LA GEOREFERENCIACIÓN DE LA POBLACIÓN DE CATALUNYA.

Eduard Suñé Luis

Área de Producció d'Estadístiques Demogràfiques. Subdirecció General de Producció i Coordinació Estadística. Institut d'Estadística de Catalunya. Via Laietana 58, 08003 Barcelona. esl@idescat.cat.

RESUMEN

El Institut d'Estadística de Catalunya (Idescat) está trabajando para elaborar un registro estadístico territorial a partir de fuentes administrativas y estadísticas, de acuerdo con el código de buenas prácticas de las estadísticas europeas. En este contexto, la asignación de coordenadas a los microdatos de las fuentes estadísticas y administrativas es de vital importancia para la obtención de datos georeferenciados en los ámbitos demográfico, económico y social. En esta comunicación se describen diferentes alternativas y los resultados obtenidos al utilizar el servicio de geocodificación del Institut Cartogràfic de Catalunya (ICC) para georeferenciar la población de Catalunya, como un primer paso para la georeferenciación sistemática de microdatos. Por último, se describen los métodos utilizados en la imputación de la posición y algunos casos de uso de la población georeferenciada.

Palabras clave: *Georeferenciación, población, geocodificiación*

ABSTRACT

The Statistical Institute of Catalonia (Idescat) is working on producing a territorial statistical record on the basis of administrative and statistical sources, in accordance with the European Statistics Code of Practice. In this context, the assignation of coordinates to the microdata from the statistical and administrative sources is of vital importance for obtaining georeferenced data in the demographic, economic and social areas. This communication describes the different alternatives and the results obtained by using the geocodification service of the Cartographic Institute of Catalonia (ICC) to georeference the population of Catalonia as the first step towards the systematic georeferencing of the microdata. Finally, it describes the methods used for imputing positions and some cases where the georeferenced population has been used.

Key words: *Georeferencing, population, geocodification.*

INTRODUCCIÓN

El Institut d'Estadística de Catalunya (Idescat) está desarrollando un sistema de información estadística basado en registros administrativos (1). Todos estos registros tienen en común la existencia de un conjunto de variables que expresan la localización espacial del microdato: la dirección postal (lugar de residencia, sede de una empresa, etc.).

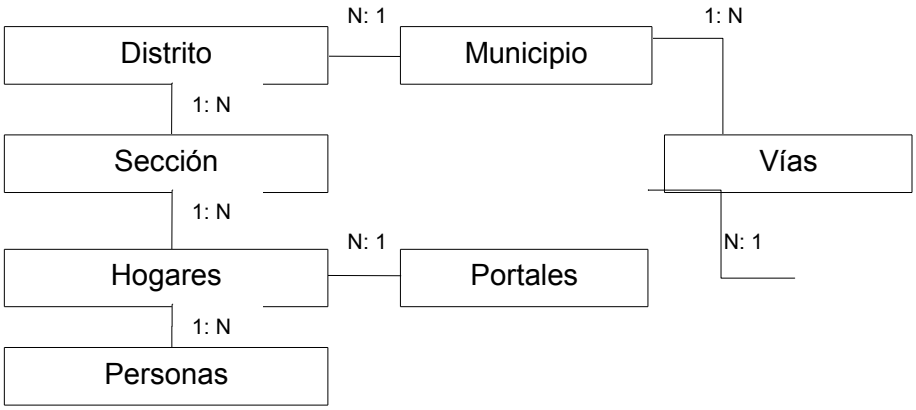
Históricamente, el tratamiento dado a las direcciones postales conducía a la asignación de zonas administrativas, como es el caso de la población y las secciones censales. El efecto final de esta asignación es que los resultados estadísticos sólo pueden obtenerse respecto de estas zonas o sus agregados, produciéndose una pérdida de información indeseable. La solución a este problema pasa necesariamente por la asignación de unas coordenadas a las direcciones postales existentes en los microdatos. Así, la posterior agregación espacial no está a priori limitada.

En este artículo presentamos los resultados obtenidos en el proceso de georeferenciación del Registro de población utilizando el servicio de geocodificación del Institut Cartogràfic de Catalunya (ICC), como un primer paso hacia la georeferenciación sistemática de los registros administrativos.

GEOCODIFICACION DEL REGISTRO DE POBLACIÓN.

El Registro de población es un fichero estadístico derivado del padrón de habitantes gestionado por los ayuntamientos. Esta información la centraliza el Instituto Nacional de Estadística (INE), realiza ciertos tratamientos de depuración y, periódicamente, cede la información relativa a la población de Catalunya al Idescat.

