Memoria Trabajo Final RXDET

Andrés Lires Saborido y Ángel Vilariño García

Curso 2025-2026

Contenidos

Filtrado y limpieza de datos

En esta sección se describirán los pasos seguidos en el archivo `preprocess.ipynb` para el filtrado y limpieza de los dos archivos con los que se trabajará en la práctica: *cows_pos.csv* y *fincas.json*.

Archivo *cows_pos.csv*

En primer lugar, se cargó el archivo *cows_pos.csv* en un *DataFrame* de pandas. La columna *time* se convirtió a tipo *datetime* para facilitar su manipulación.

A continuación, se crearon dos *timestamps* que definen el rango temporal de interés: desde el 20 de abril de 2023 hasta el 25 de abril de 2023. Se filtraron las filas del *DataFrame* para conservar únicamente aquellas cuyo valor en la columna *time* se encontraba dentro de este rango.

Posteriormente, se eliminaron 3 filas que tenían valores nulos en la columna *location*. Cabe destacar que la única columna espacial contenía datos de la forma *lat::long*, así que se definió una función para procesar estos valores, la cual realizaba las siguientes tareas:

- Comprobar que todas las filas tuvieran el formato correcto (dos números separados por "::").
- Separar los valores de latitud (*lat*) y longitud (*lon*).
- Comprobar que en el valor original la latitud era el primer elemento. Había casos en los que la longitud (que sabemos que es negativa en Galicia) aparecía en primer lugar en la columna original, por lo que verificamos que el valor *lat* obtenido en el paso anterior fuera positivo, intercambiándo los valores de *lat* y *lon* en caso contrario.
- Corregir ciertas longitudes que comenzaban por -6 (vacas situadas en Asturias), cambiándolas para que comenzaran por -7.

Se aplicó esta función a la columna *location* y se crearon dos nuevas columnas en el *DataFrame*: *latitud* y *longitud*. A partir de ellas, convertimos el *DataFrame* en un *GeoDataFrame* de geopandas, utilizando el sistema de referencia EPSG:4326 y creando una columna de puntos de *geometry*, manejables por PostGIS.

Archivo *fincas.json*

En cuanto al segundo archivo, *fincas.json*, se cargó en un *GeoDataFrame* de geopandas directamente, ya que el archivo estaba en formato GeoJSON. Se comprobó que el sistema de referencia era EPSG:4326 y se dividió su contenido en *Polygons* (uno por finca), ya que el archivo original contenía un *MultiPolygon*. A continuación, empleamos la latitud del centroide de cada finca para eliminar aquellas 3 más al sur, quedándonos con 15 fincas de las 18 originales. En este paso se aplicó una transformación al sistema de referencia EPSG:32629, para evitar posibles problemas.

Carga de datos en la BD

Para la carga de datos de las dos tablas en la base de datos PostGIS, existían varias opciones como usar QGIS, crear el esquema de la base de datos y cargar los datos directamente en DBeaver, o cargarlos mediante comandos en la terminal. Se optó por la solución que a priori parecía más sencilla y rápida: cargar los datos mediante código en *Python*.

Para realizar dicha carga, se utilizó la librería *SQLAlchemy* para gestionar la conexión con la base de datos. En primer lugar, se creó un motor de conexión utilizando la función *create_engine*, proporcionando la URL de conexión a la base de datos. A continuación, se empleó el método *to_postgis* de geopandas para cargar los *GeoDataFrames* previamente creados en la base de datos. Este método permite especificar el nombre de la tabla destino, el motor de conexión y la opción de reemplazar la tabla existente. A continuación, se realizó una comprobación empleando la función *read_postgis* de geopandas para asegurarse de que el número de filas y columnas en las tablas de la base de datos coincidía con el número de filas y columnas en los *GeoDataFrames* originales.

Este proceso se llevó a cabo también en el archivo *preprocess.ipynb*, justo después de la limpieza y filtrado de los datos.