Geocodificación múltiple de direcciones postales basada en proveedores externos

Raquel Bedia Expósito (Instituto Cántabro de Estadística)
Miguel Expósito Martín (Dirección General de Organización y Tecnología del Gobierno de Cantabria)

Resumen

Desde el Instituto Cántabro de Estadística (ICANE) se viene trabajando en el establecimiento de un marco básico de geocodificación de direcciones postales, tanto de personas como de establecimientos, para obtener información geográfica de calidad y satisfacer las necesidades del propio Instituto y otras unidades del Gobierno de Cantabria. Dicha información supone una fuente de datos fundamental para completar las publicaciones actuales (como es el caso del Directorio de Empresas y Establecimientos de Cantabria) así como para el desarrollo de nuevos productos y la realización de estudios relacionados con cuestiones tan diversas como la localización y flujos de empleo, estudios de salud pública y mortalidad, etc.

Dada la limitación de recursos existente y la ausencia de un callejero digital unificado regional, la línea de trabajo se ha basado en el uso de proveedores externos como *Google* o *Cartociudad* (entre otros). Se ha desarrollado un paquete Python que permite la lectura masiva de ficheros de entrada con direcciones que pueden estar previamente normalizadas con otros productos ya existentes en el ICANE (y, por tanto, con un tratamiento de calidad mínimo). A partir de dicha lectura, se lanzan en paralelo tantos procesos de geocodificación como proveedores existan, volcándose sus resultados a varios ficheros. Dichos resultados son fusionados y analizados posteriormente para la elección de la mejor coordenada de entre todos los proveedores utilizando algoritmos de agrupación o *clustering* (DBSCAN) y siguiendo un criterio de cercanía al centroide del clúster formado por los proveedores o de selección basado en la mejor correlación de las coordenadas devueltas por un proveedor con las del resto (en caso de no existir suficientes proveedores en el clúster). Para ello se tiene en cuenta si los proveedores son seleccionables como mejor coordenada en términos de licencia. Finalmente, se generan estadísticas básicas sobre los datos geocodificados para facilitar otro análisis posterior.

Los primeros resultados con muestras de colecciones de mil direcciones postales procedentes de diversas fuentes (Directorio de Empresas y Establecimientos de Cantabria y Tarjeta Sanitaria) indican que se puede llegar a geocodificar hasta el 90% de la muestra utilizada con una coincidencia mínima de dos proveedores en radios de 200 o 500 m, quedando las restantes en situación de revisión de cara a minimizar la aparición de falsos positivos. El objetivo inmediato del proyecto es poder publicar resultados con una calidad razonable para finales del ejercicio 2018, así como dar servicio a otras unidades que lo requieran.

Palabras clave:

geocodificación, direcciones, *python*, *clustering*

1. Introducción

La geocodificación directa o georreferenciación es el proceso consistente en encontrar identificadores geográficos o coordenadas a partir de datos textuales como direcciones postales. Generalmente, los sistemas de geocodificación disponen de un procesador o *parser* de direcciones postales que estructura y separa la dirección en sus distintos atributos (portal, calle, código postal, localidad, etc.) y de una base de datos geográfica con direcciones postales, asignando la coordenada a través de la mejor correspondencia.

Disponer de datos geocodificados no sólo puede ser útil para su visualización en mapas, sino que generalmente es el primer paso para realizar estudios más avanzados en ámbitos de la salud (mortalidad y causas de muerte) o la economía (flujos de movilidad laboral, localización de empleo).

Muchas Oficinas Estadísticas y Gobiernos Autonómicos disponen de infraestructura y recursos propios para obtener bases de direcciones geocodificadas, tales como encuestas in situ o sistemas de gestión de Callejeros. Otras, en cambio, han de nutrirse de proveedores externos como *Google* o proyectos ambiciosos como *Cartociudad*.

La utilidad última de disponer de datos geocodificados depende en gran medida tanto de la completitud como de la precisión de los procesos de geocodificación. La motivación para el presente trabajo viene dada por la necesidad de conseguir un sistema software capaz de compensar, en la medida de lo posible, las carencias de múltiples proveedores de geocodificación, ofreciendo resultados correctos con una garantía de calidad adecuada y minimizando los falsos positivos y negativos. El propósito general del desarrollo es poner el producto resultante a disposición del público liberándolo como un paquete de código abierto, y tratar de contribuir en la medida de lo posible a que las Oficinas Estadísticas con recursos limitados puedan comenzar a generar conjuntos de datos georreferenciados con un esfuerzo y coste razonables. Por otra parte, de forma más específica, el ICANE tiene prevista la utilización de los resultados proporcionados por este producto para la representación de establecimientos en mapas, localización de la población activa en Cantabria según su situación laboral, estudios de mortalidad según causa de muerte (proyecto MEDEA[1]) y localización de la población cántabra según la dirección de su Tarjeta Sanitaria.

En las secciones siguientes de la presente ponencia se discuten las características y resultados proporcionados por el geocodificador múltiple desarrollado.

[1] http://www.proyectomedea.org/

2. Trabajos previos

Otras Oficinas Estadísticas y organizaciones están inmersas en proyectos de naturaleza muy similar al presentado. Como casos destacables estudiados en el presente trabajo, cabe citar:

- Proyecto MEDEA: Medea es una acción estratégica del Centro de Investigación Biomédica en Red (CIBER) de Epidemiología y Salud Pública de España, que tiene por objetivo analizar las desigualdades socioeconómicas y medioambientales en áreas pequeñas de ciudades de España y de Europa. Varias Comunidades Autónomas participan en este proyecto y han desarrollado un protocolo de geocodificación propio para aquellas que no disponen de estadística georreferenciada. Dicho protocolo consiste, básicamente, en una geocodificación secuencial utilizando *Cartociudad* (en sus dos versiones) y *Google* con un pre-procesado de direcciones de naturaleza heurística.
- Instituto Canario de Estadística (ISTAC): el ISTAC está trabajando en un sistema general de geocodificación y un modelo general de geodatos; entre sus análisis se encuentra un estudio y descripción bastante exhaustivos del funcionamiento de APIs de geocodificación de diversos proveedores, sus términos de licencia, etc.

3. Metodología de desarrollo

Para enfocar el desarrollo del paquete de software se ha optado por utilizar una metodología de desarrollo incremental e iterativa, considerando que el proyecto en cuestión presentaba determinado nivel de incertidumbre debido a su componente en investigación del estado tecnológico actual.

Esta metodología ha permitido trabajar sobre subconjuntos de funcionalidad y emplear un enfoque de diseño adaptativo, mejorando tanto el producto desarrollado como el proceso de desarrollo.

