GEOBIG. Gestión de grandes volúmenes de datos abiertos

GEOBIG. Open data big volumes management

Alejandro Guinea de Salas, Kepa López Pérez, Daniel Navarro Cueto

REVISTA **MAPPING**
Vol. 29, 199, 46-50
enero-febrero 2020
ISSN: 1131-9100

Resumen

La información geográfica disponible está creciendo a pasos agigantados. Cada día se publican nuevos datos, a menudo a través de portales de datos abiertos, lo que implica inversiones nada desdeñables para su creación y mantenimiento. Sin embargo, son a menudo difíciles de reaprovechar, dada su heterogeneidad y la variedad de formas de acceso.

Paralelamente, la tecnología relacionada con la inteligencia artificial ha sufrido un gran avance, sin embargo, es difícil aplicar esas tecnologías a grandes conjuntos de datos abiertos, por su dispersión y la dificultad en procesarlos. Por ello, es necesario preparar los datos de forma previa a cualquier proceso de análisis.

Se expondrá cómo se han preprocesado más de 125 000 capas de todo el mundo, que contienen más de 250 millones de elementos, y cómo se han preparado e integrado en una arquitectura capaz de aplicar procesos de forma automatizada, independientemente de sus características.

Se exponen aspectos como la identificación y acceso a los datos, y los procesos seguidos hasta obtener un sistema capaz de procesar miles de capas en paralelo, como el enriquecimiento de metadatos. El sistema permite obtener el máximo valor de la ingente cantidad de datos geográficos disponibles, ya sean datos INSPIRE o no.

Abstract

The available geographic information is growing by leaps and bounds. New data is published every day, often through open data portals, which means investments that are not negligible for its creation and maintenance. However, they are often difficult to reuse, given their heterogeneity and the variety of forms of access.

At the same time, technology related to artificial intelligence has undergone a great advance, however, it is difficult to apply these technologies to large open data sets, due to their dispersion and the difficulties in processing them. Therefore, it is necessary to prepare the data prior to any analysis process.

It will be exposed how more than 125,000 layers from all over the world have been preprocessed, containing more than 250 million elements, and how they have been prepared and integrated into an architecture capable of applying processes in an automated way, regardless of their characteristics.

Aspects such as identification and access to data, and the processes followed until obtaining a system capable of processing thousands of layers in parallel, such as enrichment of metadata are exposed. The system allows to obtain the maximum value of the huge amount of geographic data available, whether they were INSPIRE data or not.

Palabras clave: Big data, metadatos, procesos en la nube, inteligencia artificial, aprendizaje automático, información geográfica, datos abiertos.

Keywords: Big data, metadata, cloud computing, artificial intelligence, machine learning, geographic information, open data.

Geograma
alejandro.guinea@geograma.com
kepa.lopez@geograma.com
daniel.navarro@geograma.com

Recepción 12/12/2019
Aprobación 13/12/2019

1. INTRODUCCIÓN

Sin datos territoriales es imposible gestionar el territorio, y por tanto tomar las decisiones correctas a la hora de afrontar el cambio climático.

El proyecto *GeoBIG* pone a disposición de los usuarios información geográfica homogénea, en términos de formato, sistema de referencia y clasificación, haciendo el uso de los datos más barato, más rápido y más fácil.

El objetivo es acceder y monitorizar conjuntos de datos de alto valor, sin la problemática de encontrarlos y transformarlos, cargándolos directamente en su herramienta GIS favorita.

La información geográfica (IG) es un tipo de información muy cara, difícil de capturar y compleja de gestionar. Existe un ingente volumen de datos disperso y heterogéneo a nivel de formato, semántica y sistema de referencia.

Los cimientos de la transformación digital ya están colocados, pero es necesario simplificar los requerimientos para que sean parte de los procesos reales y de negocio.

Ver los datos es sólo una pequeña parte del valor que se puede extraer de la información geográfica.

A pesar de su heterogeneidad, la IG es un tipo de información muy estructurada, después de altas y largas inversiones. Sin embargo, su uso está muy limitado más allá del propietario de los datos. Sólo grandes empresas con grandes recursos pueden invertir para reutilizar los datos.

