Bunker Fraud Analysis Report

Cheong Min Wei (A0157235E), Celine Lim Shi-Yen (A0171682B), May Chan Shu Zhen (A0161940J), Xie Yangyang (A0156738R)

Abstract

Bunker fraud is estimated to cost companies and governments US$3 billion a year and this drives the motivation behind this project. With the huge amounts of data being produced so quickly today, we want to drive illegal bunker detection towards a data-driven approach. We aim to use Machine Learning techniques to learn from the data and identify trends which can help us to identify abnormal behaviour. Anomaly detection has been identified as an important aspect in the maritime domain which requires intensive research and development. Using unsupervised anomaly detection algorithms coupled with point-pattern analysis and trajectory analysis, we present an extensive report on the analysis conducted and the insights drawn from the analysis.

Literature Review

We have conducted literature review on existing research in the maritime domain. Much work has been done in anomaly detection to search and predict emerging conflict situations such as collisions and suspicious activity in the seas. The most popular method used in anomaly detection within the maritime domain is clustering, in particular density-based spatial clustering of applications with noise (DBSCAN). We have also identified point-based and trajectory-based anomaly detection methods as the 2 main classes in which data is passed into anomaly detectors. For example, Kowalska and Peel (2012) highlighted individual anomalous points whereas de Vries and van Someren (2013) considered whole trajectories. Furthermore, Liu et. al (2015) presented a similarity metric for partitioning trajectories using DBSCAN. To the best of our interpretations, we re-implement existing methods and make modifications to suit our task.

Many of existing work are based on spatial or temporal data mining and few are based on spatio-temporal methods. We try to bridge this gap in our work.

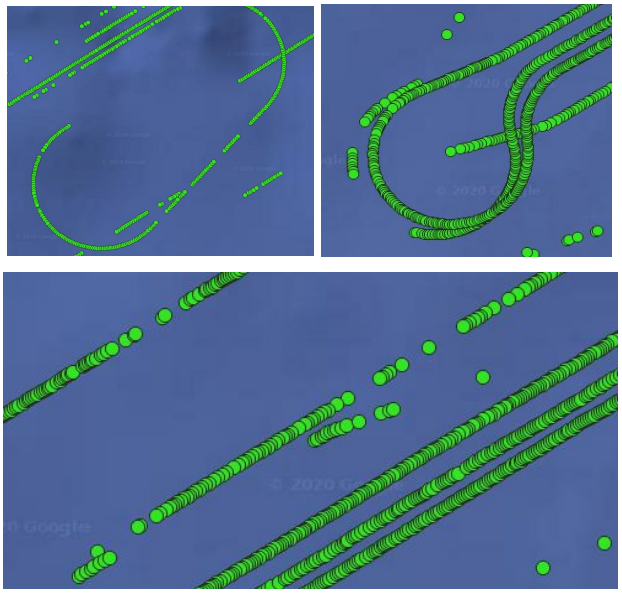
We also review some of the challenges mentioned in existing literature in order to find solutions and directions to tackle the commonly faced problems in maritime data. These include the disparate nature of the data and information, the velocity at which data is generated and the sampling frequencies. Other challenges faced include parameter initialisation such as thresholds, number of clusters, etc. We were also interested in feature engineering but observed a lack of studies on feature extraction in high-dimensional maritime data sets.

Outline of Approach

1	**Pre-processing of existing AIS datasets**
2	**Exploratory Data Analysis**
3	**Outlier / Anomaly detection** - Find vessels that exceed port limits and are located very close to another vessel for a prolonged period of time. - Use Machine Learning Techniques to identify potential anomalous vessels that deviate from normal ships in terms of certain features. (eg. *flow/direction, speed, location, time*) - Further analyse anomalous vessels' movement using a Vessel Movement Visualization to determine the potential of it being a bunker fraud case.
4	**Identifying limitations of previous methods and address them in subsequent methods.**
5	**Suggest appropriate recommendations to combat bunker fraud**

Methodology

We have a sequential series of steps conducted while defining our methodology. We first visualised the data set using exploratory data analysis. We utilised QGIS3.4 for all our visualisations and Python3 for implementing our algorithms.