La referencia metodológica más clara en este ámbito es *Extreme Programming*, (Kent Beck, 2004) *es "un estilo de desarrollo de software centrado en la aplicación excelente de técnicas de programación, comunicación clara y trabajo en equipo que permite conseguir objetivos antes impensables".* Se trata de una metodología basada en valores como la comunicación, realimentación, simplicidad, valentía y respeto, soportada sobre un cuerpo de prácticas útiles y con un conjunto de principios complementarios, además de contar con una comunidad de usuarios que comparte todo lo anterior.

Su aplicación al desarrollo de este proyecto no ha sido estricta; por ejemplo, no se han definido ciclos estrictos por la propia naturaleza inestable de la dedicación al desarrollo del mismo, pero sí que ha realizado un diseño evolutivo soportado sobre un número suficiente de pruebas unitarias, así como una planificación incremental y adaptativa a las problemáticas que iban surgiendo.

A lo largo del desarrollo, esta forma de trabajo ha permitido realizar una gestión del cambio efectiva, re-orientando los desarrollos a tiempo sin impactar en el diseño ni en la implementación existente. Finalmente, tanto la organización como la planificación de las tareas han sido lo suficientemente flexibles para conseguir un ritmo adecuado de desarrollo, reduciendo los puntos de bloqueo.

La estructura de la ponencia refleja las distintas fases seguidas en la aplicación iterativa e incremental de la metodología.

4. Análisis

De manera previa al lanzamiento del proceso de geocodificación, se llevó a cabo un estudio de las APIs[2] públicas de geocodificación ofrecidas por distintos proveedores. Para ello, se revisaron las condiciones de licencia de cada uno de ellos y posteriormente se decidió su incorporación de cara al análisis de resultados para una muestra aleatoria.

Para poder homogeneizar los resultados proporcionados por cada proveedor, se modificaron los módulos individuales correspondientes a cada uno de ellos de cara a poder ofrecer una única clasificación de estado del proceso:

- No encontrado (*not found*): no se devuelven las coordenadas de la dirección postal, sino de una entidad superior (barrio, municipio, etc.) que varía en función de cada proveedor.
- Nulo (*null*): no se devuelven coordenadas.
- Exacto (*exact*): se devuelven lo que para el proveedor son las coordenadas exactas para la dirección postal introducida.
- Cercano (*nearest*): se devuelve un portal o número de calle cercano al de la dirección introducida, pero no exacto.

A continuación, se estudia la caracterización de cada uno de los proveedores seleccionados:

4.1. Google

Google ofrece dos APIs distintas, con similares condiciones de licencia: *geocoder* y *places*. La primera responde a un geocodificador estándar de direcciones postales, mientras que la segunda está más orientada a nombres de lugares, establecimientos, empresas, etc.

Cabe destacar que las condiciones de uso de todas las APIs de *Google* cambiaron el 13 Julio de 2018, siendo reseñable el hecho de **que los resultados de la geocodificación solo pueden almacenarse durante 30 días,** independientemente del coste que supone superar el límite de la capa gratuita.

[2] Application Programming Interface, puntos de servicio en Internet generalmente implementados sobre http y sin estado.

Asimismo, los puntos geocodificados con *Google* solo pueden representarse en los productos cartográficos de *Google*.

En el caso del geocodificador, los siguientes tipos de entidades se han clasificado como *"not found"* o no encontrado: *political, country, administrative_area_level_1, administrative_area_level_2, administrative_area_level_3, administrative_area_level_4, administrative_area_level_5, locality, Ward, sublocality, sublocality_level_1, sublocality_level_2, sublocality_level_3, sublocality_level_4, sublocality_level_5, neighborhood, postal_code, subpremise, post_box, postal_town.*

Como se puede comprobar en los resultados de las tablas *Tabla 2 Comparativa de resultados individuales de geocodificación por proveedor* y *Tabla 3 Coincidencias de proveedores en un radio de 200m para una muestra de 1000 direcciones*, *Google* es, a priori, el proveedor mejor situado en cuanto a número de direcciones geocodificadas y coincidencias con el resto de proveedores en un radio de 200m.

4.2. Cartociudad

Cartociudad presenta dos versiones de servicio web de geocodificación, que para el objeto de este trabajo se han denominado *v1* y *v2* respectivamente, siendo v2 la más reciente. La principal diferencia entre ambas versiones radica en el procesador o *parser* de direcciones postales, ya que la base de datos cartográfica es única.

El proyecto *Cartociudad*[3] no presenta restricciones de licencia y está liderado y coordinado por el Centro Nacional de Información Geográfica y se genera a partir de datos oficiales del IGN, la Dirección General del Catastro, el Grupo Correos y el Instituto Nacional de Estadística. Además, colaboran en su elaboración las comunidades autónomas de País Vasco, Navarra, Comunidad Valenciana, La Rioja, Baleares y Andalucía.

En la siguiente tabla se resume la clasificación de estados de ambas versiones del geocodificador:

	Cartociudad v1	Cartociudad v2
Exacto	1	1
Cercano	2	2,3,4
No encontrado	resto	resto

Tabla 1 Clasificación de estados de cada versión de Cartociudad

[3] http://www.cartociudad.es/portal/web/guest/que-es-cartociudad

En donde los valores devueltos por *Cartociudad v1* son:

- Valor ‘1’: Portal/Pk encontrado
- Valor ‘2’: Portal/Pk no encontrado. Se devuelve el más cercano encontrado
- Valor ‘3’: Portal/Pk no encontrado. Se devuelve el inicial de la vía
- Valor ‘4’: Municipio encontrado, pero no la vía
- Valor ‘0’: Recoge diferentes situaciones. Por un lado, vías encontradas que no tienen números asociados (entendemos que en su sistema), y por otro, viales no encontrados. Devuelve el comentario El vial encontrado (NOMBRE DEL VIAL) no tiene Portal/Pk asociados o el comentario Vial no encontrado. En el primer caso devuelve una geocodificación (que corresponde a un punto central del municipio habitualmente), en el segundo no devuelve ninguna geocodificación

Y los valores devueltos por *Cartociudad v2*:

- Valor '1' para resultados exactos.
- Valor '2' si no se encuentra un portal par. Se ajusta al par más próximo.
- Valor '3' si no se encuentra un portal impar. Se ajusta al impar más próximo
- Valor '4' si no se encuentra un punto kilométrico de un vial. Se ajusta al más próximo
- Valor '5' si no se encuentran portales/puntos kilométricos asociados al vial.
- Valor '6' si el tipo de vía no coincide con el solicitado en la consulta.
- Valor ‘10’ si no encuentra resultados, pero devuelve la entidad superior.