La tecnología está de nuestro lado, permite procesar fácilmente altos volúmenes de datos de este tipo de información.

Por último, pero no menos importante, el valor más alto cuando se trabaja con IG es la dificultad de comparar datos para realizar análisis a lo largo del tiempo.

2. OBJETIVO Y FASES

El proyecto *GeoBIG* acomete las necesidades planteadas mediante:

- La localización de los conjuntos de datos geográficos.
- La transformación para su uso inmediato.
- La clasificación y resolución de cuestiones semánticas.
- La monitorización y actualización de los datos.

3. IDENTIFICACIÓN DE DATOS

El proceso comienza con la identificación y captura de datos geográficos, accesibles y descargables. Normalmente comprimidos en formato ZIP. La URL padre se almacena para su uso posterior en futuras actualizaciones de datos.

En este punto la gran dificultad es la búsqueda de los recursos, debido a su dispersión, y a la muy diversa forma que tienen los proveedores de datos de ponerlos a disposición del público, algunas de ellas son:

- Servidor HTTP.
- Servidor FTP.
- Aplicación web del tipo «añadir a la cesta».
- Interfaz gráfica con mapas para seleccionar la información.
- Envío por correo electrónico.
- Descarga fichero a fichero.

Todos ellos con diferentes formas de autenticación, desde las más sencillas vía formulario, hasta las más complejas mediante petición y envío de un *token* específico para la descarga.

Por último, la información encontrada dentro de los ficheros ZIP es de gran heterogeneidad, por ejemplo, un *ZIP* con tres ficheros en formato *SHP*, cientos de ficheros *SHP*, *GML*, *Excels*, etc.

4. PREPROCESO Y CARGA EN EL SISTEMA

Una vez localizada la información, es necesario revisarla y preprocesarla para que esté lista para procesos posteriores. Aquí se plantea el gran reto del volumen a procesar. El tiempo de cálculo es elevado, y puede ocasionar cuellos de botella. Por tanto, es necesario plantear una arquitectura que permita el proceso en paralelo. Mediante la utilización de colas compartidas, se ha conseguido una arquitectura que es capaz de utilizar decenas de ordenadores independientes trabajando en el procesado de los datos.

Los aspectos tenidos en cuenta en esta fase son:

- Revisión del contenido de los conjuntos de datos, para asegurarse de que contienen información geográfica procesable.
- Confirmación de la disponibilidad de los datos, y de la correcta descompresión de los mismos.
- Captura de metadatos adicionales disponibles en la URL
- Comprobación y registro de volumen y fecha.

Una vez superados los procesos anteriores, se puede considerar que los datos están listos para los siguientes procesos.

Actualmente se han preprocesado unos 40 000 conjuntos de datos.