Idescat, por su parte, realiza unos tratamientos para la estandarización de las direcciones postales, agregación de las personas con una misma dirección postal (hogares) y creación de la tabla portales que modela una dirección postal a nivel horizontal. El diagrama entidad-relación de esta información vendría a ser:



Así pues, la tabla portales representa el conjunto de direcciones postales a nivel horizontal en donde reside alguna persona. Esta tabla contiene un conjunto de campos que expresan una dirección postal: provincia, municipio, distrito, sección, código de vía, tipo de numeración, número inferior, número superior, punto kilométrico, etc. La tabla vías contiene los nombres normalizados de las vías, sus códigos y la clave foránea de la tabla municipios. El *target* de nuestro tratamiento será pues la tabla de portales.

El Institut Cartogràfic de Catalunya 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.

Hay que tener en cuenta que el servicio trabaja contra un grafo de calles y realiza interpolaciones para la obtención de las coordenadas que, por lo tanto, serán aproximadas. También es necesario mencionar que el resultado de la geocodificación depende, obviamente, de los parámetros de entrada y que en casos de direcciones mal formadas o incorrectas el servicio no devolverá, naturalmente, unas coordenadas.

Desde Idescat se realizaron los desarrollos de software, utilizando herramientas de desarrollo libres, para integrar este servicio y para realizar altas de las coordenadas obtenidas, utilizando para ello la base de datos espacial postGIS.

El nivel de geocodficación obtenido fue, para la información a 1 de enero de 2013, del 92,36% sobre el conjunto de portales (990.000), un nivel ciertamente alto. No obstante, la distribución en el territorio de los portales no geocodificados no es uniforme. Los percentiles de casos geocodificados, agrupados por municipios puede observarse en la tabla 1.

Tabla 1: Distribución de casos geocodificados agrupados por municipios

Percentil	**% casos geocodificados**
100	100
75	94,94
Mediana	89,01
25	76,99

Los casos de municipios con bajos porcentajes se corresponden a aquellos de tipo rural, dándose la circunstancia de que en 27 de ellos el proceso de geocodificación no devolvió ningún punto. Es el caso del municipio de Bellprat, para el que la mera observación de las direcciones postales explica este fenómeno. En efecto, todas las direcciones que aparecen en el Registro de población hacen referencia a nombres de casas y no, como sería de esperar, a una calle y una numeración. Los casos no geocodificados, unos 75.000, fueron imputados con diferentes técnicas que describiremos más adelante.

Por otro lado, como trabajo preliminar del Censo de población y viviendas del año 2011 el INE realizó el Censo de edificios. La principal novedad de este censo es que se recogió por primera vez las coordenadas de las aproximaciones postales de los edificios en aquellas secciones censales en donde se efectuó recorrido.

En el resto de secciones se utilizó la información del Catastro para obtener estas coordenadas. Idescat ha utilizado la información provisional de las aproximaciones postales de esta fuente, para un cierto número de municipios, con el fin de comparar los resultados obtenidos con el geocodificador del ICC y validar sus resultados.

Para el conjunto de portales comunes en las dos fuentes se creó una capa con un segmento definido por las dos posiciones, la obtenida mediante el geocodificador del ICC y las posiciones del Censo de edificios, tal como puede observarse en la figura 1.