4.3. Bing

Las condiciones de uso de la API de geocodificación de *Bing* permiten hasta un máximo de 125.000 peticiones al año. Asimismo, los puntos geocodificados con esta API solo pueden representarse en los productos cartográficos de este proveedor.

Para este proveedor, los estados resultantes se han clasificado como “not found” o no encontrado para valores de *AdminDivision1, AdminDivision2, ContryRegion, Neighborhood, PopulatedPlace, Postcode1, Postcode2 y RoadBlock.*

Bing no destaca por su precisión, siendo un proveedor que suele ocupar el cuarto o quinto puesto (según se geocodifiquen personas o empresas) dentro de la escala de prioridades. No obstante, aporta utilidad en el proceso de selección y agrupamiento.

4.4. Mapbox

Las condiciones de uso de la API de geocodificación de *Mapbox* permiten hasta un máximo de 50.000 peticiones al mes. Asimismo, los puntos geocodificados con esta API solo pueden representarse en los productos cartográficos de este proveedor.

Mapbox devuelve un valor numérico para el estado del proceso de geocodificación, que se ha clasificado de la siguiente forma:

- Exacto: calidad >= 0.7
- Cercano: 0.66 <= calidad < 0.7
- No encontrado: resto.

Al igual que *Bing*, su uso aporta valor en el proceso de selección y agrupamiento de coordenadas.

4.5. Arcgis

La Unidad de Cartografía y Sistemas de Información Geográfica del Gobierno de Cantabria dispone de licencias de uso de *Arcgis*, si bien no ha sido necesario utilizarlas dado que la API pública no presenta restricciones para geocodificación registro a registro. Por otra parte, los puntos geocodificados con esta API solo pueden representarse en los productos cartográficos de este proveedor. La clasificación de estados para este proveedor es la siguiente:

- Exacto: *PointAddress*
- Cercano: *StreetAddress, StreetAddressExt*
- No encontrado: resto

Arcgis se sitúa en la mejor posición de entre los proveedores “no principales” (es decir, aquellos que no son *Google* o *Cartociudad*).

4.6. Comparativa de resultados

En el siguiente cuadro se presenta un resumen de los resultados obtenidos para una muestra aleatoria de mil direcciones de personas en entorno urbano (no es necesario incluir el mismo análisis para entornos rurales puesto que arroja similares resultados a efectos del propósito del mismo). El estado *“Duplicate”* solo se utiliza para caracterizar la peculiaridad de que *Google Places* referencia con relativa frecuencia diversas direcciones al mismo punto si no se utilizan nombres de lugares o establecimientos.

Estado (%)	google	cartociudadv2	arcgis	cartociudadv1	bing	mapbox	places
Exact	99,8	77,4	75,4	65,2	85,9	74,2	41,6
Nearest	0	17,2	15,9	16,0	0	16,4	0
Duplicate	0	0	0	0	0	0	56,3
Not found	0	5,3	8,7	18,8	14,1	9,4	0
Null	0,2	0,1	0	0	0	0	2,1

Tabla 2 Comparativa de resultados individuales de geocodificación por proveedor

5. Diseño e implementación

El paquete de software de metageocodificación se ha implementado en el lenguaje Python 3. La elección de esta tecnología obedece básicamente a los siguientes motivos:

- Riqueza y variedad de bibliotecas con implementaciones algorítmicas para trabajo con datos geográficos, clasificación y APIs restful: pandas (Python Data Analysis Library, s.f.), scikit-learn (Scikit-learn: Machine Learning in Python, s.f.), shapely (Shapely, s.f.), geopy (Geopy, s.f.).
- Facilidad de mantenimiento e integración con los marcos de trabajo ETL[4] desarrollados en el ICANE.

El paquete consta de los siguientes módulos diferenciados:

- *metageocoder.py*: módulo principal del geocodificador que se vale del resto de módulos y bibliotecas para ofrecer funcionalidades de geocodificación masiva de direcciones postales y selección de la mejor coordenada de entre las ofrecidas por un conjunto de proveedores dato.
- *geoprovision.py*: contiene toda la lógica relacionada con los distintos proveedores de geocodificación, ofreciendo facilidades para lanzar llamadas a dichos proveedores y obtener resultados a través de sus APIs. Puede ser extendido fácilmente con nuevos proveedores.
- *geocoding_mode.py*: implementa distintos modos de geocodificación, diferenciándose estos entre personas (mortalidad, tarjeta sanitaria) y establecimientos (directorio de empresas). Los modos de geocodificación afectan a aspectos como la lectura y preprocesado de los distintos ficheros que contienen las direcciones postales. Puede ser extendido fácilmente con nuevos modos de geocodificación.
- *geography.py*: contiene la lógica necesaria para localización de puntos en municipios o en una región a través de un mapa en formato *geojson*[5], así como de tratamiento de las direcciones postales.

[4] Extracción, Transformación y Carga (Load)

- *geometry.py*: ofrece funcionalidades de cálculo relacionado a la geometría de los puntos especificados por las distintas coordenadas, tales como *clustering*, cálculo de distancia *Haversine*, de centroides, etc.
- *analysis.py*: el módulo de análisis examina los datos individuales por proveedor y genera resúmenes estadísticos relevantes para usar como criterio de asignación de prioridad