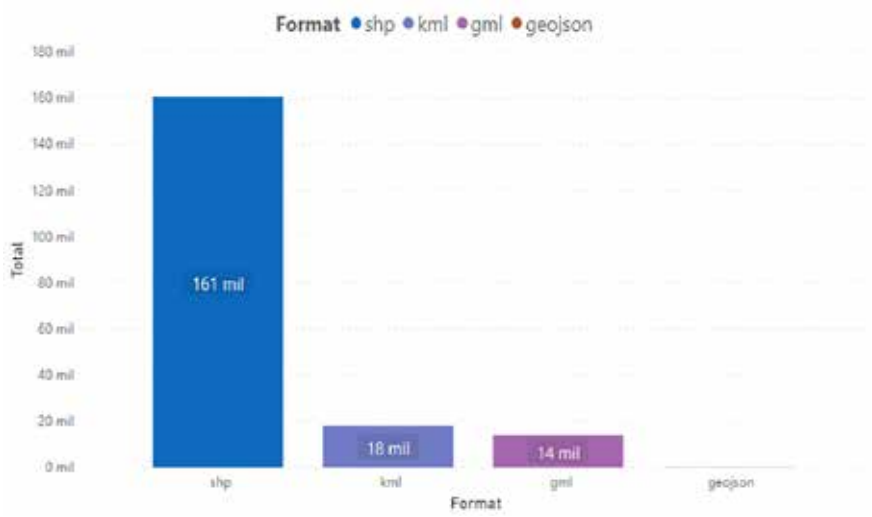

Figura 1. Distribución de capas por formato

5. TRANSFORMACIÓN

Para poder procesar y analizar realmente los datos, es necesario transformarlos en un formato y sistema de referencia homogéneo. Sólo así será posible lanzar algoritmos que permitan extraer el alto valor añadido intrínseco que contienen, como por ejemplo algo tan aparentemente sencillo como de dónde son los datos.

En este proceso de transformación se reproyectan los datos y se cambian a un formato relacional (*geopackage*), de tal forma que se almacenan en tablas independientemente de su origen GML, KML, SHP, JSON, etc., y se separan por capas. En la siguiente tabla se pueden ver las capas procesadas en función de su formato.

De esta forma el acceso a los datos está ya simplificado enormemente, y se pueden plantear procesos más complejos.

6. ANÁLISIS

El primer análisis, más inmediato y de alto interés, es la localización de los conjuntos de datos. Para ello se realizó una superposición de los datos con una capa de niveles administrativos, localizando así todos aquellos conjuntos que tuvieran definido un sistema de referencia. Para aquellos que no lo tuvieran, la localización se limitó a la que se pudo obtener de la URL de descarga, concretamente el IP y/o el país del dominio.

Esto nos ha permitido tener perfectamente localizados las casi doscientas mil capas, facilitando enormemente la búsqueda y el acceso a los mismos.

El siguiente e importante reto es la clasificación de los datos. Debido a que ya están preparados más de 300 millones de elementos, se plantean técnicas de Big Data e inteligencia artificial. Concretamente, se ha utilizado el método de clasificación no supervisada *K-means*, con el fin de encontrar estructuras en los da-