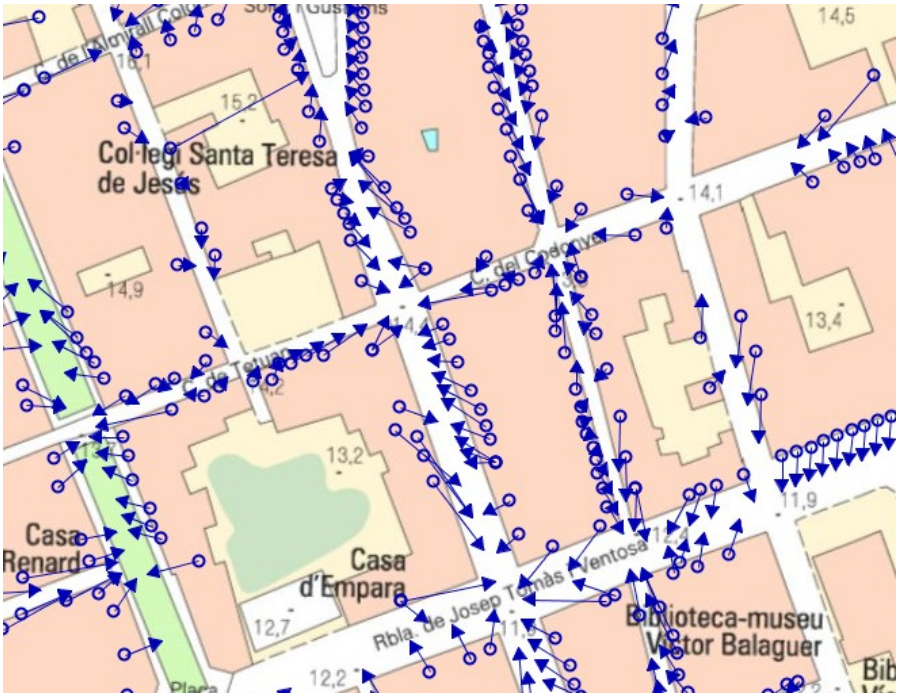


Figura 1: *Comparación de los resultados del geocodificador del ICC y del Censo de edificios 2011 en una zona del centro de Vilanova i la Geltrú. Los segmentos unen los puntos correspondientes a las dos fuentes.*

Del análisis de la distribución de distancias se deduce que los resultados del geocodificador del ICC son muy parecidos a los del Censo de edificios y de hecho el valor medio de las distancias parece consistente con el hecho de que el geocodificador del ICC utiliza un grafo de calles e interpolación para la obtención de las posiciones.

En la tabla 2 puede observarse la distribución de distancias ICC-Censo de edificios 2011 para el municipio de Vilanova i la Geltrú. Que un 75% de la distribución tenga un valor inferior a los 16 m indica la alta calidad de los resultados obtenidos con el geocodificador del ICC. Existen, ciertamente valores atípicos que, en general, se corresponden a errores en las coordenadas del Censo de edificios del 2011, aunque son muy poco frecuentes.

Tabla 2: Distribución de distancias entre los puntos obtenidos por el geocodificador ICC y el Censo de edificios 2011. Vilanova i la Geltrú.

Percentil	**Distancia (m.)**
Mínimo	0,078
25	6,482
50	10,447
75	**16,063**
Máximo	1992,496
Media	15,036

Por último resaltar que esta validación no se pudo realizar para el conjunto de portales de Catalunya debido a que la información completa del Censo de edificios 2011 no estaba disponible en ese momento.

IMPUTACION DE COORDENADAS

Para la población del Registro a 1 de enero de 2013 el porcentaje de portales efectivamente geocodificados asciende a un 92,36% (un 94,85% a nivel de personas). Queda pues un relativo pequeño número de portales cuya posición no ha podido obtenerse con el servicio de geocodificación del ICC.

A priori, para estos casos, conocemos a qué sección censal pertenecen ya que es una información que originalmente forma parte del Registro de población. Podríamos asignar una posición, típicamente el centroide, de esa sección censal, pero teniendo en cuenta que las secciones censales son divisiones administrativas que se definen precisamente por la población que incluyen, en zonas de alta densidad de población el error cometido seria pequeño pero, por el contrario, muy alto en zonas de baja densidad de población. Precisamente en las zonas de baja densidad los porcentajes de portales geocodificados son más bajos.

Así pues es necesario proceder a la imputación de las posiciones utilizando métodos más refinados.

Como quiera que la información original está georeferenciada a nivel de sección censal es necesario realizar la imputación de coordenadas sección a sección de tal forma que el valor imputado sea un punto (x,y) de su sección. Por otro lado, la posición, como tal, es una variable continua cuyos valores posibles dependerán de la forma de la sección censal.

Para simplificar el problema podemos definir un grid de lado n metros asociado al bounding box y calcular posteriormente la intersección con el contorno de la sección censal. En la figura 2 puede observarse el grid y la intersección para una sección censal (02003) del municipio de Roses.