1.0a. QGIS Trajectories

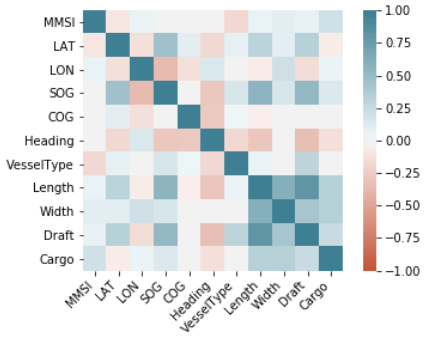
Visualization of the dataset in QGIS also showed that trajectories can take on different shapes such as linear, circular or oval.

Prior to beginning our investigation, we did some exploratory data analysis to get a better sense of what we were working with. A breakdown of vessel types in one year's worth of data (2017) is as follows:

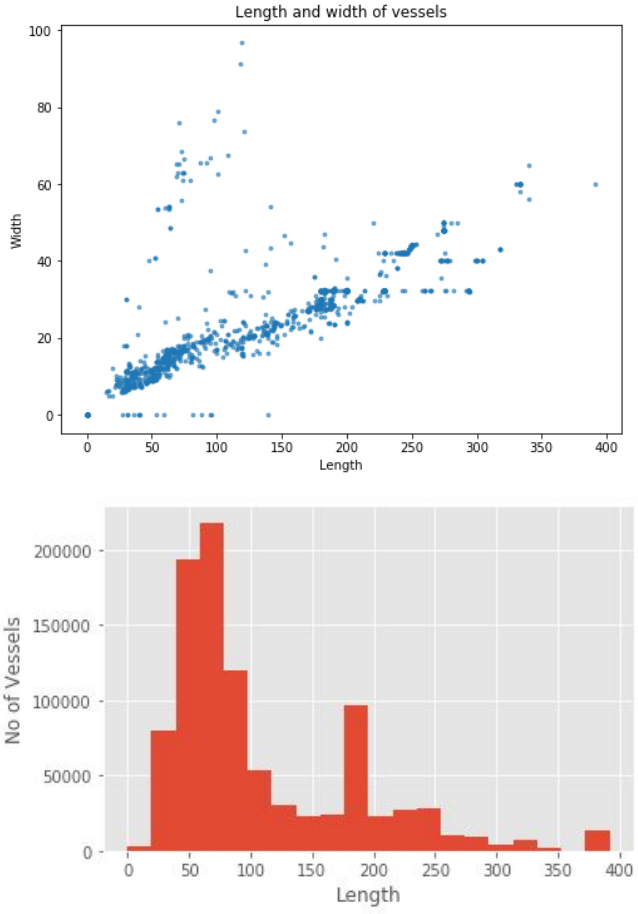
Tanker	146887
Cargo	145488
Passenger	112158
Tug Tow	42214
Fishing	4701
Pleasure Craft / Sailing	2106
Others	476719

1.0b. Tables showing number of vessels of each type

As we understand that there should be a source vessel and a recipient vessel in bunker fraud, we will be putting an extra focus on Tankers with suspicious activity involving other vessels.



1.0c. Correlation Matrix



1.0d1 and 1.0d2. Scatterplot showing ship length and width, and corresponding histogram

Our method consists of point-pattern analysis, followed by trajectory analysis, anomaly detection through clustering and finally visualization. Following this order, at every step of analysis, we evaluate the anomalies and define new measures to improve on the previous step.

Assumptions

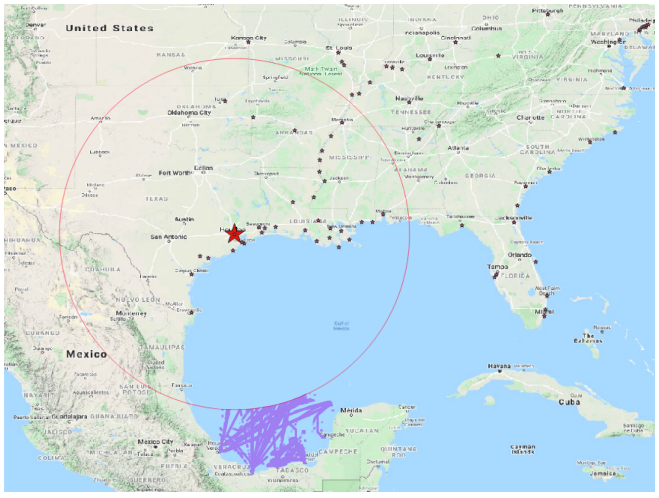
In order to perform anomaly detection, we assume for point-pattern analysis, 1% of the points are anomalous while for trajectory analysis, 10% of the trajectories are anomalous. We assume a high percentage of anomalous trajectories as compared to

points because we wanted to expand our search and include trajectories that are not extremely anomalous but could provide potential insights. Furthermore, there are fewer trajectories formed from the large number of points. Hence, we considered 10% to be a reasonable assumption.