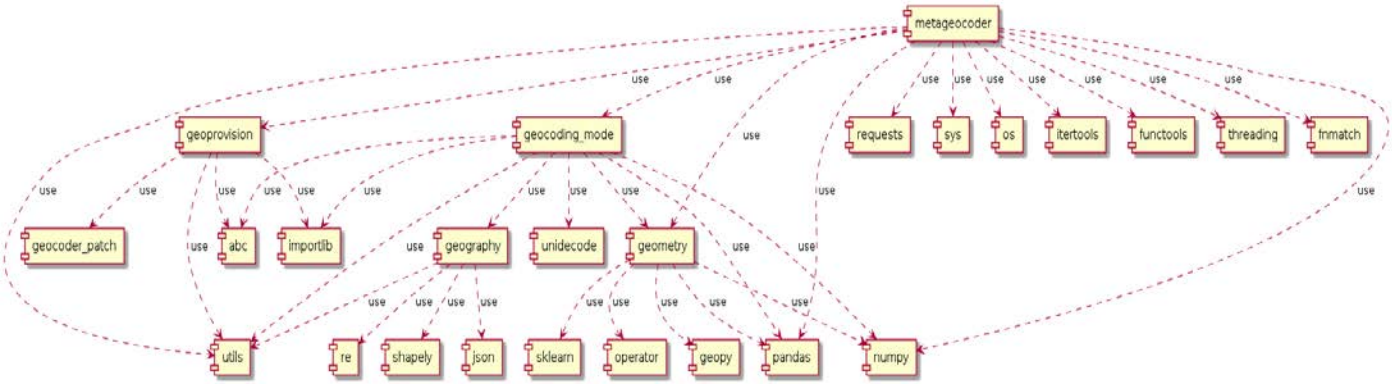

Ilustración 1 Diagrama de componentes

6. Algoritmo de decisión

6.1. Inicialización

En esta etapa se carga la configuración para el presente proceso de geocodificación. Entre los parámetros configurables más relevantes cabe citar la posibilidad de decidir si un proveedor es elegible o no por términos de licencia o cualquier otro motivo (es decir, si podrá ser seleccionado como la mejor opción o sólo contribuirá a efectos de agrupamiento). También es posible configurar la preferencia a priori de cada uno de ellos, que puede definirse por criterios estadísticos. Por ejemplo, para la geocodificación de personas se ha utilizado la siguiente lista de prioridades: *Google*, *CartociudadV2*, *Arcgis*, *CartociudadV1*, *Bing*, *Mapbox* y *Google Places*. En el caso de empresas, por otra parte, es relevante destacar que *Google Places* ocupa la primera posición de prioridad (ver análisis).

El criterio para decidir la preferencia fue el mayor número de coincidencias de geocodificación por proveedor (entendida la coincidencia en un radio inferior a 200m) para una muestra de 1000 direcciones aleatorias. Por ejemplo, para un conjunto de datos de mil direcciones urbanas, se tiene:

Coincidencia (%)	arcgis	bing	cartociudadv1	cartociudadv2	google	mapbox	places
arcgis		82,9	67,5	76,2	82,3	70,4	24,2

[5] http://geojson.org/, un formato para codificar estructuras geográficas varias.

bing	82,9		65,8	73,1	79,1	67,6	22,6
cartociudadv1	67,5	65,8		74,1	75,1	63,1	19,6
cartociudadv2	76,2	73,1	74,1		84,0	69,6	23,0
google	82,3	79,1	75,1	84,0		77,1	25,1
mapbox	70,4	67,6	63,1	69,6	77,1		25,2
places	24,2	22,6	19,6	23,0	25,1	25,2	

Tabla 3 Coincidencias de proveedores en un radio de 200m para una muestra de 1000 direcciones

Posteriormente, se leen los datos de entrada y se realiza un pre-procesado básico de direcciones postales (limpieza de tipos de vía, números de portal, códigos de municipio, etc.).

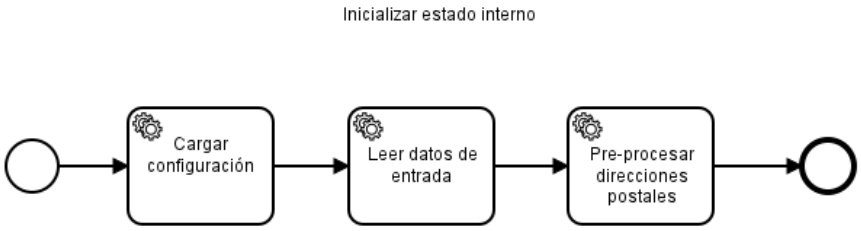

Ilustración 2 Inicialización del geocodificador

6.2. Geocodificación propiamente dicha

Una vez inicializado el proceso global, se lanzan N hilos de geocodificación (uno por proveedor). Estos se ejecutan en paralelo lanzando peticiones a las APIs correspondientes de los distintos proveedores. Al finalizar la ejecución de cada hilo, se almacenan tanto los resultados correctos como los erróneos en caso de existir. Si hay volcado de erróneos, estos se vuelven a procesar automáticamente hasta que se resuelvan correctamente; en caso de haber errores irrecuperables, estos se vuelcan en otro archivo (además de incorporarse con su resultado vacío al fichero de direcciones geocodificadas). El motivo de esta peculiaridad es mantener el mismo número de registros en cada archivo de volcado individual para facilitar posteriormente su fusión.

Finalmente, cuando todas las direcciones han sido geocodificadas, los resultados de todos los proveedores se fusionan en un único archivo para su post-procesado.

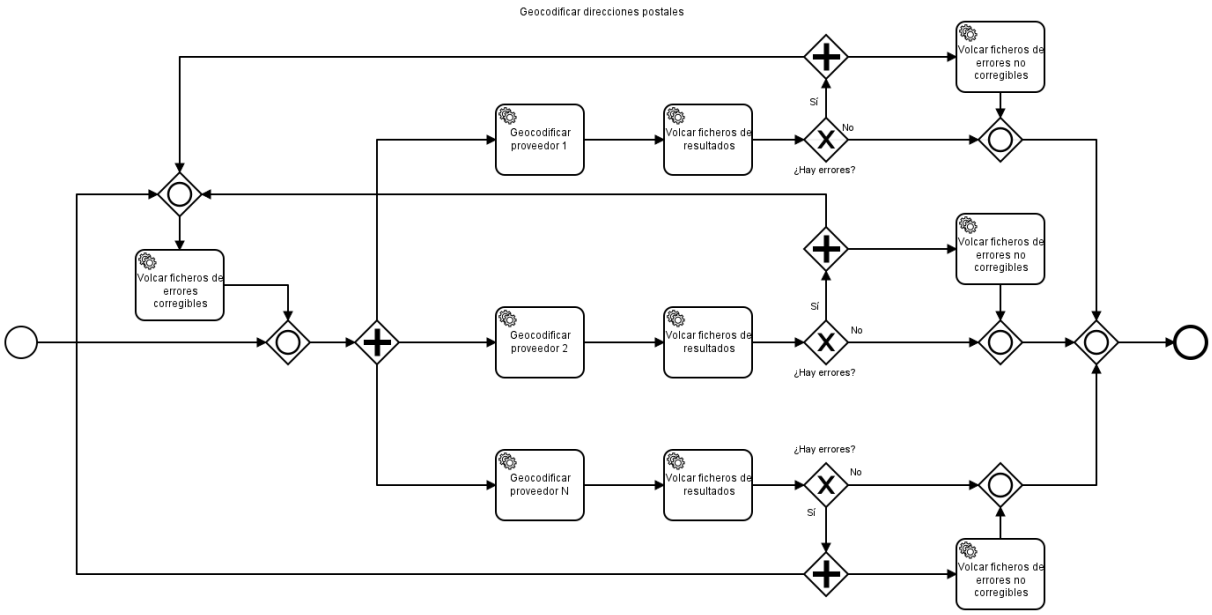

Ilustración 3 Proceso de geocodificación

6.3. Decisión de la mejor coordenada

Antes de comenzar el proceso de decisión, se llevan a cabo una serie de operaciones sobre las coordenadas en bruto. En primer lugar, se aplican filtros de calidad sobre cada una de las coordenadas proporcionadas por cada proveedor:

- Que el municipio al que pertenece la coordenada devuelta no sea distinto del municipio introducido en la dirección postal.
- Que el estado del resultado devuelto por el geocodificador sea un valor entre *"Exact"* y *"Nearest"* (ver sección *4. Análisis*).