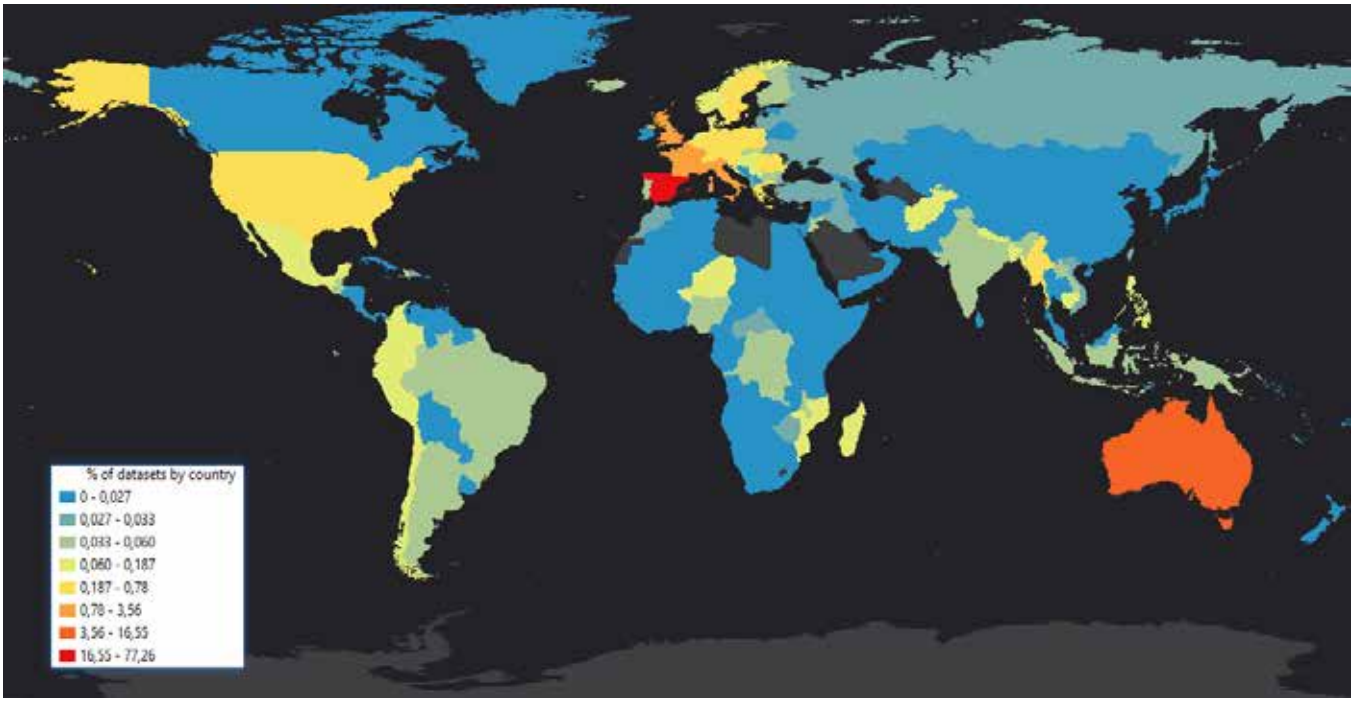

Figura 2. Capas por país

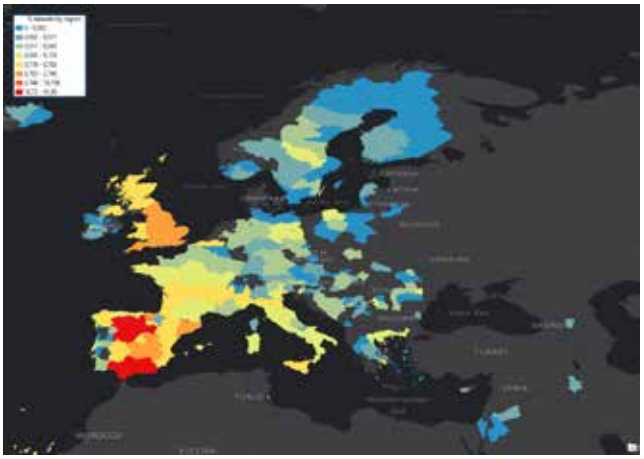

Figura 3. Capas por regiones de Europa

tos sin utilizar etiquetas.

En primer lugar, se han extraído parámetros geométricos de unas 700 capas de puntos. Estos parámetros de alguna forma analizan la dispersión de los puntos, como la distancia media al centroide, la desviación estándar de dicha distancia, y el índice de cohesión. Por ejemplo, una serie de puntos distribuidos a lo largo del área tendrían alta desviación estándar y bajo índice de cohesión, mientras que una serie de puntos agrupados en una esquina del área tendría un alto índice de cohesión, pero baja desviación estándar.

A continuación, se obtiene un diagrama de pares (fig. 4) donde puede evaluarse la relación existente entre las diferentes variables.