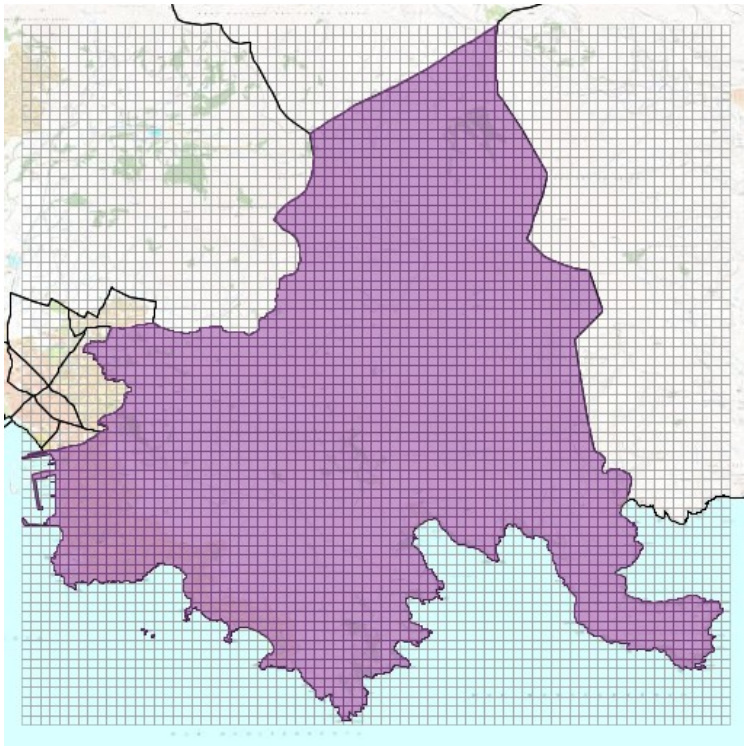


Figura 2: El grid y su intersección *con la sección censal 02003 de Roses.*

Así pues, podemos simplificar el problema de la imputación considerando que las posibles posiciones a imputar son los centroides de los elementos resultantes de la intersección del grid y la sección censal. De esta forma, la imputación lo es de una variable discreta: sólo es necesario asignar un elemento del grid para obtener una posición, su centroide o más exactamente un punto dentro del elemento.

El problema, una vez simplificado, se reduce a establecer cómo se realiza esa asignación, que en todo caso debería realizarse con criterios probabilísticos. En efecto, suponiendo que los valores *missing* se repartieran uniformemente en el territorio, deberíamos asignar a cada elemento del grid una probabilidad de ser escogido igual la superficie relativa de ese elemento:

$$P(k) \sim s(k) \quad [1]$$

siendo s la superficie relativa del elemento k del grid

El segmento [0,1] (véase figura 3), cuya longitud representa un suceso cierto, quedaría pues dividido en tantas partes como elementos del grid con longitudes iguales a su superficie relativa.

Desde el punto de vista algorítmico, la asignación de unas coordenadas pasaría por la obtención de un numero aleatorio dentro del intervalo [0,1], la búsqueda de a qué elemento del grid le corresponde ese valor en la tabla de superficies relativas acumuladas y la asignación final de un punto del elemento del seleccionado.

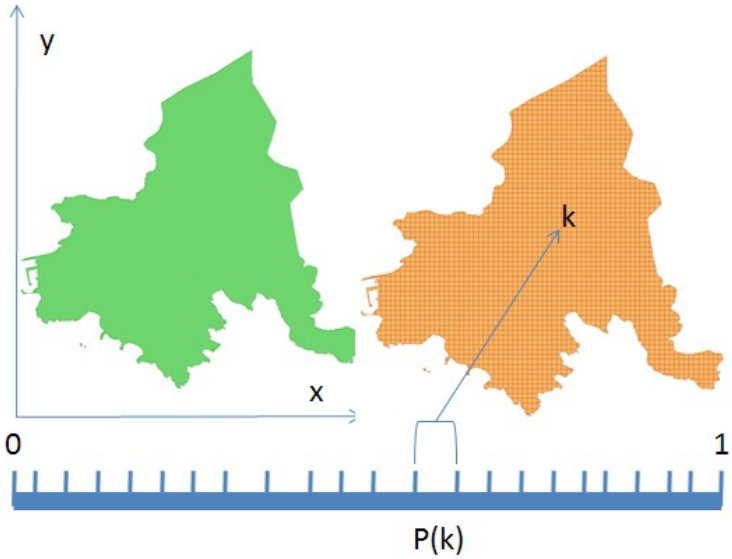


Figura 3: *Mediante la construcción de un grid la asignación aleatoria de coordenadas puede simplificarse utilizando el centroide de los elementos del grid. Queda estimar la probabilidad asociada a cada elemento que, a falta de más información, seria función de su superficie relativa.*

Otra hipótesis plausible es que el no consta se reparte en las zonas en que sí existen puntos geocodificados y de forma proporcional al número de portales geocodificados. Esta hipótesis suele asumirse frecuentemente en ciertos métodos de imputación, como por ejemplo en los métodos *hot deck,* en los que los valores a imputar se toman aleatoriamente del subconjunto de datos no missing. En esta

hipótesis, las probabilidades estimadas para los elementos del grid serían (véase figura 4):

$$P(k) \sim f(k) \quad [2]$$

siendo f la frecuencia relativa de portales geocodificados en el elemento k del grid