De no cumplir con estas condiciones, la coordenada se sobre-escribe con un valor nulo.

Si tan solo se dispone de una coordenada candidata, el registro se marca para su revisión. En caso de disponer de más de una coordenada candidata, se lleva a cabo un primer agrupamiento o *clustering* de las coordenadas de los proveedores en un radio de 200m. Es decir, a todos aquellos puntos de cada proveedor que se encuentren en un radio de 200m o menos, se les asigna una misma etiqueta. Para todas aquellas direcciones a las que no se ha asignado etiqueta en el primer agrupamiento, se repite el proceso para un radio de 500m. Finalmente, se seleccionan las mejores coordenadas en base al criterio de selección resumido en la sección *7.5. Criterio de selección*.

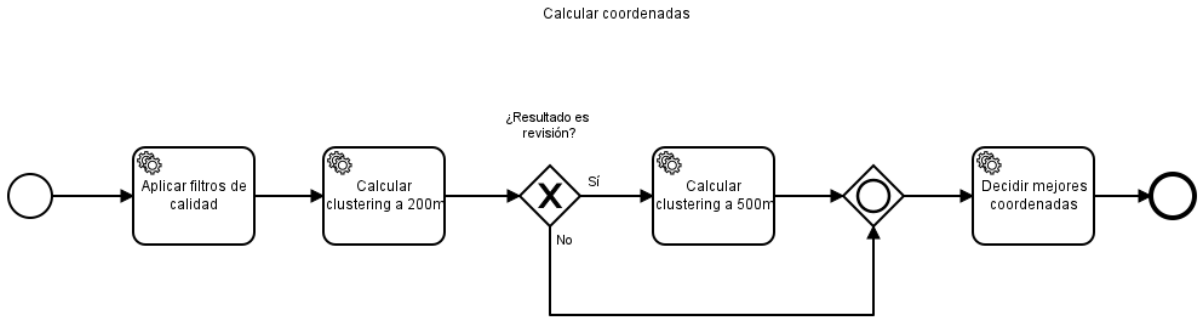

Ilustración 4 Procesado de coordenadas

6.4. Algoritmo de *clustering*

Para realizar los agrupamientos de puntos en el mapa se ha utilizado el algoritmo DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*). Dado un conjunto de puntos en el espacio, los agrupa en función de su distancia de *Haversine*[6], marcando como atípicos aquellos que se encuentran aislados en regiones de baja densidad. Presenta muchas ventajas, entre las que se puede destacar que no es necesario especificar el número de *clusters* o agrupamientos, pudiendo encontrar *clusters* de forma arbitraria de forma relativamente rápida. El algoritmo se parametriza con los siguientes parámetros:

- Épsilon (ξ): máxima distancia entre dos muestras para considerarlos en la misma agrupación.
- Min_samples: número de muestras en una agrupación para que un punto se considere como "núcleo" (incluido el propio punto).

En el presente trabajo, se ha utilizado un valor de 2 para *min_samples* y de 200 ó 500 para ξ, según el caso.

Resumidamente, su funcionamiento es el siguiente (Ester, Kriegel, Sander, & Xu, 1996):

1) Para cada punto del conjunto de datos, se calcula una esfera n-dimensional de radio épsilon alrededor de dicho punto.
2) Si el número de puntos dentro de la esfera es mayor que *min_samples,* se fija el centro de la esfera como un *cluster* y todos los puntos dentro de dicha esfera pertenecen a este *cluster*.
3) Se iteran los dos pasos anteriores a lo largo de todos los puntos dentro de la esfera y se expande el *cluster* cuando se satisfagan ambas.
4) Los puntos que no pertenezcan a ningún *cluster* se ignoran o tratan como atípicos.

[6] La fórmula de *Haversine* determina la distancia de círculo máximo entre dos puntos de una esfera dadas sus longitudes y latitudes. Básicamente, se utiliza para calcular distancias entre pares de coordenadas (lontigud, latitud).

6.5. Criterio de selección

Para seleccionar la mejor coordenada de entre las candidatas, en primer lugar se selecciona un *cluster* de entre los calculados (Ver *Ilustración 5 Selección del cluster*). Si no existe ninguno, el registro se marca para revisión. Si sólo hay un *cluster*, este es seleccionado. En caso de haber más de uno, el criterio de selección de *cluster* es elegir aquel al que pertenece el primer proveedor elegible con coordenada no nula.

6.5.1. Proceso de decisión para empresas

Para el caso de direcciones pertenecientes a establecimientos de empresas, se utiliza un procedimiento especial. Este procedimiento pretende aprovecharse de la información introducida por los propios usuarios de *Google Places*, que en muchos casos extiende y mejora la proporcionada por el geocodificador simple. *Google Places* permite obtener coordenadas a partir de una dirección postal asociada a su razón social, nombre de la empresa o del establecimiento. En concreto, si *Google Places* ha tenido éxito, su coordenada se selecciona como la mejor. En caso contrario, se verifica si la dirección cuenta con el campo de número de portal. De no tenerlo, se marca para revisión; si por el contrario lo tiene, la dirección se envía al proceso común de decisión descrito en el siguiente apartado (Ver *Ilustración 6 Proceso de decisión para empresas*).

