Approach to Detecting Maritime "Loitering" Events

By: José Luis Jarpa Parra

josejarpaparra@gmail.com

February 2024

Requirements:

Background:

You are tasked with identifying vessels engaged in suspicious behavior. As a data scientist, your role is to devise a strategy to identify vessels that are "loitering" in an area known for smuggling.

Objective:

Present your approach to identifying vessels that are loitering in the designated zone. Focus on the methodology you would employ, including data collection, preprocessing, and analysis techniques.

Tasks:

Data Collection:

You will have a database of ais data covering the designated zone. AIS data will contain information such as lat/long, course over ground, speed, and heading.

Data Preprocessing:

Briefly explain any preprocessing steps you would undertake to clean and prepare the AIS data for analysis, if any.

Analysis Techniques:

Present the analytical methods or algorithms you would use to analyze vessel movements and detect loitering activity.

Discuss any statistical or machine learning approaches you would employ for behavior detection or pattern recognition.

Presentation:

Prepare a concise presentation outlining your approach to monitor for loitering vessels. Focus on clarity, brevity, and relevance, and be prepared to answer any questions or provide additional details during the presentation.

Introduction:

In the realm of maritime security, the detection of vessels engaged in suspicious behavior, such as loitering, plays a crucial role in maintaining safety and preventing illegal activities. The manual examination of vessel behavior to identify anomalous ships is labor-intensive, especially for nationwide surveillance. To address this, researchers have proposed computational methods. However, many of these methods are region-dependent and **require a normal behavior profile to detect anomalies**. As a data scientist, my task is to devise a strategy to identify vessels that are loitering in a specific area known for smuggling. This document outlines the methodology I employed, focusing on detailing data collection, preprocessing, and analysis techniques.

It is important to note that this approach is based on studies such as the one conducted by Wayan Mahardhika Wijaya and Yasuhiro Nakamura, titled "Loitering behavior detection by spatiotemporal characteristics quantification based on the dynamic features of Automatic identification System (AIS) messages." They propose **a region-independent method to automatically detect loitering without training normal instances and produces a ranked list of loitering vessels to facilitate further anomaly investigation**.

Next, we will detail each of these points in the following sections, providing a comprehensive strategy for detecting loitering vessels in the designated area.

I generated the complete code necessary to test this method. All files including Code, Dataset and validation videos are available at: https://github.com/jjarpa/LoiteringDetector

Initial Concepts:

This experiment proposes a method to automatically detect loitering behavior without the need for training normal instances. The method is based on the definition of spatiotemporal characteristics of loitering and quantifies them using AIS message data.

In this approach, loitering behavior is adopted as a type of anomalous behavior in the context of vessel movement. The voyage of a ship is a deliberately planned event that should assure safety and efficient navigation. Thus, in normal circumstances, any ship shall not take any maneuver that endangers people's life at sea, they shall navigate as efficiently as possible in a safe manner, which means that they are to take the shortest and fastest route whenever it is safe to do so, and they shall not conduct any activity that causes pollution or damage to the marine environment. **Any ships that behave oppositely are to be considered anomalous**. This experiment recognizes loitering as a type of anomalous behavior, especially for ships of types cargo and tanker, and develops a loitering detection method that overcomes the remaining issues of the known methods.

Considered what was mentioned in above paragraph, is expected that only few vessels have anomalous behavior, if compared with the total of the vessels in a specific time. The method then formulates parameters to determine a loitering (anomalous) trajectory and calculates an anomaly (loitering) score for each trajectory. The Isolation Forest algorithm is used to establish a threshold and rank the loitering vessels, with geographic visualization for intuitive evaluation.

In order to quantify the spatiotemporal characteristics of loitering behavior, this experiment utilizes the dynamic information of AIS messages: course over ground (COG), speed over ground (SOG), position, heading, and timestamp (basedatetime). COG is the actual direction of a vessel's movement between two points and SOG is the actual speed, with respect to the surface of the earth. COG is expressed in 360° angular direction, where 0° ≤ COG < 360°. SOG is in knots, which is nautical miles (Nm) per hour. The heading is the direction of the ship's bow as indicated by the compass of the ship, which is also expressed in 360° angular direction within the range of 0° ≤ Heading < 360°. Ship position is expressed in longitude and latitude coordinates, thus ship position at a timestamp can be specified as point(x, y) where x is the longitude and y is the latitude. AIS devices broadcast this dynamic information as a message timestamped in UTC. The broadcasts are transmitted discretely on a certain time interval proportional to the speed of the ship.

Stem from the definition, the spatiotemporal characteristics of loitering can be elaborated as follow: **(1) movement of frequent course change, with a certain speed, within a certain spatial range, (2) movement of frequent course change within traversed geodetic distance, (3) might demonstrate frequent extreme turning, and (4) extreme turning produces a significant discrepancy between the course over ground (COG) and the heading of the ship.** These spatiotemporal characteristics of the loitering behavior correspond to the dynamic information of the AIS messages. Course change is the absolute difference between COG at two consecutive points:

$$\boxed{\Delta C_k = \left|COG_{t_k} - COG_{t_{(k-1)}}\right|}$$

Speed is the speed over ground (SOG). The spatial range is the bounding box that encloses the ship's trajectory. Geodesic distance is the shortest distance between the starting point and the ending point of the trajectory. The discrepancy between COG and heading is the absolute value of COG–Heading:

$$\Delta H_k = \left|COG_{t_k} - Heading_{t_k}\right|$$

The frequent change of courses in a loitering trajectory may include minor ones to the relatively large change of courses that occur in extreme turnings. The rate of course change can be described by comparing ΔC with the maximum course change (180°). Considering the area of the bounding box B enclosing the loitering trajectory, the speed S of the ship, and the rate of course change, the score of loitering F(c) can be expressed as:

$$F\left(c\right) = \frac{\sum_{k=1}^{n} \Delta C_k \times \sum_{k=0}^{n} S_k}{180 \times B}$$

Equation (1)

The unit for B and S is Nm2 and knots respectively. Here, the score of loitering F(c) is proportional to the rate of course change and inversely proportional to the area of the enclosing bounding box.

Another approach to describe loitering behavior is to additionally consider the discrepancy between COG and heading (ΔH) and the geodesic distance (G) between the starting and ending points of the trajectory. In other words, it is to take into account all of the loitering behavior spatiotemporal characteristics as expressed by:

$$F\left(c, h, d\right) = \frac{\sum_{k=1}^{n} \Delta C_k \times \sum_{k=0}^{n} \Delta H_k \times \sum_{k=0}^{n} S_k}{B \times G}$$

Equation (2)

The unit for G is Nautical miles (Nm).

Then, the complete process to obtain the loitering scores is summarized in the next steps:

1. Order the AIS dataset by MMSI and TimeStamp in order to easy the next calculations/operations.
2. Calculate ΔC and ΔH of the equations (1) and (2): We will import all relevant AIS data ('MMSI', 'LAT', 'LON', 'COG', 'SOG' and 'Heading') in a pandas DF and generate 2 new columns named 'cog_change' and 'delta_cog_heading' (ΔC and ΔH from Equations

(1) and (2) respectively. We will save this new DF in a CSV file for further analysis and backup.