Se puede observar la existencia de una correlación

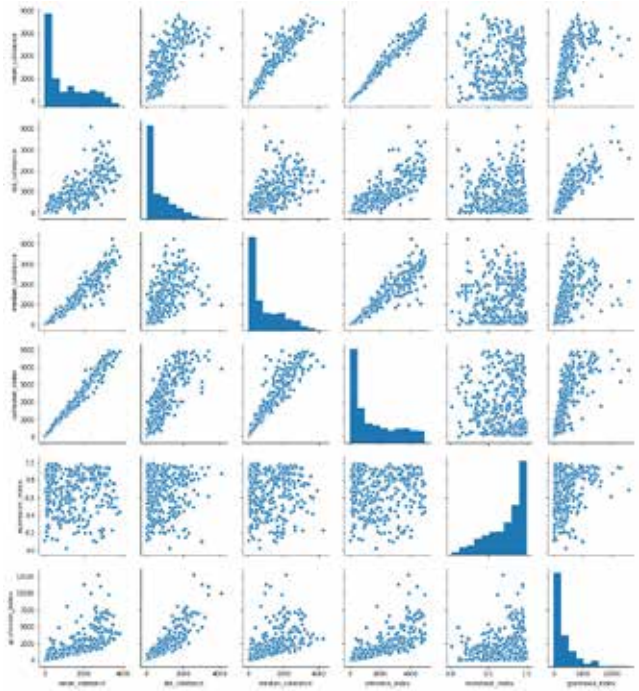

Figura 4. Gráfico de pares entre las variables evaluadas

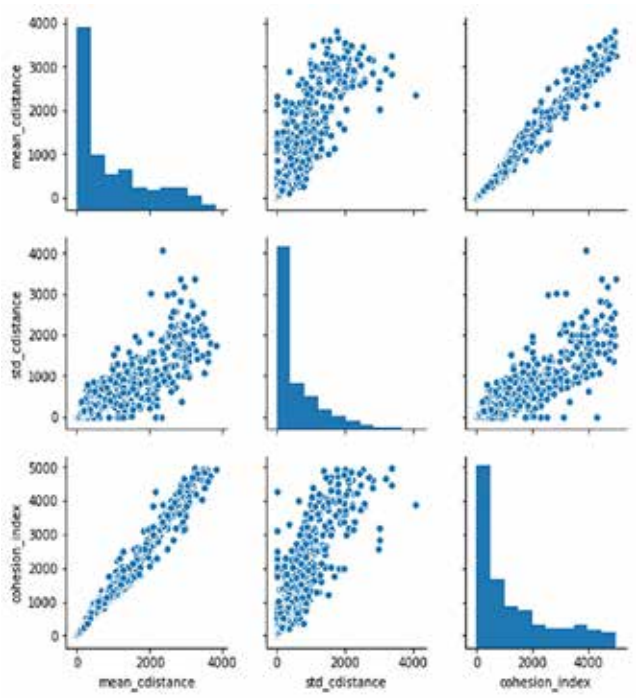

Figura 5. Gráfico de pares entre las seleccionadas

bastante significativa entre las variables de las capas analizadas, concretamente en los nueve primeros gráficos que corresponden a las variables distancia media al centroide, desviación estándar e índice de cohesión

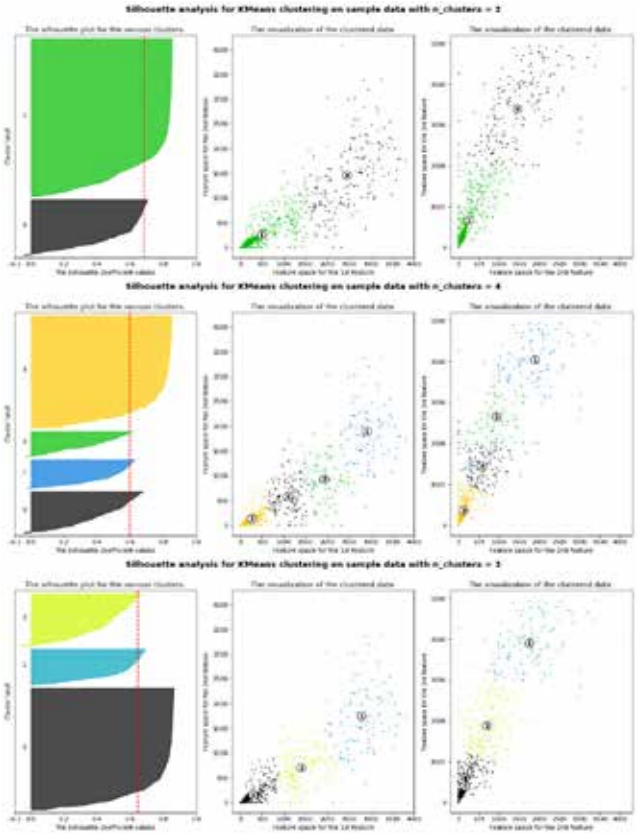

Figura 6. Resultados del clustering mediante KMeans

(fig. 5). Estas variables son las seleccionadas para el análisis posterior.

Una vez elegidas y contrastados los pares que tienen correlación, se ha aplicado un proceso de aprendizaje no supervisado mediante la técnica de *K-Means*, primero en dos, tres y finalmente cuatro clústeres. Se puede observar cómo se obtiene una clasificación con un índice de fiabilidad razonable en cuatro grupos.