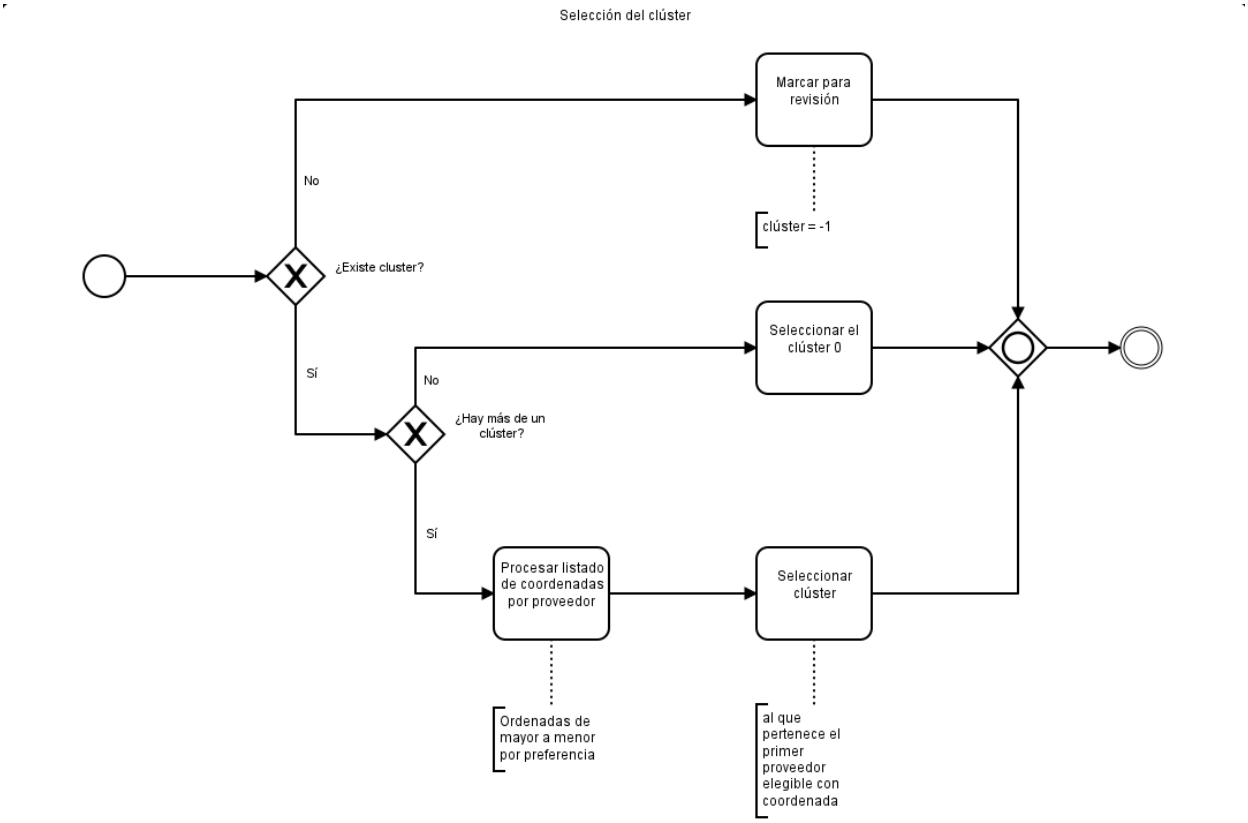

Ilustración 5 Selección del cluster

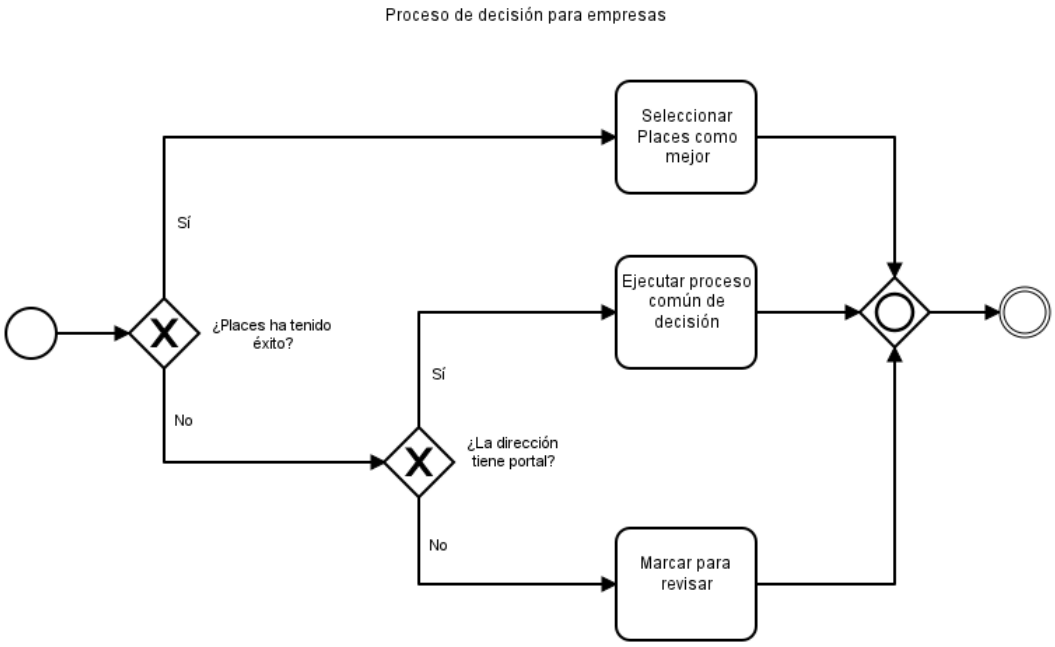

Ilustración 6 Proceso de decisión para empresas

6.5.2. Proceso de decisión común

Una vez seleccionado el *cluster*, se procesa el listado de coordenadas por proveedor, ordenándolas por preferencia. Si hay menos de 3 proveedores en la lista, se extrae el primero elegible (que será tomado como mejor coordenada). En caso contrario, se ordena el listado de proveedores de mayor a menor por su distancia al centroide del *cluster* al que pertenecen y se extrae el primero elegible (que será tomado como mejor coordenada).

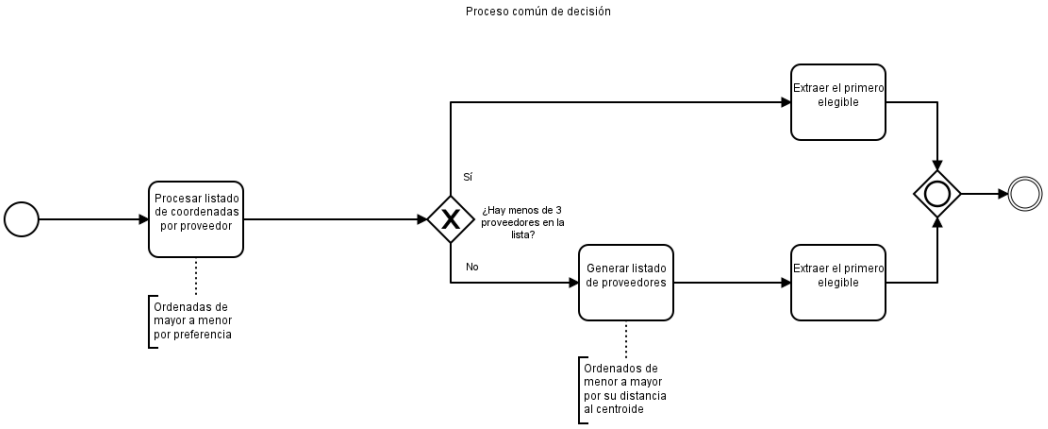

Ilustración 7 Proceso de decisión común

7. Resultados

Para estudiar la calidad de los resultados del geocodificador, se ha lanzado un proceso con una muestra aleatoria de 1000 direcciones de personas residentes en Cantabria en el entorno urbano (fuente: Tarjeta Sanitaria) para todos los proveedores y se han comprobado manualmente sus resultados. Este mismo ejercicio se ha repetido con otra muestra de igual tamaño de residentes en entorno rural (Alegría Montes, Bedia Expósito, Expósito Martín, & Mayoral de la Vega, 2018).