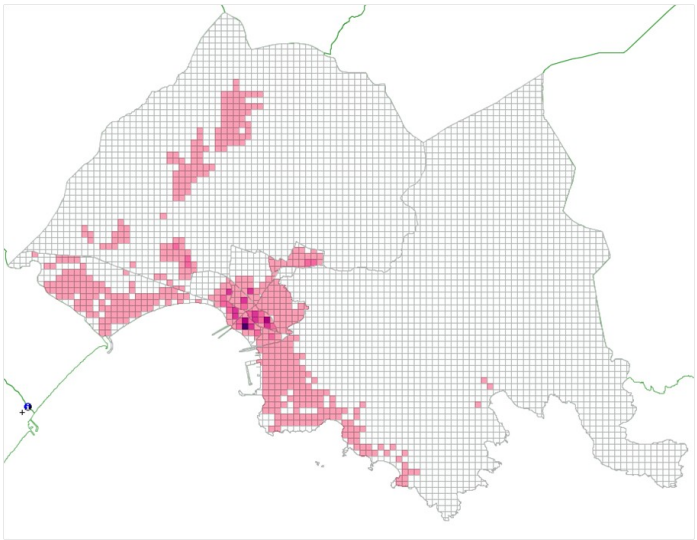


Figura 4: *Podría estimarse la probabilidad mediante la frecuencia relativa de los portales geocodificados. En esta imagen puede observarse la distribución de casos geocodificados para el municipio de Roses.*

No obstante hay que tener en cuenta que pueden haber secciones con un nivel de portales geocodificados relativamente bajo. En estos casos la hipótesis anterior pierde fuerza, llegando incluso al caso extremo en que no exista ningún portal geocodificado.

Por lo tanto es necesario establecer un límite a partir de cual asumiremos la hipótesis anteriormente mencionada y aplicaremos la ecuación 2 para estimar las probabilidades. Este límite lo hemos fijado en un 70%, es decir, en secciones con un nivel de geocodificados igual o superior al 70% estimamos la probabilidad mediante las frecuencias relativas de geocodificados en los elementos del grid. Este método de imputación lo denominaremos método A.

En el caso que el límite del 70% no se alcance necesitamos otra solución al problema. Esta solución no pasa directamente por estimar la probabilidad mediante la superficie relativa de cada elemento del grid, ya que sabemos, a priori, que existen zonas en las que esta probabilidad es cero: zonas boscosas, zonas de cultivo, etc.

Es necesario delimitar el territorio en donde es posible que exista un portal y donde no. Por suerte el SIGPAC [3] delimita las zonas urbanas del resto, de tal forma que si en vez de trabajar con el grid trabajamos con la intersección del grid y las zonas urbanas del SIGPAC, eliminamos todos aquellos elementos del grid con probabilidad cero. En la figura 5 puede observarse que la distribución de portales geocodificados están situados precisamente en las zonas urbanas delimitadas en el SIGPAC.

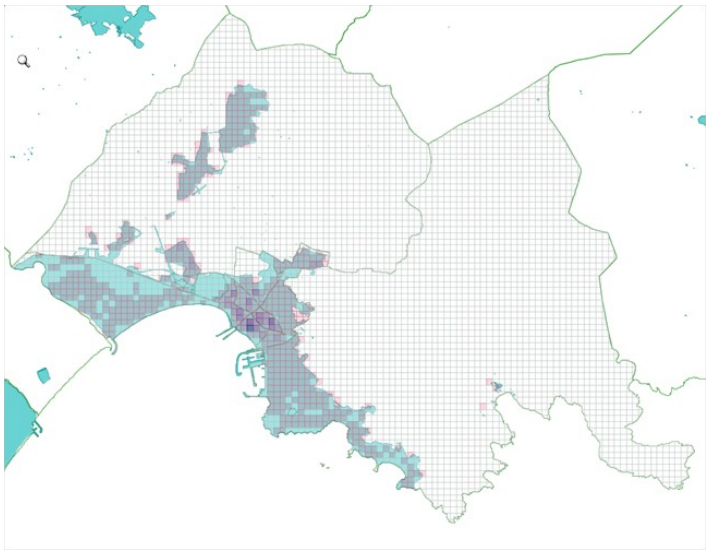


Figura 5: *Los elementos del grid con frecuencias no nulas coinciden con la información obtenida del SIGPAC para zonas urbanas*

Podemos fijar nuestra atención en un caso extremo: un municipio en donde no se ha geocodificado ningún portal, como por ejemplo Bellprat. Para este municipio, que es de sección única, operaremos como hasta ahora: calculamos un grid de un cierto tamaño asociado a su bounding box, obtenemos la intersección con la sección censal y, como novedad, calculamos la intersección con la información del SIGPAC correspondiente a las zonas urbanas.