3. Iterate over all MMSI's. First of all, is necessary to check if the time amplitude of the data we have is bigger than the time window of analysis (24hrs). If not, we can skip this MMSI (which will not apport information for our algorithm). In case the time amplitude of the data is bigger than 24hrs, we can continue to execute the process. Also is needed to consider anchored vessels. The filter here is simple: just calculating the Area of Bounding Box of the 'trajectory' of an anchored vessel, we will see that this area is relatively small if compared with a normal trajectory. So for small areas we will discard the vessel from the analisis.
4. Yet iterating over MMSI's, we will slide the time window across all the data of the MMSI. You can imagine as if each time window data is a frame of a movie the complete movie is the complete trajectory of the vessel. So, we will move time by time (row by row) on the AIS dataframe for this specific MMSI. At each time we will calculate B (Total Area of the Time Window in [Nm]2) and G (Geodesic distance between all points of MMSI dataframe, in Nm) from equations (1) and (2), like showed in next figure:

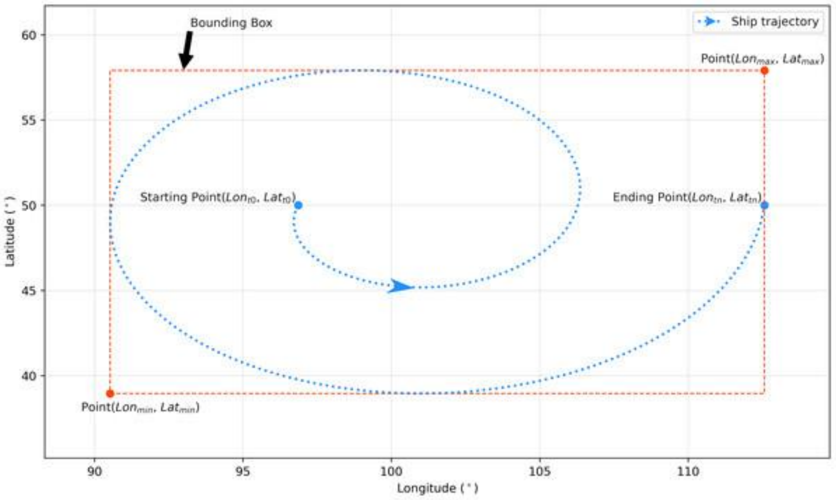


Figure (1)

Note that B is the area inside the red bounding box of the ship trajectory. And G is the length of the blue line (calculated as geodesical distance in Nm).

5. Yet iterating over the MMSI's, once we have all the members: ΔC, ΔH, B and G, we can calculate (in each step of the slide window) the value of F(c) and F (c,h,d) of equations (1) and (2). We will conclude sliding across all AIS data of the MMS, so we will end the loop with a set of F(c) and F(c,h,d) –one per slide step--. We will choose the biggest of these parameters and store them in a dictionary where the key will be the MMSI and the values will be F(c) and F(c,h,d).
6. Once we finish the iteration over all MMSI's we will obtain a dictionary with all the vessels and their corresponding F(c) and F(c,h,d) parameters.
7. In order to estimate the degree of anomaly (Loitering) we apply the Isolation Forest algorithm to the set of parameters F(c) and F(c,h,d). The Isolation Forest algorithm provides the capability of detecting anomalies without the requirement to build a profile of normal instances. The idea of the algorithm is that anomalies are 'few and different'. Few means anomalies are minority in number of instances, and different means anomalies have values that are significantly different from those of normal

ones. The idea corresponds to the reality that vessels of types cargo and tanker do not normally loiter. Loitering tanker or cargo ships should be rare, and loitering trajectories is obviously different from the normal voyage routes that follow the guidelines of voyage planning.

8. Also as a way to enhance the method, we will combine the F(c) and F(c,h,d) parameters using the Entropy Weight Method (EWM): The anomaly scores returned by the Isolation Forest algorithm of both parameters (in the previous case) are integrated to return only one anomaly score. Let sf1 and sf2 be the scores of anomaly of the parameters F(c) and F(c, h, d) respectively returned by the Isolation Forest algorithm, and wf1 be the weight of F(c) while wf2 is of F(c, h, d), then the integrated anomaly score I(f1, f2) of the parameters F(c) and F(c, h, d) is obtained with the formula:

$$\boxed{I\left(f_1, f_2\right) = \frac{s_{f_1} \times w_{f_1} + s_{f_2} \times w_{f_2}}{w_{f_1} + w_{f_2}}}$$

Equation (3)

In this case, the Entropy Weight Method (EWM) is adopted to set the weight of each parameter. A parameter with higher entropy, which means a higher differentiation degree, is given the heavier weight.

9. Once we have all the anomaly scores of F(c), F(c,h,d) and I(f1,f2) we store all them in a dictionary and save them in a csv file for further analysis. Using this scores we can rank the most anomalous MSSI's and review them with the support of an human expert. The expert will label them and we can obtain the accuracy degree of this method.
10. Optionally, or as a final enhancement of this method is possible use the subset of labeled (by the human expert) trajectories and apply it to a Convolutional NN, so we can train this network to emulate the human expert knowledge.

For experiment with this method, I constructed a functional python code using Spyder 5.4.3 under Anaconda, with this environment/packages:

1) Python 3.11.7: For general environment.
2) Pandas 2.2.0: For data manipulation and storage.
3) Cartopy 0.22.0: For geographic display and charts.
4) Geopy 2.4.1: For Geodesic Distance Calculation.
5) Matplotlib 3.8.3: For geographic display and charts.
6) Python-date-util 2.8.2 : For timestamp/Date conversion.
7) Numpy 1.26.4: For data manipulation and calculations
8) Scikit-learn 1.4.1: For isolation Forest algorithm.