We also assumed that bunkers can be identified through vessels that are tankers. In our analysis, we bounded our search for bunker vessels within vessels with type labelled as tankers.

For the bunkering process, we assumed the average time taken for a bunker to finish bunkering a vessel is between 14 to 18 hours.

Finally, we assume that port limits are constant and are within 24 Nautical Miles from Houston Anchorage Point as seen by the red star in the diagram below.



1.0e. Assumption of port limits

Experimental Results

1. Using TimeRatio variable in Cluster-Based Local Outlier Factor Algorithm

With our current working knowledge, we decided on the most trivial method and we considered how it translates to ship movement. Since fraudulent ships have the tendency to go off the grid, i.e. their AIS logs have a longer lag time, we used a time ratio to compare the lag times between each signal sent out by the different types of ships.

The formula is relatively straightforward:

$$r = \frac{x - y}{\Sigma(x - y)}$$

where the r is the ratio being calculated, x is the time of first signal and y is the time of the next signal. We removed the first point since there is no lag time calculated and the ratio is 0 for that point. We clustered the data points using the Cluster-Based Local Outlier Factor (CBLOF) algorithm to identify the possible anomalous time ratios. It clusters the data into small clusters and large clusters. The anomaly scores are calculated based on the size of the cluster the point belongs to and the distance to the nearest largest cluster. Outliers will have the highest scores. This algorithm was created by He et. al (2002) in order to balance out computation time as well as cost of computation. In our context, we chose it due to its ability to cluster non-anomalous points, as well as identify potential outliers in the dataset.

Prior to clustering, we scaled down both features of time ratio and SOG to a range between 0 and 1 to create an explainable visualization. The reason why we chose SOG, is based on the assumption that the SOG of ships should be a key factor in allowing us to assume that the ship is almost stationary and not in motion.

Upon running the results with an outlier factor of 1% (assuming that we only have 1% of outliers), we obtain the following figure that shows possible outliers.

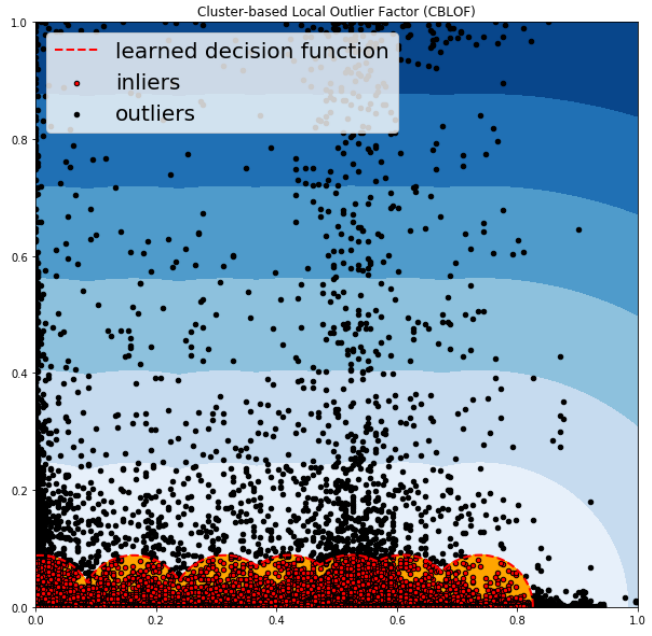


Fig 1.1 CBLOF Outlier Figure

Labelling the outliers obtained, we then plot out these outlying points on a geographical map for better visualisation.

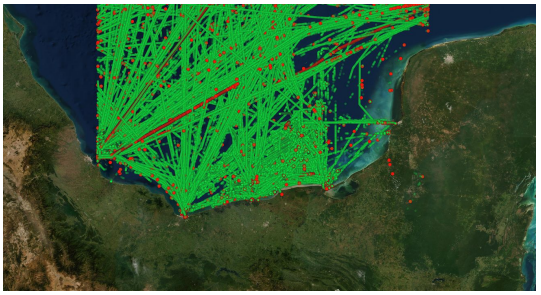


Fig 1.2 - Red points showed an anomalously long time where ship stayed almost stationary

We extract the anomalous points and visually find neighbouring points that could be involved in suspicious activity with the anomalous ship. We filtered the data for a shorter time frame so that we could take a closer inspection, as well as labelled the MMSI's of the various ships.