En la *Tabla 4 Evaluación de resultados de tarjeta sanitaria en entorno urbano por proveedor* y la *Tabla 5 Evaluación de resultados de tarjeta sanitaria en entorno rural por proveedor* se han resumido los siguientes datos por proveedor:

- OK: direcciones geocodificadas correctamente.
- Cercano: direcciones cuya geocodificación devuelve una coordenada a menos de 50m (aproximadamente) de la correcta.
- Incorrecto: direcciones cuya geocodificación devuelve una coordenada a más de 50m (aproximadamente) de la correcta.
- Coordenada nula o de calidad insuficiente: direcciones que, o bien no han obtenido coordenada en el proceso de geocodificación, o bien no han superado los filtros de calidad aplicados a las mismas.
- No confirmable: direcciones cuya coordenada geocodificada no se ha podido confirmar correcta.

7.1. Entorno Urbano

Evaluación (%)	google	cartociudad v2	arcgis	cartociudad v1	bing	mapbox	places
OK	91,5	77,6	78,3	70,6	77,4	64,4	13,2
CERCANO	2,9	6,2	5,4	4,0	3,8	13,9	12,7
INCORRECTO	2,8	8,7	5,1	4,3	3,4	10,1	13,1
COORDENADA NULA O DE CALIDAD INSUFICIENTE	0,3	5,4	9,3	18,8	14,1	9,9	59,4
NO CONFIRMABLE	2,5	2,1	1,9	2,3	1,3	1,7	1,6
TOTAL	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**

Tabla 4 Evaluación de resultados de tarjeta sanitaria en entorno urbano por proveedor

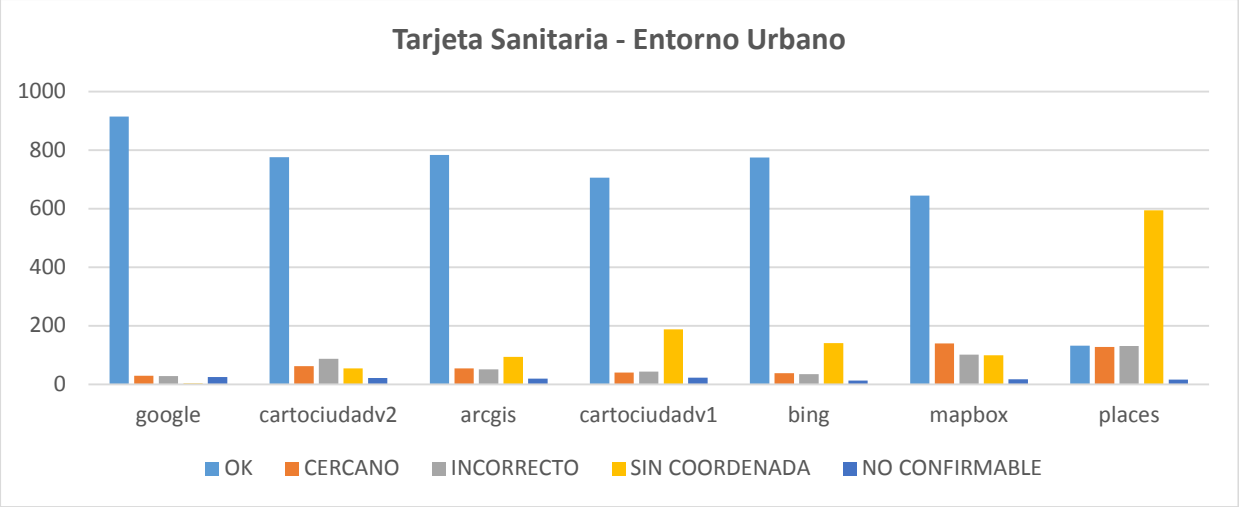

Ilustración 8 Evaluación de resultados de tarjeta sanitaria en entorno urbano por proveedor

7.2. Entorno Rural

Evaluación (%)	google	cartociudad v2	arcgis	cartociudad v1	bing	mapbox	places
OK	71,3	53,7	38,1	51,4	36,2	38,8	20,4
CERCANO	7,1	9,6	9,6	8,1	5,1	11,2	14,4
INCORRECTO	6,8	8,0	14,6	6,9	8,2	11,3	14,0
COORDENADA NULA O DE CALIDAD INSUFICIENTE	3,5	23,1	31,4	27,9	45,8	34,2	46,3
NO CONFIRMABLE	11,3	5,6	6,3	5,7	4,7	4,5	4,9
TOTAL	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**

Tabla 5 Evaluación de resultados de tarjeta sanitaria en entorno rural por proveedor

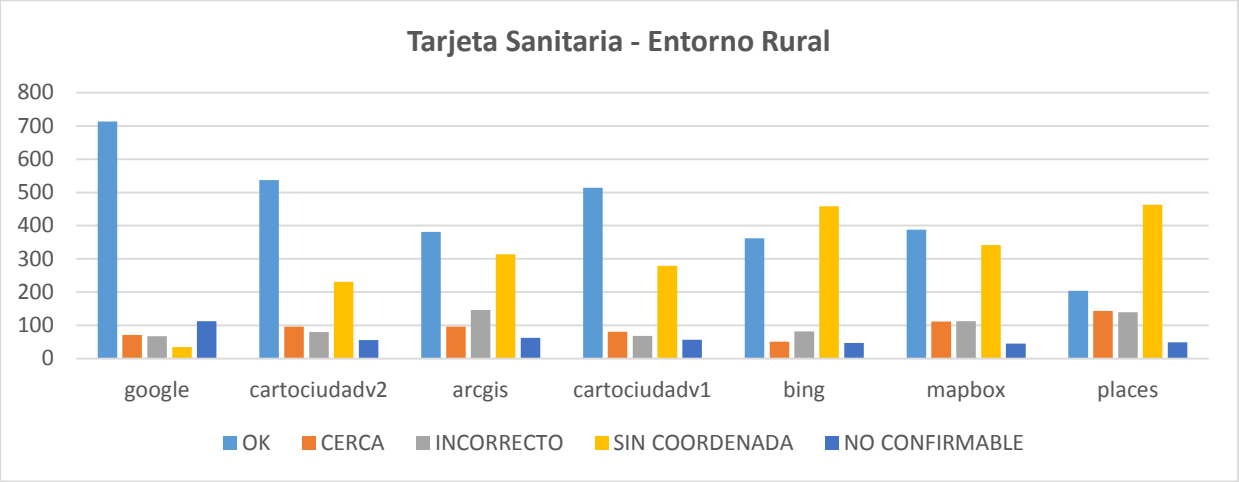

Ilustración 9 Evaluación de resultados de tarjeta sanitaria en entorno rural por proveedor