See below the workflow of this loitering detection method:

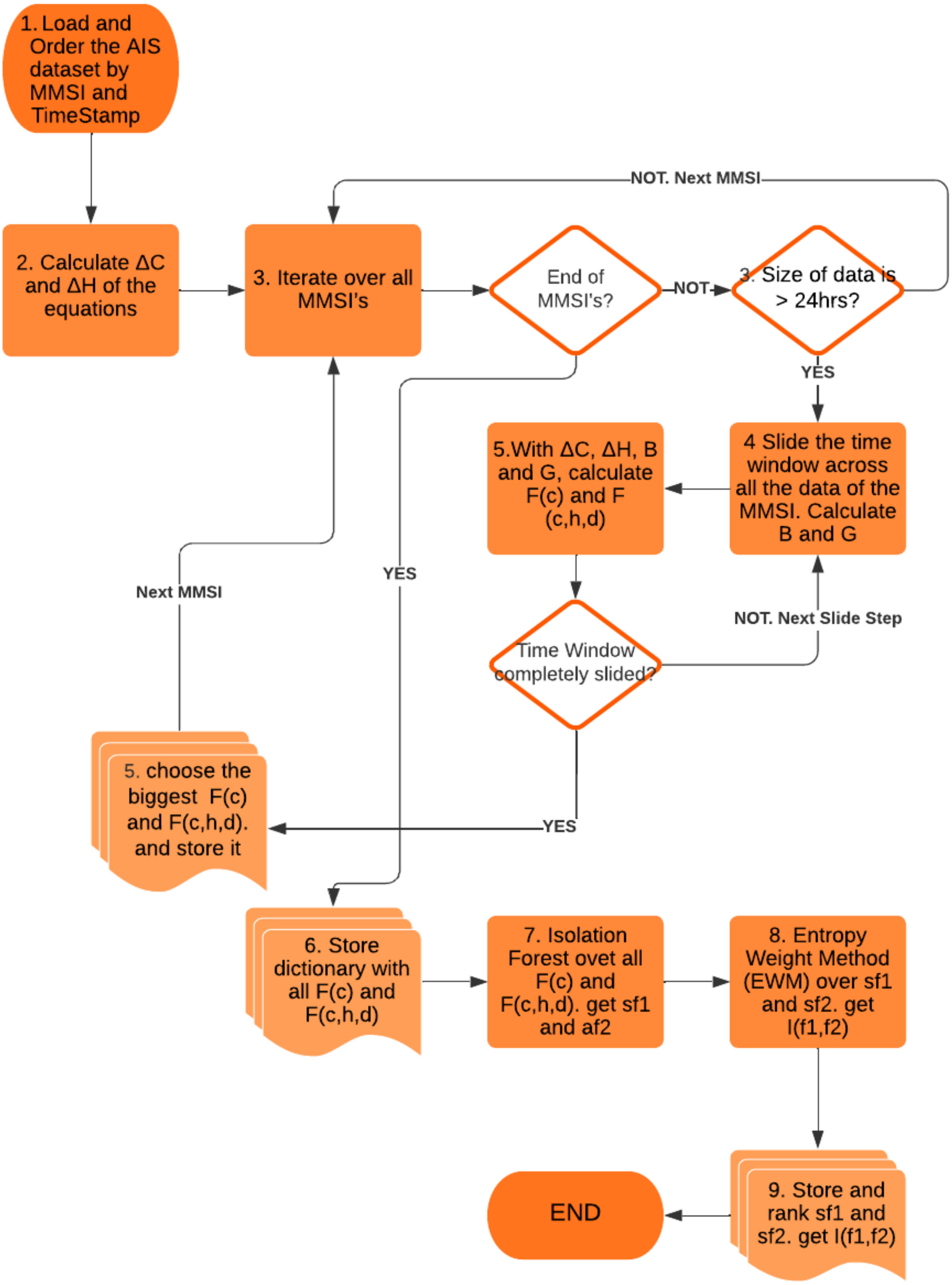

1. Load and Order the AIS dataset by MMSI and TimeStamp
2. Calculate ΔC and ΔH of the equations
3. Iterate over all MMSI's
End of MMSI's?
NOT
3. Size of data is > 24hrs?
NOT. Next MMSI
YES
4 Slide the time window across all the data of the MMSI. Calculate B and G
5.With ΔC, ΔH, B and G, calculate F(c) and F (c,h,d)
Time Window completely slided?
NOT. Next Slide Step
YES
5. choose the biggest F(c) and F(c,h,d). and store it
Next MMSI
YES
6. Store dictionary with all F(c) and F(c,h,d)
7. Isolation Forest ovet all F(c) and F(c,h,d). get sf1 and af2
8. Entropy Weight Method (EWM) over sf1 and sf2. get I(f1,f2)
9. Store and rank sf1 and sf2. get I(f1,f2)
END

Figure (2)

Data Collection & Preprocessing:

For this section we will use the same workflow detailed in **Figure (2)**.

1) Load and Order the AIS dataset by MMSI and TimeStamp:

The data of interest for this experiment is TimeStamp (BaseDateTime), MMSI, LAT, LON, COG, SOG and Heading.

Data received from AIS must be filtered in order to avoid the impact on data processing infrastructure (due to High volume of information) and also these filters must be aligned with company or customer interest. Depending on the strategy, AIS data can be directly filtered or can be downloaded from a Database (Real Time or Near Real Time, respectively). In any case, must be necessary, at least, consider the next parameters:

TimeStamp, MMSI, Type of mobile, Navigational status, Ship Type and Cargo type. Other parameters can be included, depending on the interest of the company and/or customer. Below is an example of SQL query that will filter the data on a typical AIS dataset. A brief explanation of each Filter field is also giving:

```
SELECT *

FROM "aisdk_2024_02"

WHERE "Type of mobile" IN ("Class A", "Class B")

AND "Navigational status" IN ("Under way using engine", "Under way sailing",
"Reserved for future amendment [HSC]")

AND "Ship type" IN ("Cargo", "Tanker");
```

Type of Mobile: They use the same frequencies and can access the TDMA slots. However, some special messages will be reserved for certain types of AIS and different channel access schemes may be used.

Class A (IEC 61993-2)

Shipborne mobile equipment intended to meet the performance standards and carriage requirements adopted by IMO. Transmission power normally is 12.5W. Class A stations report their position (message 1/2/3) autonomously every 2-10 seconds dependent on the vessel's speed and/or course changes (every three minutes when at anchor or moored). The vessel's static and voyage related information (message 5) is transmitted every 6 minutes. Class A stations are also capable of text messaging safety related information (message 12/14) and AIS Application Specific Messages (message 6/8/25/26), such as meteorological and hydrological data, electronic Broadcast Notice to Mariners, and other marine safety information (see IMO Safety of Navigation Circular 289, Guidance on the use of AIS application-specific messages (ASM) or the IALA Application Specific Message Collection). In professional shipping the AIS should be switched on at all times according to most regulations, also when the vessel is stationary or loading.

One of the most successful class A transponders fulfilling Inland and Solas requirements is our Nauticast A2 as it fulfills all international standards and is operating flawlessly in maritime, coastal and inland waters.

Class B (IEC 62287-1 & IEC 62287-2)

Shipborne mobile equipment which is interoperable with all other AIS stations but follows simpler performance standards. The transmission power of 2W for Class B devices is significantly lower than for Class A or Base stations, consequently the position of the VHF antenna is by far the most decisive factor for a good transmission range. Similar to Class A stations they report their position (message 18) every three minutes when at anchor or moored. But it is reported less often than Class A when moving. Likewise, they report the vessel's static data (message 24A and 24B) every 6 minutes, but not any voyage related information. They can receive safety related text and application specific messages, but cannot transmit them.

