
Data Analyst Assignment

Background

At Windward, we often work on datasets related to the maritime domain. In this exercise, you will get a glimpse of Windward's unique databases. While at sea, vessels are required to transmit frequently their location, along with other data such as speed, direction and identifying numbers (such as IMO and MMSI). These transmissions are gathered and translated over time to vessel entities.

Vessels are commercial creatures that are used to drive global trade. Each vessel type (class) is used for different commodities and purposes, and demonstrates specific behavior at sea when sailing legitimately. Windward leverages this domain knowledge to create behavioral models that track vessels and flag anomalies from the expected sailing patterns such as transmission gaps, anchoring, fishing and meetings.

In this home assignment you will try to identify **GPS Jamming** and how it can affect our data - the following blog explains this phenomenon more thoroughly:

<https://windward.ai/glossary/what-is-gps-jamming/>

In part A you are required to differentiate between actual meetings and data noise induced by GPS jamming, while in part B you will focus on vessel tracks of specific vessels, filtering out GPS jamming **transmissions** within these tracks.

Please make sure to dedicate enough time for both parts.

Part A - False Meetings in the Red Sea:

A meeting between vessels can be defined as two vessels that are co-located (spatial proximity) at low speed for a sustained period, typically to transfer cargo. Key determinants are distance between vessels at the beginning and end of the activity, overlap duration, classification of meeting - commodity, fueling (bunkering), etc., and location context (anchorage, high seas, offshore terminal zones).

It can be quite challenging to distinguish between legitimate meetings, an intentional illicit commodity transfer, and data noise.

Task:

In your task you are given a dataset containing meetings between vessels in the Red Sea area during the past year. **Your task is to identify which of these meetings were actually data noise caused by GPS Jamming.**

Please keep the time constraints in mind and avoid over-complicating the solution.

Deliverables:

1. A CSV file with additional information about each meeting whether it is a jamming induced meeting or not.
2. Presentation (no more than 3 slides) with research steps, assumptions and findings. a. Including a hidden slide with a link to chat or prompts if used.
3. IPYNB file (Jupyter notebook / Google Colab) including the code you used for the analysis (.py file is also acceptable)

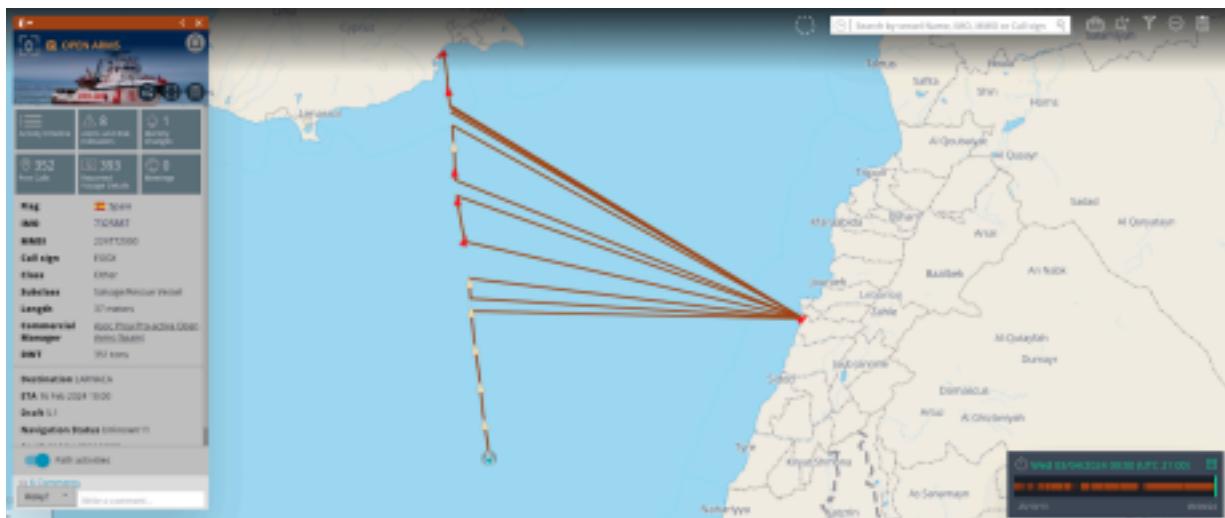
You will be measured on:

1. Conciseness - focus on important findings only and stay within the 3 slides limit.
2. Presentation skills - present information that is visually appealing, accurate, clear, and tell a story.
3. Methodology - make sure your metrics and calculations are valid, reasonable and explainable.

Try to find domain / statistical justifications for your decisions.

Part B - “menifot”

GPS Jamming can also impair vessel track reliability. Following the Israel-Gaza war, vessels approaching Israel often transmit false locations of Beirut Airport or Cairo while sailing. The image below outlines the path of a vessel heading from Israel to Cyprus on April 3rd, 2024 (marked in brown). The path is based on real transmissions received from this vessel (including occasional false transmissions of the Beirut Airport), where consecutive transmissions are connected in a line.



Task:

Using the two examples provided in part_b folder (menifot1.csv, menifot2.csv), please write a Python function that receives a sequence of location points, filters out the faulty data, and returns only the clean track.

Important guidelines:

1. Indicate what assumptions the function was based on
2. The filtering function should not use the is_menifot field. This field is a label we provide to help you understand what the false locations are, out of the entire data set.
3. The filtering function should be as general as possible. For example, will your solution still work if starting from next week the false transmissions would appear in the Red Sea?

Deliverables:

1. IPYNB file (Jupyter notebook / Google colab) including the code you used for the analysis
2. Prompts or link to the chat with LLM model itself can be added in a separate text file or attached as markdown in the python code

You are allowed to use AI tools (e.g., GPT) to assist in the assignment. Please explain the reasoning behind your analysis and any assumptions you make. Please add to the presentation in part A a hidden slide including a link to the chat or the prompts you used.

Feel free to reach out if something is not clear - we are available :)

Ro'i - Roi.makov@wnwd.com

Good luck!