimos a disponer de un sistema que sea capaz de mantener los estados por los que un individuo o empresa pueda pasar a lo largo de su existencia. En el caso de la población, como todo el mundo sabe, se nace, se obtiene una educación, se pasa por una vida laboral, se forma una nueva familia, etc. En todos estos cambios la administración intervine generando cierta información. Solo sería necesario gestionarla adecuadamente para obtener información estadística.

GEOLOCALIZACIÓN Y VALIDACIÓN DE LAS DIRECCIONES POSTALES

El registro estadístico de territorio (RET) es el subsistema responsable de la gestión de las direcciones postales que aparecen en el resto de subsistemas, con una doble finalidad: geolocalizar los microdatos y validar las direcciones postales completas.

Hay que tener en cuenta que una dirección postal es una información compleja resultante de la combinación de una serie de campos: unos dan información a nivel horizontal y otros la dan a nivel vertical. Por ejemplo: calle Unión, 25, 4.º 3.ª de Vilanova i la Geltrú. En este caso concreto, el RET facilitaría las coordenadas (X, Y) del portal de la calle Unión nº 25 (geolocalización de la dirección horizontal) y comprobaríamos que existe un inmueble en la planta 4 puerta 3 (validación de la dirección vertical).

Al alcanzar esta doble finalidad, la información del resto de subsistemas tendrá la máxima precisión espacial posible y será válida, con lo que desde el punto de vista territorial las estadísticas que puedan generarse serán de alta calidad.

La obtención de las coordenadas y la validación de la parte vertical de las direcciones postales se realizan utilizando sistemas de información externos al RET: por un lado, para la geocodificación se utiliza un servicio web del Instituto Cartográfico y Geológico de Cataluña (ICGC) y, por otro lado, para la validación de la dirección postal completa se utiliza la información alfanumérica de la Dirección General Catastro (DGC), considerándose una dirección válida cuando es posible encontrar en el Catastro un bien inmueble relativo a esa dirección.

En cuanto a la geocodificación, se obtienen las coordenadas del 96% de las direcciones postales cuando la información a tratar proviene del Registro estadístico de población (REP). Este porcentaje se explica por la alta calidad de las direcciones postales de entrada facilitadas por el Instituto Nacional de Estadística (INE), que incluye el código de vía. Para otras fuentes, los porcentajes pueden ser sensiblemente inferiores debido, entre otras razones, a la no disponibilidad de los códigos de vía o porque la dirección postal no está estructurada y se debe preprocesar antes de ejecutar el servicio web de geocodificación.

En cuanto a la validación de la dirección postal completa, que recordamos consiste en comprobar que existe un bien inmueble relativo a esa dirección, presenta la dificultad derivada de que los valores de campos como tipo de vía, código de vía, nombre de vía, bloque, escalera, planta y puerta presentan valores diferentes en los dos conjuntos de datos (la información del Registro estadístico de población (REP) procedente del INE y la información de la DGC). Por lo tanto, ha sido necesario construir un conjunto de tablas de equivalencias para cada uno de los campos que configuran la dirección, siendo el que se refiere a las vías el más problemático, dada su elevada cardinalidad. En efecto, en Catalunya existen aproximadamente unas 100.000 vías y solo un 50% de ellas tienen nombres exactamente iguales en los dos conjuntos de datos.

Para resolver este problema, evitando soluciones que impliquen el cálculo de métricas entre

literales, hemos utilizado métodos geométricos: una vez geocodificada la fuente de entrada (REP), obtenemos un conjunto de puntos que pueden ser agrupados según el código de vía. Así, para cada vía con más de dos portales geocodificados puede construirse un polígono mediante sus envolventes cóncavas obteniendo un polígono que es una buena aproximación a la geometría de la vía.

Por otro lado, dado que la información alfanumérica de las fincas de la DGC contiene sus centroides, puede realizarse también para esta fuente la operación descrita anteriormente, obteniendo así dos conjuntos de geometrías: las envolventes cóncavas de las vías INE y Catastro. La entidad física vía, tenga la descripción que tenga en las dos fuentes, tendrá asociados dos polígonos muy similares, es decir, dos polígonos con un elevado porcentaje de intersección y una orientación espacial (azimut) muy similar.

Con este criterio y bajo supervisión manual se logró construir una tabla de equivalencias entre vías INE y Catastro que cubría un porcentaje altísimo de portales, abriendo así el camino hacia la validación de las direcciones postales completas.

Así mismo, se construyeron las tablas de equivalencias del resto de campos de la dirección postal completa (horizontal y vertical) y en estos momentos hasta un 80% de las direcciones postales del RET son válidas, ya que han podido identificarse en la información de bienes inmuebles de la DGC.

MÉTODOS DE CONTROL DE CALIDAD DE LA DIRECCIÓN HORIZONTAL

Al generar las envolventes cóncavas de la fuente DGC observamos la presencia de puntos que sin duda eran erróneos debido a que su posición era anormalmente distante del resto, como se ilustra en la Figura 1.