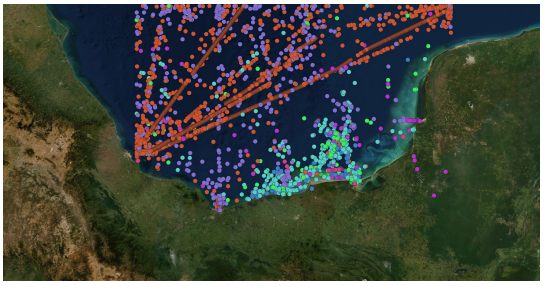


Fig 1.3 Outliers classified by ship type

Fig 1.4a Legend of points depicting the various bunkers

Fig 1.4b points of anomalous points found in a particular zone

Picking out two of the points that were close together and of different vessel types, we plot their trajectories to have a better picture of the scenario and possibly what could be illegal bunkering.

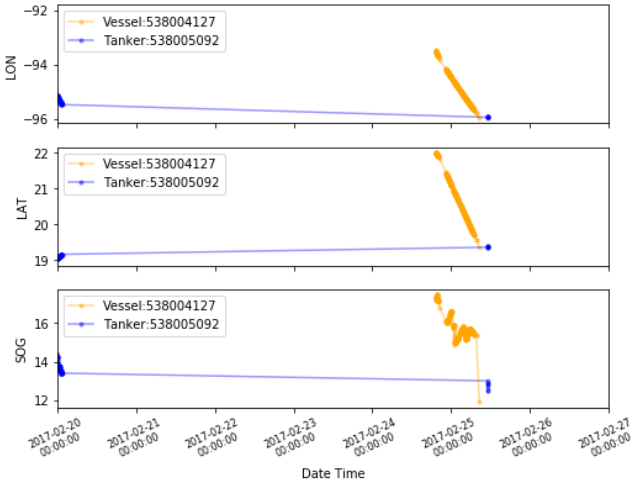


Fig 1.5: Trajectories belonging to suspicious activity

From Fig 1.5, we identified some suspicious activity as it can be observed that the Cargo ship was slowly moving towards the trajectory of the Tanker during the same period of time. These observations may indicate that they were bunkering at sea during this period as both ship's signals were turned off when they were close to each other, with their SOGs dropping rapidly till last transmission. However, since this was identified using just points, we could not be precise about the activity of the ship as compared to using a trajectory. As a result, we moved on to trajectory based calculations.

2. Density-Based Spatial Clustering of Applications with Noise

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm that seeks to separate clusters of high density from clusters of low density. Compared to traditional unsupervised clustering algorithms such as K-Means which is only good at capturing spherical clusters, the DBSCAN introduced by Daniel Wu (2017) is able to discover clusters of various complex shapes (e.g spherical, elongated or linear). Due to these properties of DBSCAN, we decided to use this method to cluster our trajectories as initial analysis has shown that most trajectories tend to follow a similar path and can have different shapes.

Our idea is that normal vessels tend to have similar routes and they will belong to dense clusters while suspicious vessels tend to take routes that deviate from the norm, hence will be in clusters of low density or may not belong to any cluster at all.

To represent the data points as trajectories, we turned to the MovingPandas library implemented in Python. This library implements a Trajectory class and methods based on the existing GeoPandas library. Rather than coding out the process of identifying a trajectory, we used existing libraries since they have been extensively reviewed and are efficient. In order to identify a trajectory, we set the minimum length of a trajectory to be 2 metres. Using the trajectory class methods in GeoPandas library, we created trajectories connecting the vessel's entire movement in the data set for each unique vessel.

To take into account both space and time as well as speed of the vessels, we extracted the features Latitude, Longitude, SOG and time difference between consecutive points in a trajectory and Hausdorff distance is used to compute pairwise distances between them. This gives us a distance matrix. We passed this precomputed distance matrix into sklearn's DBSCAN algorithm to identify different clusters of trajectories as well as trajectories that do not belong to any cluster. Trajectories that do not belong to any cluster are considered highly suspicious and were flagged for further analysis.