Ahora podríamos estimar la probabilidad utilizando el criterio de la superficie relativa aplicada al resultado de esas intersecciones. No obstante el criterio de la superficie no parece suficientemente bueno ya que tiene implícita la suposición de que la densidad de población es uniforme en el territorio.

Con la finalidad de mejorar los criterios de estimación de probabilidades hemos utilizado imágenes satélites nocturnas obtenidas por la NOAA (4). Estas imágenes contienen los valores de luminosidad con un tamaño de pixel que se corresponde a un área de 900 m x 900 m, aproximadamente. La capa raster se ha convertido, utilizando Sextante, a una capa de puntos con los valores de la luminosidad (figura 6).

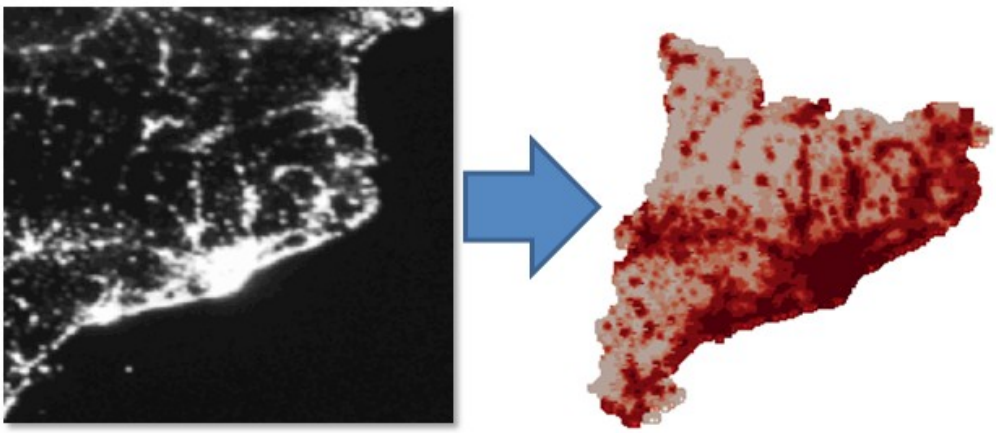


Figura 6: Mediante fotografías satélite nocturnas puede obtenerse una capa de puntos con la luminosidad asociada

Con todo esto podemos estimar la probabilidad como:

$$P(k) \sim L(k)*S(k)/\sum L(k)*S(k) \quad [4]$$

donde S(k) es la superficie y L(k) la luminosidad del elemento k del grid, calculada mediante la media móvil de n valores de luminosidad situados a una cierta distancia.

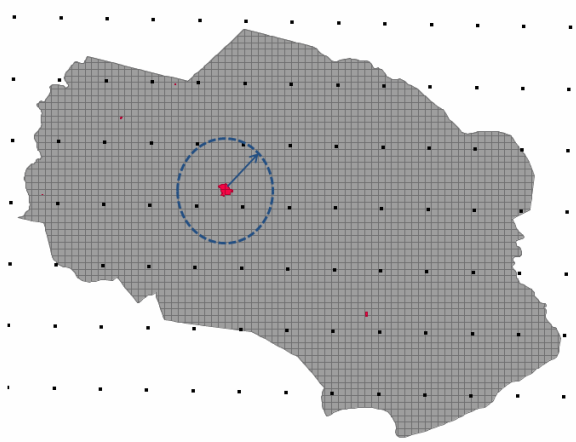


Figura 7: *La probabilidad puede estimarse mediante el producto de la superficie relativa de la intersección grid-SIGPAC (rojo) y la luminosidad en el centroide del elemento calculada mediante la media móvil de los puntos de la capa de luminosidad a una cierta distancia.*

Este método de imputación, que denominaremos B, se ha aplicado a todas aquellas secciones censales con un porcentaje de portales geocodificados inferior al 70%.

Por último es necesario comentar que la disponibilidad de la geometría asociada a las secciones censales está condicionada por sus frecuentes modificaciones. Así, en el momento en que se realizaron los tratamientos de la población a 1 de enero de 2012, Idescat disponía de la geometría de las secciones censales del año 2010. Para el tratamiento de la población a 1 de enero de 2013 disponía de la geometría de las secciones censales del año 2011 y, por lo tanto, la imputación que hemos descrito (tanto A como B) se realizó para aquellas secciones que permanecieron inalteradas. Para el resto se realizó la imputación a algún punto de la capa SIGPAC o al centroide del municipio. La siguiente tabla indica el número de portales y personas según los métodos utilizados para la obtención de coordenadas para el caso de la población a 1 de enero de 2013.