Interpretamos que, en la mayor parte de los casos, esto ocurre cuando el código de vía y alguna de las posiciones son incoherentes, es decir, cuando el código de vía asignado a una finca es incorrecto.

Para asegurar la calidad de las posiciones obtenidas en el RET era necesario disponer de algún indicador que permitiera detectar polígonos con algún punto erróneo, mediante métodos de control supervisados.

Del conjunto de envolventes cóncavas obtenidas con la información de fincas de la DGC (unos 80.000 polígonos), se extrajo una muestra discrecional de 1.500 polígonos para obtener los grupos de aprendizaje y test. Cada uno de los polígonos de la muestra se inspeccionó y etiquetó como correcto o erróneo como paso previo a la exploración de indicadores.

En primer lugar, se realizó un estudio univariante utilizando variables derivadas de los polígonos (número de puntos, longitud de los lados, etc.) usando como método de comparación las áreas bajo la curva COR[1].

Finalmente, se evaluaron indicadores multivariantes y de ellos el de mejor comportamiento fue el siguiente indicador de calidad del polígono (ICP_p):

$$ICP_p = CV_p^2 \cdot max\left\{longitud_i - longitud_{i+1}\right\}_{i=1...v_p-1}$$

Donde "i" recorre los lados del polígono "p" según su longitud decreciente, "v_p" es su número de vértices y "CV_p" es el coeficiente de variación de las longitudes de los lados. **Todo ello sin tener en cuenta el lado de mayor longitud**.

[1] La curva COR (ROC en inglés, fórmula generalmente utilizada) es un gráfico que compara la sensibilidad (verdaderos positivos dividido entre los verdaderos positivos y los falsos negativos) y la especificidad (verdaderos negativos dividido entre los verdaderos negativos y los falsos positivos). En este estudio, se entiende como verdadero positivo un polígono etiquetado como erróneo y que realmente lo es.

Figura 1.

La razón de excluir el lado más largo es clara: un polígono que se corresponda con una vía que solo tiene fincas a un lado, un paseo marítimo, por ejemplo, tendrá necesariamente un lado mucho más largo que el resto (el que cierra el polígono). Un polígono erróneo tendrá uno o más puntos alejados del resto, pero tendrá al menos dos lados anormalmente largos.

Para cada polígono puede evaluarse el indicador de calidad del polígono (ICP) y obtener un subconjunto con una probabilidad muy alta de que sean incorrectos, tal y como puede observarse en la Figura 2.

Figura 2.

Los polígonos marcados en rojo son muy probablemente incorrectos, es decir, contienen algún punto incorrecto ya sea porque las posiciones sean erróneas o porque el código de vía asignado a la finca no es correcto.

Un método alternativo que puede utilizarse en relación al control de calidad de datos geocodificados consiste en detectar incoherencias entre el número del portal (número de policía) de una dirección postal y la posición de los portales. En efecto, en función del sistema de numeración de una calle, pueden construirse polígonos uniendo ordenadamente los puntos geocodificados. En principio, un polígono así construido debería ser geométricamente válido, es decir, que no se autointersecte. Debemos advertir aquí que este tipo de controles puede dar lugar a falsos positivos dependiendo de la distribución de portales en los dos lados de la calle y que hemos localizado casos de vías, como la calle de Can Ros en Barcelona, en los que la numeración es inconsistente pero correcta (Figura 3), crece en el lado de impares mientras que decrece en el lado de pares.

Por otra parte, en relación a la validación de las **direcciones verticales**, los resultados dependen del contenido de las tablas de equivalencias: tipos de vía, vías, bloque, escalera, planta y puerta. Para contrastar la calidad se construyeron tablas de equivalencias alternativas usando datos del modelo de direcciones de la Administración General de Estado (MDAGE), observando diferencias poco significativas entre las tablas de equivalencias construidas en Idescat y las que se derivan de los datos de la MDAGE.

Figura 3.

La diversidad de las descripciones de las direcciones postales en las distintas fuentes utilizadas en el Registro Estadístico de Territorio dificulta la correcta identificación espacial de las unidades estadísticas

CONCLUSIONES

El Registro Estadístico de Territorio incorpora diferentes fuentes con el fin de asignar y validar un identificador territorial para cada una de las unidades estadísticas, como población, hogares, empresas, entidades, etc.

La diversidad de las descripciones de las direcciones postales en las distintas fuentes utilizadas dificulta la correcta identificación espacial de las unidades estadísticas. Para ello, el uso de métodos espaciales y geométricos ha demostrado ser una herramienta muy poderosa para establecer tablas de correspondencias entre las diferentes fuentes, así como para validar, detectar errores y establecer indicadores de calidad para las direcciones georreferenciadas.