There are two types of Class B AIS, those using Carrier Sense Time Division Multiple Access (CSTDMA) technology and those using Self Organizing Time Division Multiple Access Technology (SOTDMA). Class B "SO" uses a more elaborated time slot reservation methodology and needs more processor power for slot calculations. Class B "CS" is generally less expensive. In practice the differences are rather negligible. The Nauticast B2 offers class A like reception of 50 km and higher and will be received at distances of up to 20 km under good conditions. With its reliability, a variety of communication options and a powerful management software it is the choice of the ambitious and prudent yachtsman.

Search and Rescue Aircraft (partly IEC 61993-2)

Aircraft mobile equipment, normally reporting its position every ten seconds (message 9). Static data can be transmitted using messages 5, 24A and 24B.

AIS AtoN station (Aid To Navigation) (IEC 62320-2)

An AtoN station may be fixed (shore) or floating (maritime or river) providing location and status of an Aid to Navigation (AtoN). Normally it reports (message 21) every three minutes, but here the standard offers a lot of flexibility for adjustment to individual requirements. These stations may also broadcast Application Specific Messages (message 12/14). The station may also transmit messages 21 for virtual or synthetic AtoNs.

AIS search and rescue transmitter (SART) (IEC 61097-14)

Mobile equipment to assist homing to itself (i.e. life boats, life raft). An AIS SART transmits a text broadcast (message 14) of either 'SART TEST' or 'ACTIVE SART'. When active the unit also transmits a position message (message 1 with a 'Navigation Status' = 14) in a burst of 8 messages once per minute.

AIS SARTs are also used in maritime survivor locating devices (MSLD) or man overboard (MOB) devices, as specified in RTCM 11901.1, Standard for Maritime Survivor Locating Devices as well as for AIS locating beacons on 406 MHz EPIRBs. Standard AIS SARTs can be identified by MMSI's beginning with the numbers "970", AIS maritime survivor locating devices or MOBs with MMSIs beginning with "972", and AIS EPIRB with MMSIs beginning with "974". All categories of AIS SARTs will be displayed on IMO-mandated shipboard navigation displays.

AIS base station (IEC 62320-1)

Shore-based station providing AIS channel management, text messages, time synchronization, meteorological or hydrological information, navigation information, or position of other vessels. Normally reports (message 4) every ten seconds. These stations are normally operated by the responsible authorities for the area. With the spread of applications like Marinetraffic.com, vesselfinder.com, shipfinder.com a lot of privately run "base stations" came into operation. However, these are more like large "receiver networks" with huge databases and will not transmit or assign and control other AIS stations as base stations run by the authorities would do.

For this experiment **only Class A and Class B** type of devices must be considered.

Navigational Status: The navigational status takes different values, Below we'll find the possible values and their significance with meaning to each status:

0 = underway using engine

The vessel is underway using the engine when :

It is not aground

It is not at anchor

It was not attached to a dock, the shore, or any other stationary object.

1 = at anchor

The vessel is considered at anchor when it is in one position held by an anchor so it is prevented from drifting away to another position.

2 = not under command

The "under command" means that the vessel unable to manoeuvre due to some circumstance.

3 = restricted maneuverability

The term "restricted maneuverability" means that the vessel is unable to keep out of the way of another vessel.

4 = constrained by her draught

The vessel that works with power, is restricted from drifting away from the course due to the draught power.

5 = moored

Limiting the vessel movements by several lines or cables while securing it at a pier.

6 = aground

The vessel that's aground onto a strand or underwater.

7 = engaged in fishing

This term refers to that the vessel used for fishing with nets, trawls...

8 = underway sailing

We can say that the vessel is underway sailing when:

It is not aground

It was not attached to any object such as a shore or a duck.

It is not at anchor

9 = reserved for future amendment [HSC] of navigational status for ships carrying dangerous goods (DG), harmful substances(HS), or IMO hazard or pollutant category C, high-speed craft (HSC). Those types of vessels are reserved specifically for the ships carrying dangerous good that can be harmful.

10 = reserved for future amendment of navigational status for ships carrying dangerous goods (DG), harmful substances (HS) or marine pollutants (MP), or IMO hazard or pollutant category A, wing in the ground (WIG). The type of vessels are carrying dangerous good, it can be also marine pollutants.

11 = power-driven vessel towing astern.

A power-driven vessel usually is when the lights are in a vertical line. The towing light is in a vertical line instead if stern light. In addition, usually the description applies when the length exceeds 200 meters.

12 = power-driven vessel pushing ahead or towing alongside. A power-driven vessel when pushing ahead or towing alongside, usually consist: two masthead lights in a vertical line, sidelights and two towing lights in a vertical line.

13 = reserved for future use. For the vessels whose function is not currently defined, but it will be defined after some future enhancement. It should not be used until the enhancement has been further defined.

14 =AIS-SART Active (Search and Rescue Transmitter), AIS-MOB (Man Overboard), AIS-EPIRB (Emergency Position Indicating Radio Beacon): An AIS-SART (Automatic Identification System – Search And Rescue Transmitter) is a device that sends an emergency message with the position based on the Automatic Identification System (AIS) protocol.

15 = undefined = default (also used by AIS-SART, MOB-AIS, and EPIRB-AIS under test).

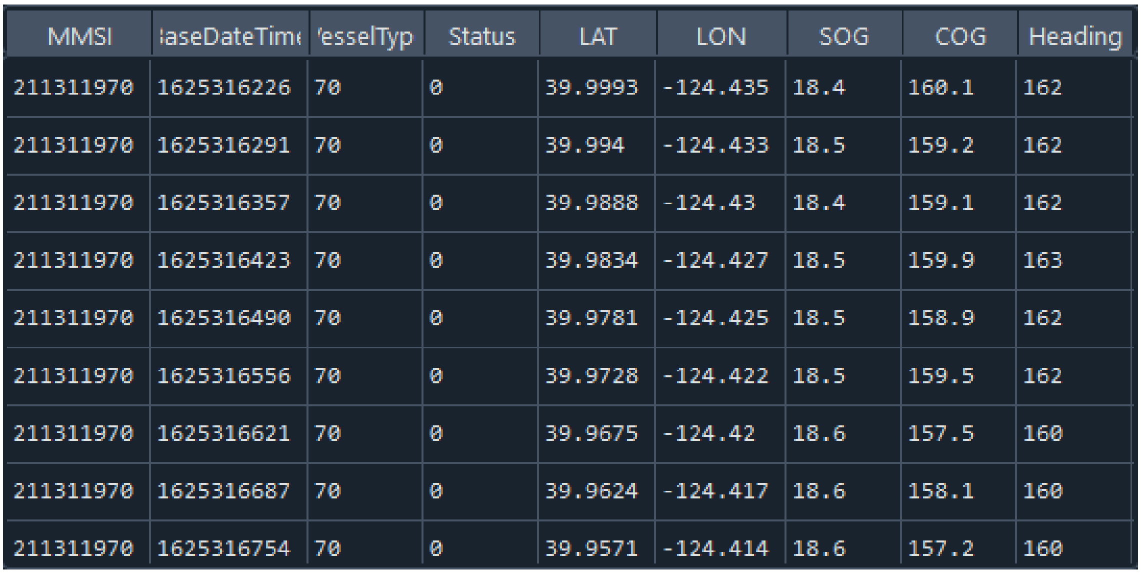
For this experiment **Under way using engine, Under way sailing, Reserved for future amendment [HSC]** Navigational Status must be considered.