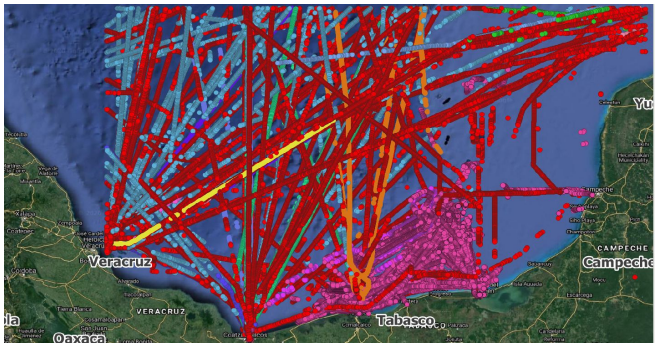


Fig 2.1: Trajectories belonging to different clusters

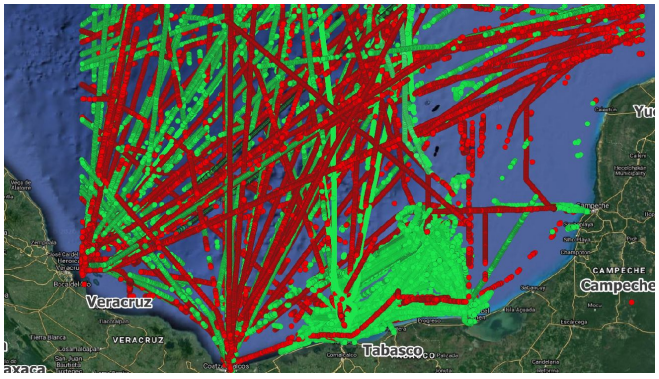


Fig 2.2: Normal (Green) and Abnormal (Red) Trajectories

Fig 2.1 shows the trajectories coloured by the clusters that they belong to. To identify potentially suspicious vessels, trajectories that do not belong to any cluster are coloured red while trajectories that belong to a cluster are coloured green in Fig 2.2.

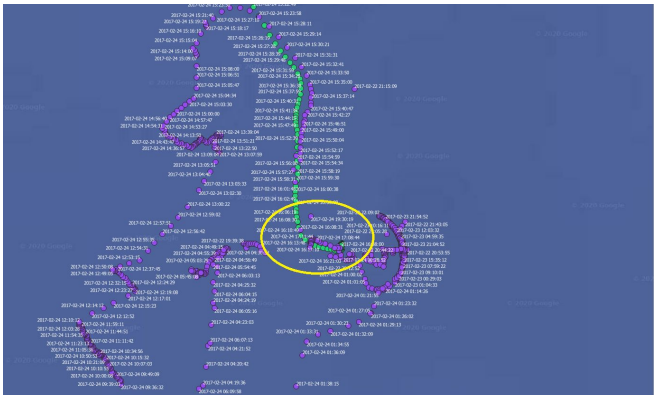


Fig 2.3: Potential Illegal Bunkering Process between Tug Tow (Purple) and Tanker (Green). Yellow oval indicates the end of AIS signal transmission for both vessels.

Further analysis is done on the abnormal trajectories to identify if these trajectories are indeed suspicious. We filtered out the anomalies, and tracked trajectories pairs which are close to each other both in time and space with greater focus placed on pairs that include a tanker. We visualised their latitude, longitude and SOG over time using our Vessel Movement Visualization to see if there are any overlaps.

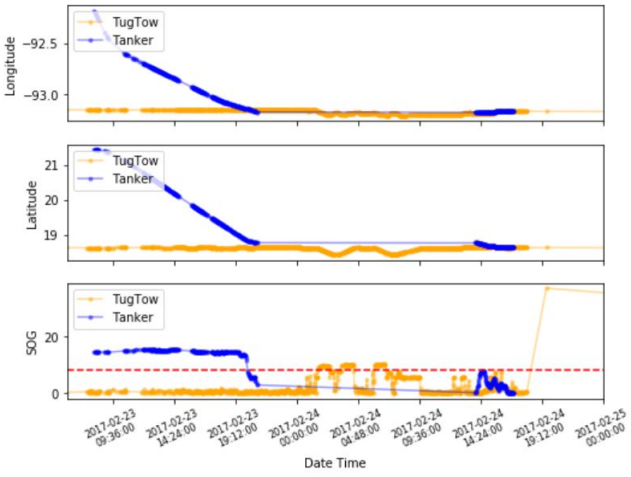


Fig 2.4: Vessel Movement of potential illegal bunkering

From Fig 2.4, we identified a potential bunkering activity as it can be observed that the Tanker approaches the Tug Tow that was already waiting in place for the bunkering during the same period of time. During the periods 2017-02-23 19:12:00 to