7. CONCLUSIONES Y PRÓXIMOS PASOS

Se ha podido comprobar el potencial de la aplicación de algoritmos de inteligencia artificial a la información geográfica, una vez que ésta se encuentra bien estructurada y organizada, lo que abre la puerta a una nueva forma de enfocar la organización y procesado de la misma. Mediante un análisis de las variables implícitas en los datos, se ha observado cómo existen correlaciones y clústeres en los datos, que muestran indicios de la posibilidad de utilizar los algoritmos de inteligencia artificial para la clasificación de los datos, primer paso para asegurar la usabilidad y facilitar el estudio de los mismos.

Lo siguientes pasos van en la línea de seguir identificando y preprocesando recursos, y a analizar los resultados de la clusterización para seguir avanzando en el modelo que permita la clasificación automática de capas.

A nivel de geometría, se seguirán buscando datos estadísticos geométricos cada vez más significativos y relevantes para buscar patrones o propiedades comunes entre los conjuntos de datos. A nivel de atributos, se están realizando acciones para conocer la semántica de los datos (idioma, palabras clave, etc.)

REFERENCIAS

Lloyd, Stuart P. (1982). Recuperado de: "Least squares quantization in PCM"(PDF). http://www-evasion.imag.fr/people/Franck.Hetroy/Teaching/ProjetsImage/2007/Bib/lloyd-1982.pdf

Open Geospatial Consortium. (2017) OGC® GeoPackage Encoding Standard. Recuperado de: http://www.geopackage.org/spec121/index.html

Parlamento Europeo y Consejo de la Unión Europea. Directiva INSPIRE. (2007). Recuperado de: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:ES:PDF

Sobre los autores

Alejandro Guinea

Director y Consultor GIS en Geograma. Ingeniero Técnico en Topografía y Máster en Geotecnologías, tiene más de 20 años de experiencia en desarrollo y gestión de proyectos de cartografía y SIG. Es responsable de contenidos en el nodo de acceso de datos de referencia de Copernicus in situ (CORDA), miembro del pool de expertos del marco de mantenimiento e implementación de INSPIRE (MIF-MWIP-8), miembro de los grupos de trabajo de la Infraestructura de Datos Espaciales de España y participante en la actualización de las guías técnicas de metadatos. Ha participado en el estudio INSCope Study of Copernicus & INSPIRE. Es miembro de EuroGI y Presidente de la Asociación Española de Geómetras Expertos.

Kepa López

Consultor Senior GIS y Programador en Geograma. Licenciado en Ciencias Ambientales, Máster GIS (ESRI España - mención especial al mejor expediente de la promoción y Mejor Proyecto en el ámbito del Desarrollo de Aplicaciones GIS - ACUA) y Máster en Desarrollo de Aplicaciones Java. Con 3 años de experiencia como inspector medioambiental y 2 años como gestor técnico de proyectos de I+D+i, finalmente asentado en el sector GIS los últimos 4 años, mejorando continuamente las aptitudes en Big Data y ciencia de datos.

Daniel Navarro

Analista GIS y Data Scientist en Geograma. Doctorando en Geografía, Postgrado en Análisis de la Geoinformación, Máster Oficial en Gestión y Ordenación del Desarrollo Territorial y Local, Licenciado en Antropología Social y Cultural e Ingeniero Técnico en Informática de Sistemas. Experiencia profesional en contextos internacionales en la realización de análisis geoespaciales incluyendo España, Ecuador, Haití, Japón, Perú, JRC (Centro Común de Investigación de la Comisión Europea) y EEA (Agencia Europea de Medio Ambiente). Ha publicado artículos en revistas científicas de impacto y congresos internacionales acerca de procesos socioeconómicos implicados en la vulnerabilidad y exposición al riesgo de desastres y el cambio climático, utilizando diversos modelos espaciales cuantitativos y cualitativos. Dispone de un alto conocimiento en análisis espacial, tecnologías GIS, data science y bases de datos utilizando una amplia gama de tecnologías.