Ship Type: A vessel's type can be deducted using the information contained in the AIS-transmitted messages that she is emitting. The vessel's crew or the accountable officer are responsible for correctly entering this piece of information to the vessel's AIS transponder. AIS SHIPTYPE usually consists of two digits. The first digit represents the general category of the subject vessel:

1 = Reserved
2 = Wing In Ground
3 = Special Category
4 = High-Speed Craft
5 = Special Category
6 = Passenger
7 = Cargo
8 = Tanker
9 = Other

In certain cases (e.g. Cargo Vessels, Tankers), the second digit provides additional information regarding the subject vessel's type of cargo, according to IMO.
1 = Category X
2 = Category Y
3 = Category Z
4 = Other Substances (OS)
For this experiment **Cargo and Tanker"** Ship type must be considered. **Cargo type** field will not be used in this experiment.

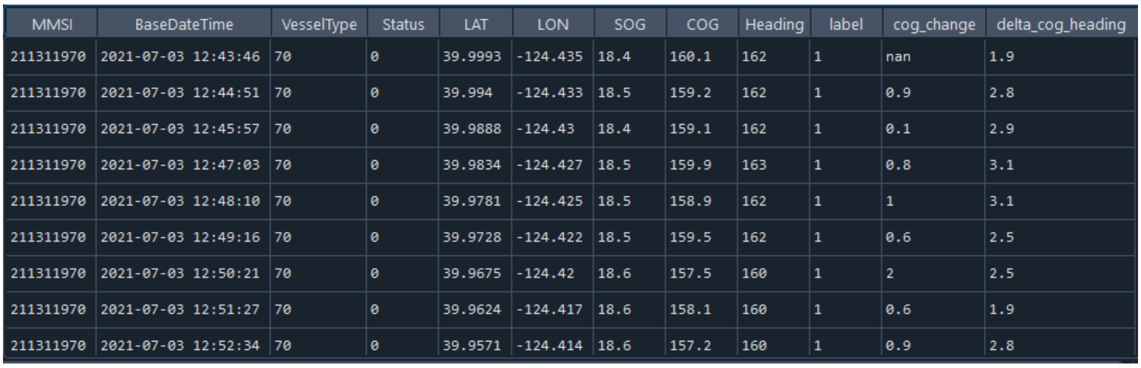
Once data is selected and ordered from database, it will be loaded into a python pandas's DF:



MMSI	aseDateTim	esselTyp	Status	LAT	LON	SOG	COG	Heading
211311970	1625316226	70	0	39.9993	-124.435	18.4	160.1	162
211311970	1625316291	70	0	39.994	-124.433	18.5	159.2	162
211311970	1625316357	70	0	39.9888	-124.43	18.4	159.1	162
211311970	1625316423	70	0	39.9834	-124.427	18.5	159.9	163
211311970	1625316490	70	0	39.9781	-124.425	18.5	158.9	162
211311970	1625316556	70	0	39.9728	-124.422	18.5	159.5	162
211311970	1625316621	70	0	39.9675	-124.42	18.6	157.5	160
211311970	1625316687	70	0	39.9624	-124.417	18.6	158.1	160
211311970	1625316754	70	0	39.9571	-124.414	18.6	157.2	160

2) Calculate ΔC and ΔH of the equations:

This step is simple, and we only need to create a function which generate the 2 columns (: ΔC and ΔH):



MMSI	BaseDateTime	VesselType	Status	LAT	LON	SOG	COG	Heading	label	cog_change	delta_cog_heading
211311970	2021-07-03 12:43:46	70	0	39.9993	-124.435	18.4	160.1	162	1	nan	1.9
211311970	2021-07-03 12:44:51	70	0	39.994	-124.433	18.5	159.2	162	1	0.9	2.8
211311970	2021-07-03 12:45:57	70	0	39.9888	-124.43	18.4	159.1	162	1	0.1	2.9
211311970	2021-07-03 12:47:03	70	0	39.9834	-124.427	18.5	159.9	163	1	0.8	3.1
211311970	2021-07-03 12:48:10	70	0	39.9781	-124.425	18.5	158.9	162	1	1	3.1
211311970	2021-07-03 12:49:16	70	0	39.9728	-124.422	18.5	159.5	162	1	0.6	2.5
211311970	2021-07-03 12:50:21	70	0	39.9675	-124.42	18.6	157.5	160	1	2	2.5
211311970	2021-07-03 12:51:27	70	0	39.9624	-124.417	18.6	158.1	160	1	0.6	1.9
211311970	2021-07-03 12:52:34	70	0	39.9571	-124.414	18.6	157.2	160	1	0.9	2.8

'cog_change' and 'delta_cog_heading' are respectively ΔC and ΔH.

The next steps are happening inside the main loop of the program:

3) **Iterate over all MMSI's:**
4) **Slide the time window across all the data of the MMSI. Calculate B and G**
5) **With ΔC, ΔH, B and G, calculate F(c) and F (c,h,d). Choose the biggest F(c) and F(c,h,d) and store it.**
6) **Store dictionary with all F(c) and F(c,h,d)**

After steps 3 to 6, we obtain 2 dictionaries: **max_FC{MMSI:F(c))** and **max_Fchd(MMSI: F(c,h,d)**, like this:

max_Fc:
211327410: 564.1014812973774
211335760: 361.66151317743146
215071000: 3.9544335376954116
219031000: 0
220593000: 1374.2619555112494

max_Fchd:
211327410: 417076085.79720855
211335760: 155669829.10367316
215071000: 6699.901962041226
219031000: 0
220593000: 873617051.8396381

Note the variability of these scores. This will be fixed with isolation forest algorithm.