8. Conclusiones

Como se desprende de los datos mostrados procedentes de las direcciones postales de tarjeta sanitaria, *Google* es el proveedor que presenta un mayor acierto en la geocodificación y *Google Places* no ofrece buenos resultados cuando no se introducen nombres de empresa o establecimientos (como ha sido el caso). Estos resultados confirman que el orden de preferencia establecido a priori es adecuado para el proceso de geocodificación múltiple, siendo también destacable la diferencia de acierto entre las dos versiones de *parsers* de *Cartociudad*, superando la nueva versión a la anterior.

En el caso de la muestra de direcciones postales de personas residentes en el entorno urbano, únicamente han necesitado revisión manual un 0,2% del total. De los resultados obtenidos con éxito (un 99,8%), en un 94,5% del total de la muestra se han evaluado manualmente como exactos o cercanos al punto real, en un 2,6% pueden existir dudas y en un 2,7% el resultado ha sido erróneo.

Mejor (%)	OK	CERCA	DUDA	ERROR	REVISIÓN	Total
google	36,6	1,5	1,3	0,8	0,0	**40,2**
cartociudadv2	29,0	1,0	0,6	1,6	0,0	**32,2**
cartociudadv1	25,5	0,9	0,7	0,3	0,0	**27,4**
revisión	0,0	0,0	0,0	0,0	0,2	**0,2**
Total	**91,1**	**3,4**	**2,6**	**2,7**	**0,2**	**100,0**

Tabla 6 Evaluación de resultados del proveedor elegido para direcciones urbanas

En el caso de direcciones postales de datos personas residentes en el entorno rural, en el proceso de geocodificación se habían obtenido resultados para un 89,1% de los registros, necesitando un 10,9% de las direcciones revisión manual. De estas últimas, en un 24,8% no se han obtenido resultados individuales válidos de ningún proveedor, en un 38,5% hay varios resultados lo suficientemente separados como para no poder elegir ninguno de ellos ni siquiera en el proceso manual y únicamente en el 36,7% de los casos a revisión (un 4,0% del total de la muestra) se podría asignar un proveedor de forma manual.

En cuanto al análisis de los casos en los que el proceso había dado un proveedor como el mejor, un 77,0% de la muestra total tiene resultados exactos o cercanos al punto real, en un 8,0% puede haber dudas entre las obtenidas y en un 4,1% se ha detectado que la coordenada elegida es errónea (3,0% en los que otro proveedor tiene la coordenada correcta, 0,2% sobre los que hay dudas y 0,9% sobre los que no se ha encontrado ninguna coordenada correcta).

Mejor (%)	OK	CERCA	DUDA	ERROR	REVISIÓN	Total
google	32,2	3,6	4,4	1,7	0,0	**41,9**
cartociudadv2	19,6	2,7	2,0	1,8	0,0	**26,1**
cartociudadv1	16,2	2,7	1,6	0,6	0,0	**21,1**
revisión	0,0	0,0	0,0	0,0	10,9	**10,9**
Total	**68,0**	**9,0**	**8,0**	**4,1**	**10,9**	**100,0**

Tabla 7 Evaluación de resultados del proveedor elegido para direcciones rurales

Tal y como se esperaba a priori, la calidad del proceso de geocodificación es sensiblemente mejor en entornos urbanos que en entornos rurales. Esto es consistente con la calidad observada en las distintas bases de datos geográficas de proveedores relevantes como *Google* o *Cartociudad*.

Con los datos presentados, es posible calcular la precisión de cada proceso de geocodificación. Concretamente, comparando el producto desarrollado con el mejor proveedor único (*Google*), se tendría:

Precisión = VP / (VP + FP),

donde VP son los verdaderos positivos y FP los falsos positivos.

Precisión (%)	Google	Metageocodificador
Entornos rurales	92,0	95,0
Entornos urbanos	97,1	97,2

Tabla 8 Comparativa de precisión entre el mejor proveedor y el metageocodificador

No obstante, la precisión del metageocodificador podría mejorarse una vez incorporados los resultados de aplicar una revisión manual a aquellos casos que así hubieran sido clasificados.

Se puede concluir que el producto desarrollado mejora levemente los resultados del principal proveedor de geocodificación, proporcionando también el marcaje de determinadas direcciones para su revisión posterior y así facilitar la reducción de falsos positivos. Además, su diseño permite utilizar distintos proveedores y aglutinar sus resultados, siendo especialmente interesante esta característica para superar problemas derivados de condiciones de uso restrictivas de las diferentes APIs de geocodificación (véanse, por ejemplo, las nuevas condiciones de licenciamiento de *Google*[7]). Es especialmente relevante recalcar que, si bien estos son los resultados para direcciones postales de personas físicas a las que se ha aplicado un pre-procesado, los análisis que se están llevando a cabo en el momento de redacción de la presente ponencia no parecen indicar que para el caso de establecimientos de empresas estos resultados vayan a ser similares (debido fundamentalmente a una peor calidad de las direcciones y a la probable peor calidad de las bases de datos geográficas para estos casos).

[7] https://cloud.google.com/maps-platform/

Bibliografía

Alegría Montes, L., Bedia Expósito, R., Expósito Martín, R., & Mayoral de la Vega, B. (2018). *Informe interno de evaluación de los resultados de geocodificación.*

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorith for Discovering Clusters in Large Spatial Databases with Noise. *KDD-96 Proceedings*, 6.

Geopy. (s.f.). Obtenido de https://pypi.org/project/geopy/

Kent Beck, C. A. (2004). *Extreme Programming Explained: Embrace Change.*

Python Data Analysis Library. (s.f.). Obtenido de https://pandas.pydata.org/

Scikit-learn: Machine Learning in Python. (s.f.). Obtenido de http://scikit-learn.org/

Shapely. (s.f.). Obtenido de https://github.com/Toblerity/Shapely