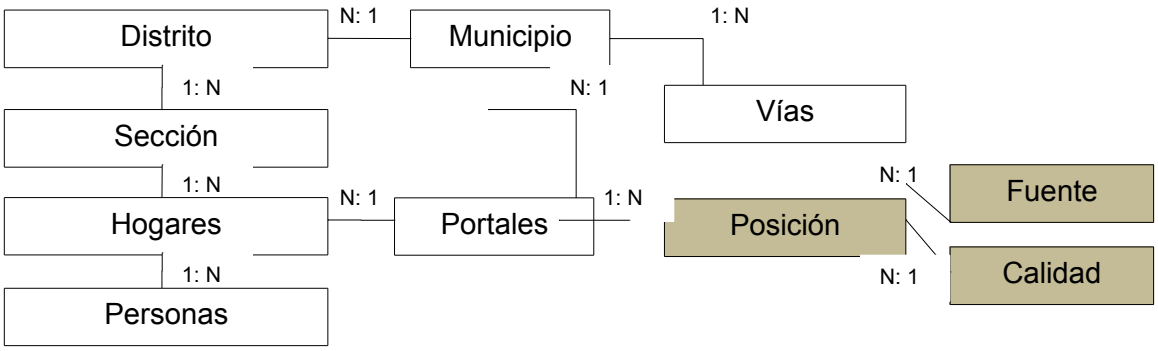
Tabla 3: Resumen del proceso de geocodificación e imputación para el Registro de población a 1 de enero de 2013.

Calidad/Tratamiento	**Portales**	**%**	**Personas**	**%**
1/ Geocodificados ICC	914.316	92,36	7.196.043	94,86
2/ Imputados método A	59.538	6,01	327.156	4,31
3/ Imputados método B	15.105	1,53	48.777	0,64

4/ Imputados zona urbana SIGPAC	731	0,07	13.214	0,17
5/ Imputados al centroide del municipio	310	0,03	1.701	0,02
Total	**990.000**	**100**	**7.586.891**	**100**

RESULTADOS. CASOS DE USO.

El esquema entidad relación del registro de población es después de los procesos de geocodificación e imputación:



Como puede observarse, a cada portal del registro de población se ha asignado unas coordenadas, una fuente y un indicador de calidad que hemos definido en función del tratamiento y que se corresponde con el orden decreciente de filas en la tabla 3. De esta forma, los usuarios finales de esta información pueden escoger el nivel de calidad de los datos e incluir en sus cálculos aquellos que consideren más fiables. Asignar unas coordenadas a un portal y por consiguiente asignarlas al lugar de residencia permite evaluar la población dentro de cualquier área de interés. Además dado que el conjunto de variables recogidas para la población en el registro incluyen entre otras el sexo, la edad y la nacionalidad, el cálculo de la población puede llegar a ser extremadamente útil para la toma de decisiones por parte de la administración.

Un caso paradigmático lo encontramos en los trabajos de planificación que debe realizar Protección Civil: planes por riesgo de inundación (INUNCAT), transporte de materias peligrosas (TRANSCAT) o planes de emergencia exterior del sector químico en Catalunya (PLASEQCAT). En todos estos casos, con una metodología concreta, se deben evaluar poblaciones en zonas de riesgo.

En la figura 8, a modo de ejemplo, puede observarse la población georeferenciada (capa en negro) y la zona de inundación correspondiente a desembocadura del rio Llobregat (capa azul). La intersección de la dos capas permite evaluar con precisión la población por edades dentro de la zona de riesgo, tal como es necesario según la metodología que Protección Civil aplica en este plan.