7) Isolation Forest ovet all F(c) and F(c,h,d). get sf1 and sf2

For the estimation of anomaly score, I used python sckit-learn package which implements this function.

```
# calculate the anomaly score of max_Fc & max_Fchd dictionaries
anomaly_scores_max_Fc = calculate_anomaly_scores(max_Fc)
anomaly_scores_max_Fchd = calculate_anomaly_scores(max_Fchd)
```

here, sf1 = anomaly_scores_max_Fc and sf2 = anomaly_scores_max_Fchd

```
anomaly_scores_max_Fc:
211327410: -0.08175186734594309
211335760: -0.04898182812653673
215071000: 0.16609184498976992
219031000: 0.16065462762119842
220593000: 0.01828600914472539
anomaly_scores_max_Fchd:
211327410: 0.03170486143387358
211335760: 0.03273140423366433
215071000: 0.1911303619677942
219031000: 0.1919257700368605
220593000: -0.0079442886256601
```

8) Entropy Weight Method (EWM) over sf1 and sf2. get I(f1,f2)

Here we combine the scores **anomaly_scores_max_Fc and nomaly_scores_max_Fchd**, to obtain a weighted one. I created another function to do this:

```
EWM_anomaly_scores_max_Fc_max_Fchd =
combine_columns(anomaly_scores_max_Fc, anomaly_scores_max_Fchd)
```

The result is:

```
EWM_anomaly_scores_max_Fc_max_Fchd:
211327410: -0.025023502956034754
211335760: -0.0081252119464362
215071000: 0.17861110347878206
219031000: 0.17629019882902947
220593000: 0.005170860259532645
```

9) Store and rank sf1 and sf2. get I(f1,f2):

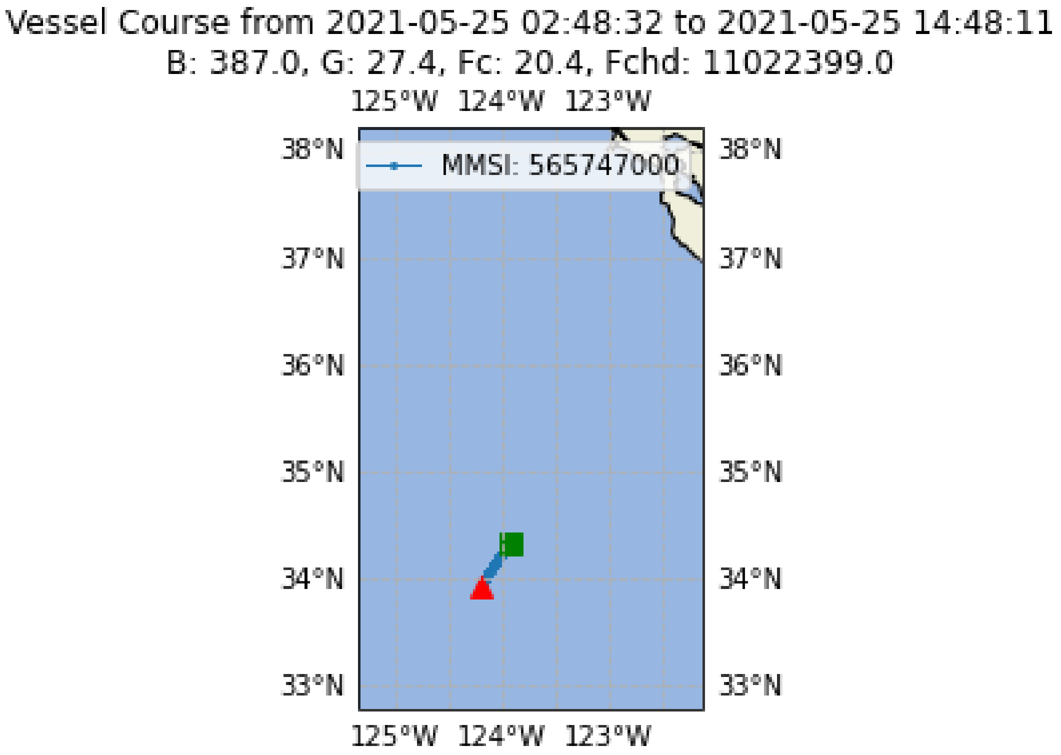
Finally, we group all this values in a DF dictionary where the MMSI is the key and save the results in a cvs file for further analysis. The result is this:

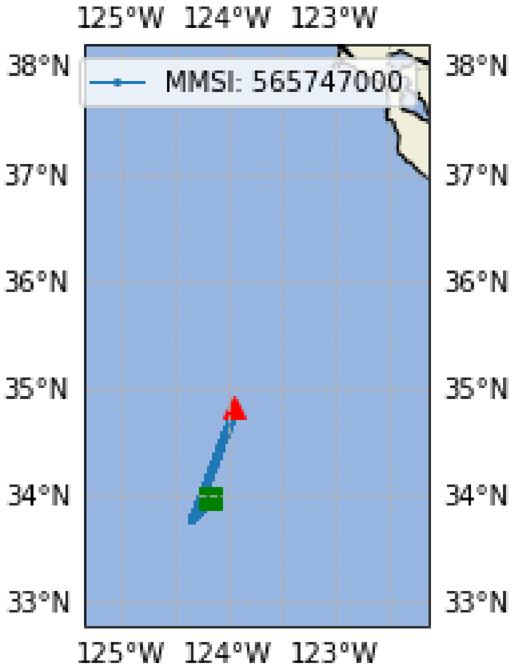
```
5 most negative rows based on 'EWM_Anomaly_Score_Max_Fc_Max_Fchd':
     MMSI  ...  EWM_Anomaly_Score_Max_Fc_Max_Fchd
46  565747000  ...                  -0.236336
32  373932000  ...                  -0.231074
23  357051000  ...                  -0.203917
```

```
9  303352000 ...              -0.157054
35  431496000 ...              -0.153165

[5 rows x 6 columns]

5 most positive rows based on 'EWM_Anomaly_Score_Max_Fc_Max_Fchd':
     MMSI  ... EWM_Anomaly_Score_Max_Fc_Max_Fchd
13  319200700 ...                0.179389
14  319819000 ...                0.179462
52  636019627 ...                0.179715
33  374077000 ...                0.179817
12  311000421 ...                0.179875
```

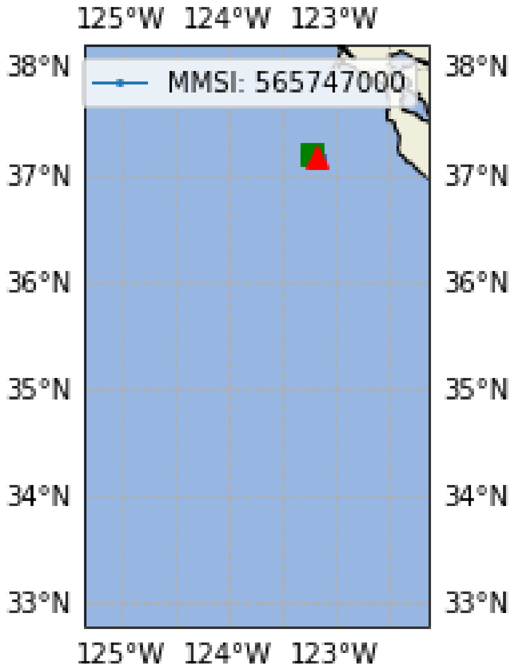
Now, once the program was end, let analyze the results.