2017-02-24 14:24:00, the latitude and longitude of both vessels are aligned and they largely remained in place with their SOG being below 5 knots. These observations may indicate that they were bunkering at sea during this period as this was a 19 hour duration of little movement by both vessels which is a time period within the average time taken for a bunkering process. Moreover, at 2017-02-24 19:12:00, the Tug Tow started accelerating to a SOG of above 20 knots, suggesting that the bunkering process has ended and it is starting to move to another location.

While this may seem like a normal bunkering activity, there were a few suspicious points that could indicate this as a potential case of bunker fraud to investigate.

Firstly, we note that the Tug Tow arrived at the location much earlier than the Tanker. The high concentration of points by the Tug Tow in a circular pattern before the Tanker arrives could indicate that it was waiting at that spot for it to arrive. According to Wu, D (2017), bunker barges can use time as a leverage by deliberately arriving later than ordered to cause stress. As the crew onboard the vessels have a strict schedule to follow, by deliberately arriving later than ordered, it causes stress among the crew, leading to lower awareness around the bunkering process.

Moreover, even if the captain of the vessel that is being bunkered suspects a case of cheating, the bunker barge can simply refuse to detach itself from the vessel, delaying their schedule further, leaving them no choice but to accept the suspicious bunkering. Hence, the delay in arrival time of the Tanker could signal a potential illegal bunkering activity.

Secondly, during the bunkering process, we also note that there were no AIS transmission by the Tanker as indicated by the flat blue line in Fig 2.4. After the bunkering process, there is also little to no further AIS transmission by both vessels. This could possibly be a deliberate move by both vessels to go dark to avoid detection.

These observations suggest that this could be a potential case of bunker fraud to investigate.

3. Trajectory Clustering Model

Trajectory Clustering Model (TCM) introduced by Sheng P. (2018) is a model that seeks to cluster trajectories. Instead of passing all trajectories into DBSCAN like the previous approach, TCM selects the data points in a more concise and precise way.

Firstly, it calculates the rate of change of SOG and COG for each data point in time T with respect to the data point in time T-1. If both change rates exceed a threshold, the point is considered to be a characteristic point to represent in a trajectory. Due to time constraints, we did not apply the minimum description length principle to include other possible characteristic points.

Secondly, the characteristic points are used to calculate distance between each pair of sub-trajectories. Hausdorff distance is used to compute spatial distance. Directional distance is computed by using the sub-trajectory with shorter length and the angle between 2 sub-trajectories. Speed distance is calculated by difference of the average SOG of 2 sub-trajectories. After that, a synthetic distance function normalizes 3 distances to represent the distance between 2 sub-trajectories.

We then passed the computed distance matrix into DBSCAN to cluster all the sub-trajectories.

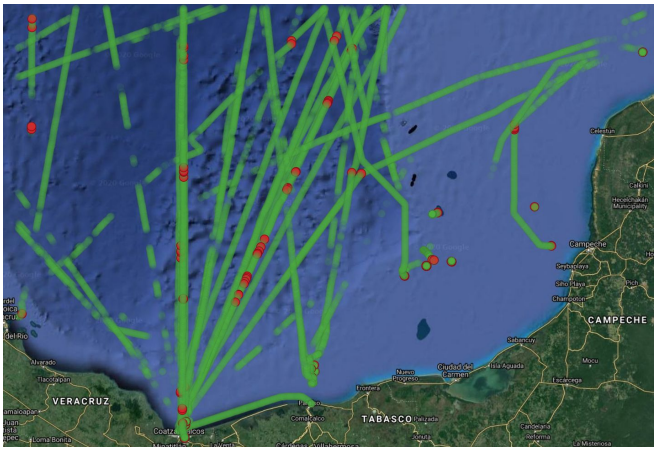


Fig 3.1

The red points in Fig 3.1 above are labelled as abnormal points according to DBSCAN. After we examined closely, the tanker in red showed some suspicious behaviour shown in Fig 3.2 below. On 25 Jan 2017, It appeared to refuel a vessel labeled in green. The tanker actually departed a few days ago on 21 Jan 2017 and stayed because there was no AIS data

reported after that until 25 Jan 2017. The vessel also went dark for a while from 06:17 to 06:49.

