Project on processing and querying for the fishing company use case
Part 2: Processing and Querying Data

Iva Bokšić, Brice Écotière

Part 1 : Data cleaning

The temperature and chlorophyll datasets contain a high proportion of missing values, with 47.81% and 54.11% respectively. Three approaches were used to impute the missing values: (1) averaging the readings from the previous and next 4 days for each missing value, (2) averaging the readings from the four nearest sensors by longitude for a fixed date and latitude, and (3) averaging the readings from the four nearest sensors by latitude for a fixed date and longitude. The AIS data was downsampled using the resample() method with a time interval of 20 minutes, resulting in a reduction in the number of rows from 81,907 to 52,576.

Part 2 : Data normalization/unification

The objective of this part was to normalize the data of ais and fishing data sets. To do this, we created a panda dataframe that contained a grid of square polygons. Using the Geopandas library, we performed a spatial join between our initial dataframe and the polygon dataframe (grid). In this way, we were able to process the data for each coordinate point per polygon. For AIS data, we averaged the velocities per polygon, for fishing we averaged the temperatures and the sum of the fishing times and the mass of fish caught. Furthermore, each polygon (square) has a GridID as (i,j) with i and j positive integers.

Part 3 : Data integration

In this part, we wanted to bring together the fishing, temperature and chlorophyll data sets. The difficulty lay in the column(s) that would allow us to perform the merge() function of panda. Thus, we assumed that the date and geolocation of the vessels would be sufficient data to join our databases properly. First, we had to find a way to match all the geolocation data. Thus, we created a function "function_id" that allows normalized longitude and latitude data by defining an identifier in the same way as for GridID in the previous part. Thus, the temperature and chlorophyll databases are now provided with the same GridID. It is thus possible to join them thanks to this column. Finally, it was necessary to ensure that the dates were in the same format between the three data sets. With these two columns correctly constituted ['Day','GridID'], we were able to join the three datasets in order to compare and cross-reference their data.

Part 4 : Querying

After plotting the trajectories of the fishing and AIS datasets before preprocessing, we observed three distinct fishing areas, as well as some fishing data that was not recorded in the AIS data.In March of 2020, the boat named Korbin traveled the furthest distance compared to other boats in the dataset. Boat Rodney caught the highest amount of fish per month in April 2020, with a total of 969,890 kilograms. However, we found negligible correlation between the quantity of fish caught and the temperature or chlorophyll measurements.