The most negative score, the most anomalous behavior (Loitering). The most positive score, the most normal behavior. So, we can conclude, for this simple experiment that 565747000, 373932000, 357051000, 303352000, 431496000 are good candidates to be investigated by Loitering. Let's see. I developed a function which create a char by each time frame of a given MMSI. These are the results:


Vessel Course from 2021-05-25 02:48:32 to 2021-05-25 14:48:11
B: 387.0, G: 27.4, Fc: 20.4, Fchd: 11022399.0
125°W 124°W 123°W
38°N
37°N
36°N
35°N
34°N
33°N
MMSI: 565747000



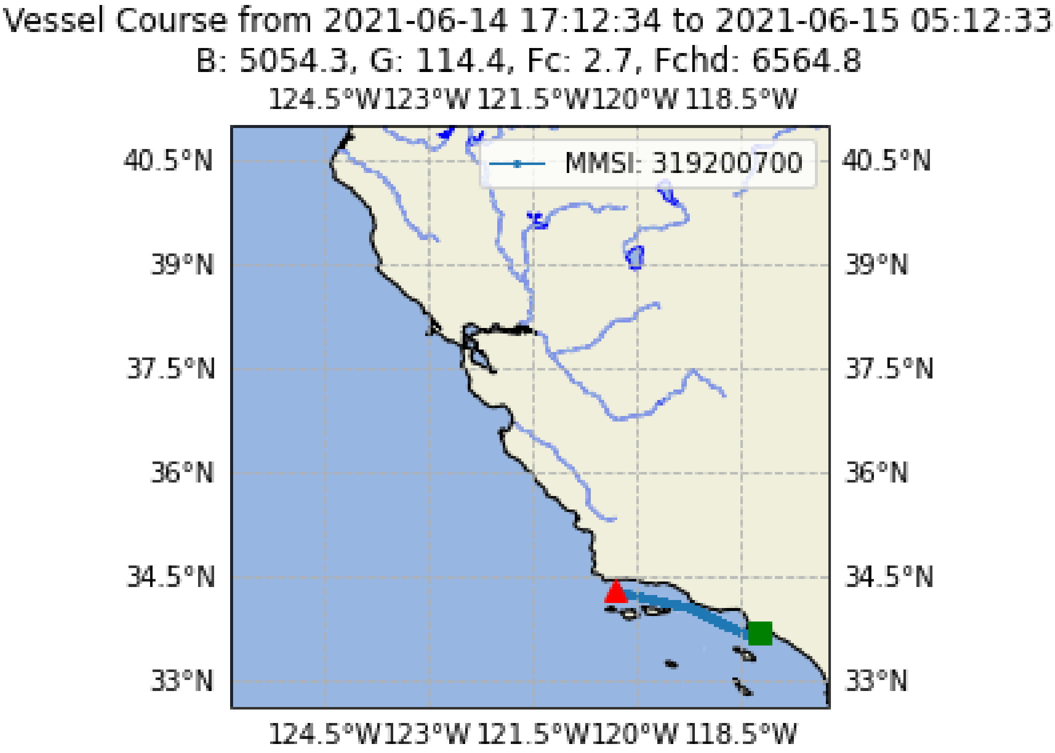
Vessel Course from 2021-05-25 13:35:50 to 2021-05-26 01:35:00
B: 1491.4, G: 83.0, Fc: 11.4, Fchd: 1076727.6