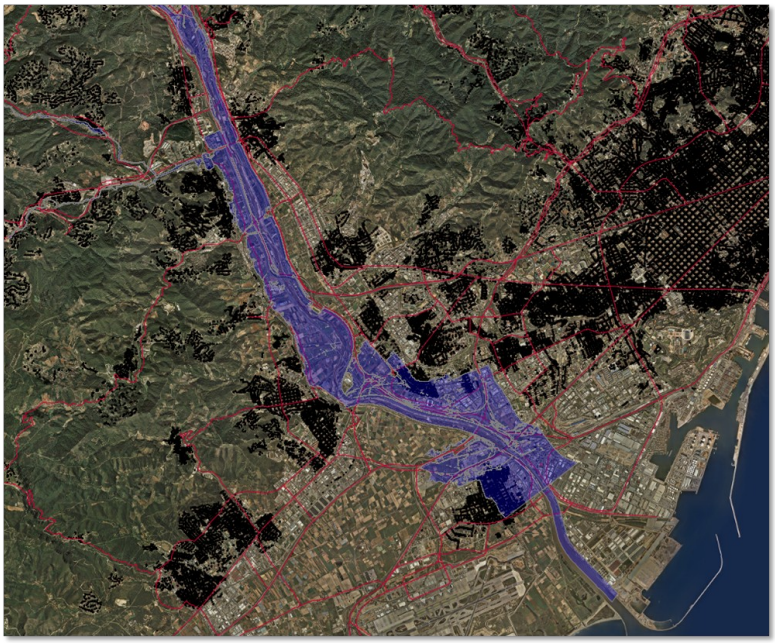


Figura 8: *Detalle de la zona del desembocadura del rio Llobregat. La capa de puntos se corresponde a la población georeferenciada a 1 de enero de 2013.*

Otro caso típico consiste en evaluar las zonas de influencia de ciertas infraestructuras y el cálculo de a cuanta población, de un determinado tipo, incluyen. A modo de ejemplo, en la figura 9, puede observarse las posiciones de ciertos colegios de Vilanova y la Geltrú (puntos azules) y las zonas de influencia (polígonos de Voronoi). De nuevo el cálculo preciso de a cuanta población de un determinado tipo afectan estas infraestructuras es posible.

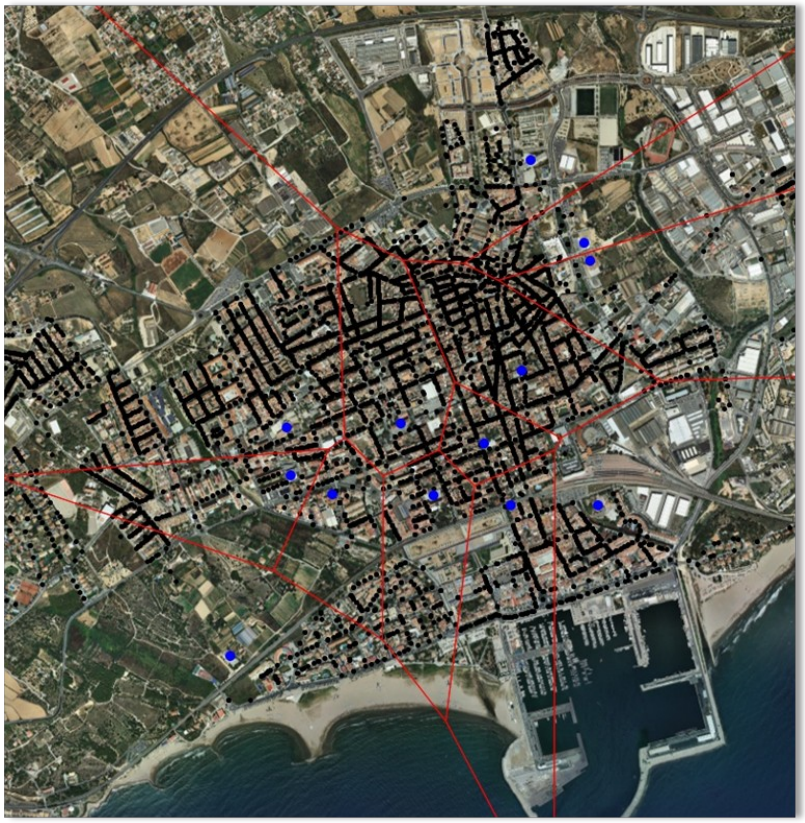


Figura 9: *La capa de puntos azul nos presenta las posiciones de ciertos colegios de Vilanova i la Geltrú. La capa de puntos negros se corresponde a la población georeferenciada. El cálculo de los polígonos de Voronoi permitiría evaluar la zonas de influencia y la población afectada*

CONCLUSIONES

La georreferenciación de los microdatos en los diferentes registros estadísticos de población permite su evaluación precisa en zonas de interés y, por lo tanto, la toma de decisiones por parte de la administración. El servicio de geocodificación del ICC, junto con técnicas de imputación permiten en este momento la obtención de esta valiosa información. En un futuro no muy lejano, iniciativas como la base de datos de direcciones de Catalunya (BDMAC) (5) o el modelo de direcciones de la administración del estado permitirán la validación y georeferenciación sistemática de los microdatos del sistema estadístico basado en registros administrativos.

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
- Especificacions tècniques de la Base de dades municipal d'adreces de Catalunya v1.0.