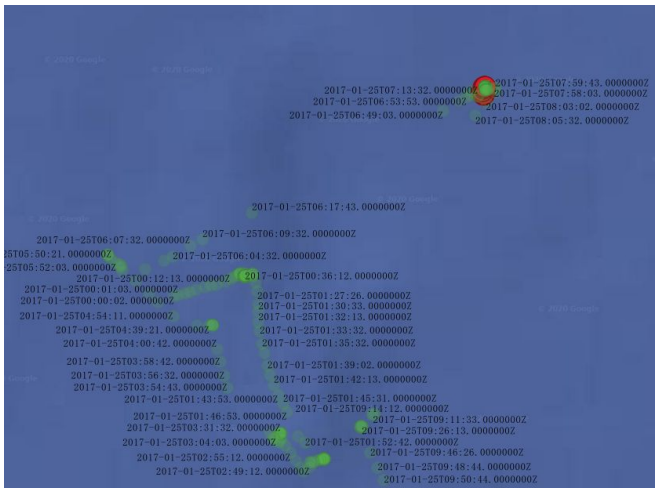


Fig 3.2

4. Dynamic Time Wraping

Previously, trajectories of an entire vessel's movement were passed into DBSCAN. However, from our assumptions, bunkering will take place within 18 hours. In order to account for this assumption, we decided to split the trajectories into individual trips with a maximum observation gap of 18 hours.

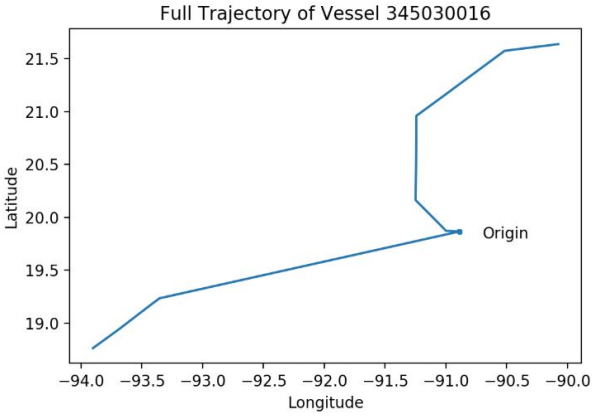


Fig 4.1: Full Trajectory of Vessel MMSI 345030016

In order to perform trajectory clustering, we first used Dynamic Time Warping (DTW), a similarity measure, to measure the distances between different trajectories. As compared to using the Haversine distances to compare distances between points on a sphere, we used the DTW measure as we wanted to incorporate time as a factor in our analysis. DTW compares two time series and warps the route from feature to feature to calculate the minimum distance which makes it easier to compare the shapes of trajectories. It can be used to measure similarity between two sequences which may vary in time or speed. However, DTW has a quadratic time and space complexity which is a major drawback in implementing a fast anomaly detector system. In order to resolve this issue, Salvador and Chan (2007) have implemented FastDTW which runs in linear time and space complexity. We utilise and credit their implementation in our work. The distance matrix obtained was derived from features Longitude, Latitude, SOG, COG and the time difference between consecutive points in a trajectory.

In the correlation plot below, we see that the DTW distances are very large in magnitude and we take note of this during the tuning of the eps parameter.