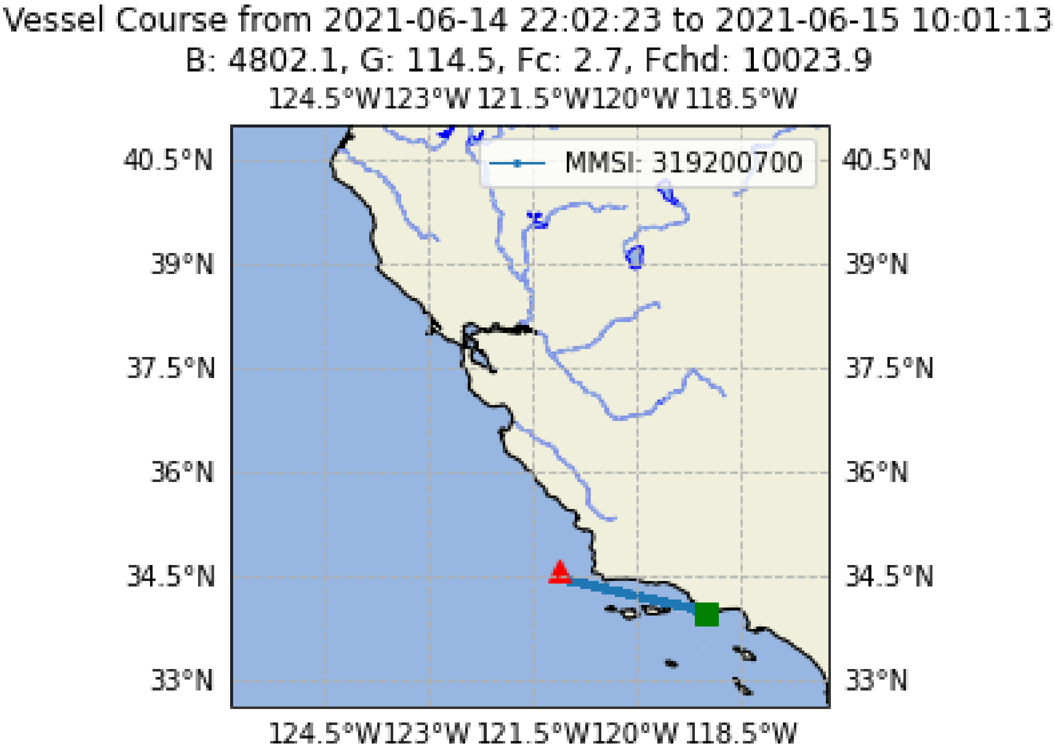


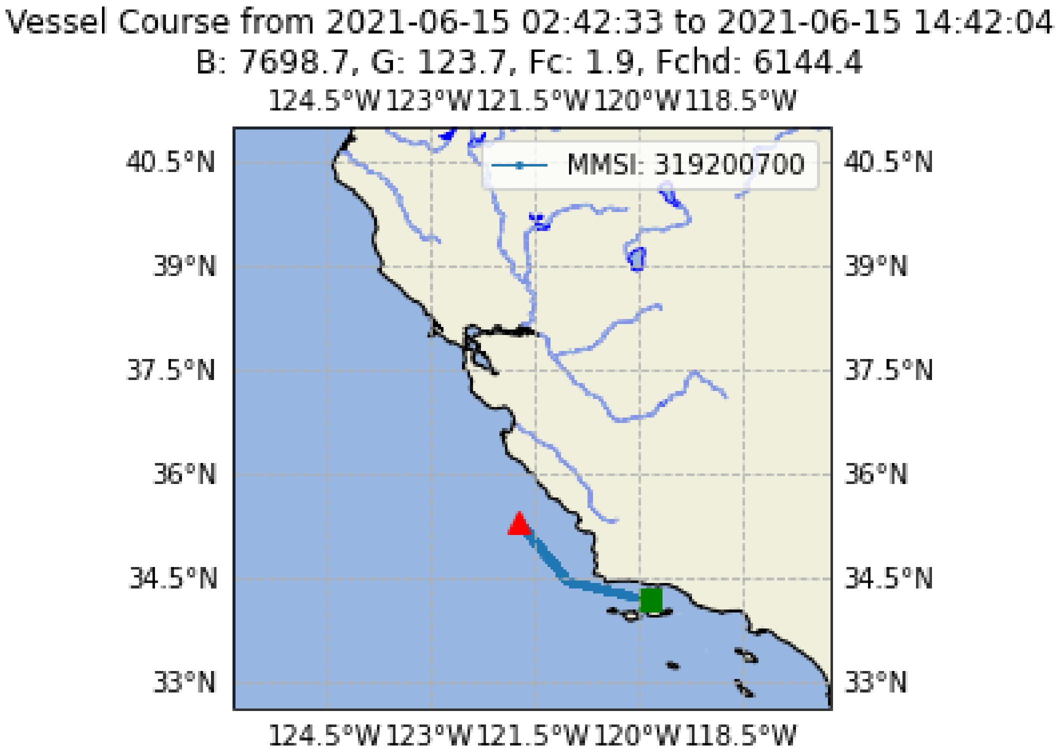
Vessel Course from 2021-05-26 17:01:11 to 2021-05-27 05:01:01
B: 21.7, G: -32.9, Fc: 3274.2, Fchd: -1291418954.6

In each figure, green square is the start point of the vessel (for the 24 hours time-frame) and red triangle is the end point. We can see that the vessel 565747000 (most negative score) starts navigating and one direction and then return to the neighborhood (remaining there lot of time) then the vessel move in direction to the coast and then remains in almost the same position lot of time.

By the other side. MMSI: 319200700 (most positive score) have this behavior:


Vessel Course from 2021-06-14 17:12:34 to 2021-06-15 05:12:33
B: 5054.3, G: 114.4, Fc: 2.7, Fchd: 6564.8
124.5°W 123°W 121.5°W 120°W 118.5°W
40.5°N
39°N
37.5°N
36°N
34.5°N
33°N
MMSI: 319200700


Vessel Course from 2021-06-14 22:02:23 to 2021-06-15 10:01:13
B: 4802.1, G: 114.5, Fc: 2.7, Fchd: 10023.9
124.5°W 123°W 121.5°W 120°W 118.5°W
40.5°N
39°N
37.5°N
36°N
34.5°N
33°N
MMSI: 319200700


Vessel Course from 2021-06-15 02:42:33 to 2021-06-15 14:42:04
B: 7698.7, G: 123.7, Fc: 1.9, Fchd: 6144.4
124.5°W 123°W 121.5°W 120°W 118.5°W
40.5°N
39°N
37.5°N
36°N
34.5°N
33°N
MMSI: 319200700

In each figure, green square is the start point of the vessel (for the 24 hours time-frame) and red triangle is the end point. We can see that the vessel 319200700 have a 'normal' trajectory, moving in efficient way straight forward to his destiny.

Analysis Techniques:

This study proposes a method to automatically detect loitering without the need for training normal instances. It creates a ranked list of loitering vessels to aid further anomaly investigation. The method defines loitering spatiotemporal characteristics and quantifies them using dynamic AIS message information. Parameters are formulated to determine a loitering trajectory by comparing various factors. The loitering score for each trajectory is calculated, and the Isolation Forest algorithm is used to establish a threshold and rank. Geographic visualization is then created for intuitive evaluation.

Observing equations below:

$$F\left(c\right) = \frac{\sum_{k=1}^{n} \Delta C_k \times \sum_{k=0}^{n} S_k}{180 \times B}$$

Equation (1)

Is easy to conclude that F(c) is directly proportional to the cumulative change of course (ΔC) and cumulative speed (S), as well as inversely proportional to trajectory Bounding Area (B). Intuitively, we can expect that loitering vessels are those ones have frequent changes of course and are navigating above a regular speed. Also, we can expect that in vessels with a loitering behavior, the covered area (B) during the trajectory will be lesser than a “normal” trajectory (because covered less nautical miles in the same time period). So, the factor 1/B will be bigger in case of loitering and magnify the total score. In summary the bigger the result of this equation the bigger the probability to have a loitering vessel.

$$F\left(c, h, d\right) = \frac{\sum_{k=1}^{n} \Delta C_k \times \sum_{k=0}^{n} \Delta H_k \times \sum_{k=0}^{n} S_k}{B \times G}$$

Equation (2)

By the other side, Equation (2) also consider the cumulative effect of de course and heading difference in a time-frame (ΔH). Again, is expected that loitering vessels have frequently changes between course and heading, big values of this parameter are indicative of loitering. Also considering G (geodesical distance) inside the time-frame will give us a good idea if vessel is moving in a “normal” trajectory (big G) or if is moving only a little (small G). So, analog to equation (1), the bigger the result of this equation (2) the bigger the probability to have a loitering vessel.

In order to have more “similar” values/scales to compare between the vessels, we apply the Isolated Forest algorithm to the this F(c) and F(c,h,d) scores. The isolation forest algorithm is useful to determine ‘outsiders’, that means the scores that are more different form the universe of scores. This is aligned with what we intuitively are looking for: Loitering vessels are the ‘outsiders’, that is we expected that most part of the vessels are using ‘standard’ speed, don’t have too much course/heading changes and cover an area and distance accordingly with the speed. Loitering vessels must be only few ones with characteristics of trajectory “very different” of the rest.

Also, the experiment proposes to combine these two scores In a weighted one. In order to capture a combined set of behaviors that can be useful to discover more anomalous cases. For this reason, the experiment uses EWM which will magnify the information with more entropy (higher differentiation degree).

Conclusions:

1. The method is reliable to be able to find loitering vessels.
2. Is possible to implement it in a “real-time” or “near-real-time”, depending on the availability of the data sources (AIS).
3. The method is and region-independent one to automatically detect loitering without training normal instances and produces a ranked list of loitering vessels to facilitate further anomaly investigation.
4. The method can be classified as an unsupervised one. That means that implementation is easy and isn’t dependent on labeled/verified dataset for training.
5. F(c) and F(c,h,d) are to good scores to measure the loitering behavior of vessels.
6. The negative the scores, the bigger the probability of loitering behavior.
7. Isolation forest performance well to discover the outsiders (loitering) behaviors.
8. Both can be combined using Entropy Weight Method in order to magnify the Loitering information.
9. The result of this method can be used to decrease the volume of work of operators who need to analyze tons of data in order to discover loitering behavior. The algorithm can suggest a very reduced rank of highly probable loitering candidates that can be submitted to operator analysis posteriorly.
10. Operator can label correctly the vessels with confirmed loitering behavior and then these datasets can be used to a machine learning algorithm (Convolutional NN probably) in order to learn the operator knowledge, reducing much more the human dependency for this task. (This can be discussed posteriorly and be added to this work as an enhancement).