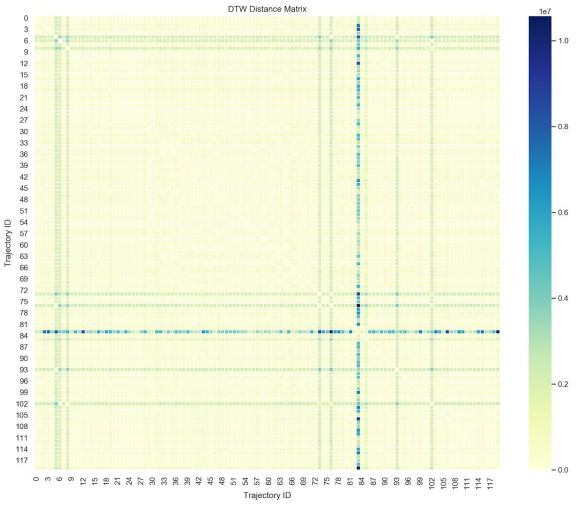


Fig 4.2: Correlation Plot of the DTW Distances

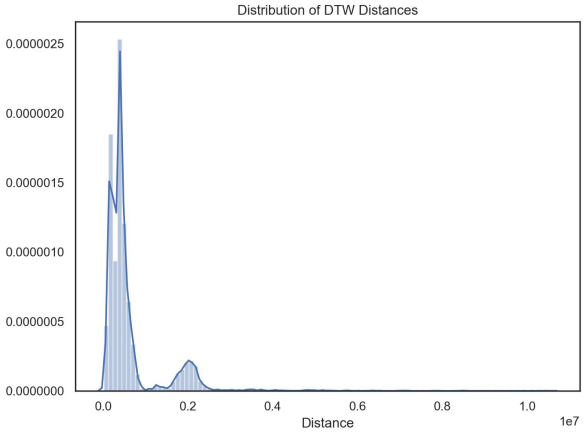


Fig 4.3: Distribution of the DTW Distances

We decompose the distance matrix using Principal Component Analysis and we see that 2 components explain at least 95% of the variance. We conducted

PCA to first examine the number of clusters we can expect from clustering.

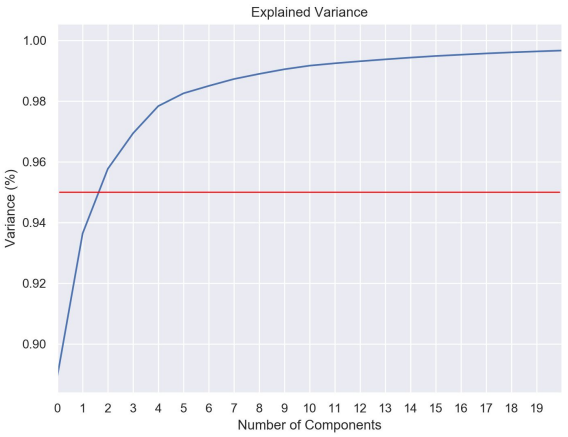


Fig 4.4: Proportion of Variance Explained by PC

Inspired by the works of Liu et al. (2015), we use the obtained distance matrix to filter for anomalous points. While Liu et al. created an algorithm based on a set of rules to identify anomalous points, we adapted their idea of using the distance matrix into a clustering algorithm. With the precomputed distance matrix at hand, we used DBSCAN to separate the anomalous and non-anomalous trajectories. To find the optimal value of the eps parameter in DBSCAN, we assume that 10% of trajectories are anomalous and select eps = 300000. The algorithm finds 2 clusters and this is expected as seen in PCA.

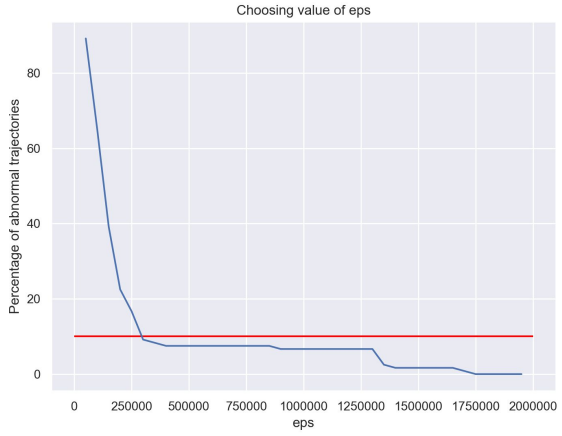


Fig 4.5: Fine Tuning the Parameter eps in DBSCAN

We obtain the following anomalous trajectories marked in red. In order to identify fraudulent behaviour, we visually examine neighbouring trajectories close in proximity to the anomalous trajectories. We note that while this was able to help us find suspicious vessel behaviour, we had to take note of several factors such as the date and time of trajectories being compared, and the distance of the trajectories from each other.

Fig 4.6: Full plot of anomalous trajectories (red) and normal trajectories (green)

We found that there were a few instances of suspicious illegal activities based on the time coincidence of two neighbouring trajectories, one of which has been labelled as anomalous by the DBSCAN algorithm. In the figure below, the two trajectories overlap in time and there is a clear gap in both trajectories which could imply that the transmission of the AIS signals have been switched off deliberately. We identify this as an instance of a vessel going dark on purpose.

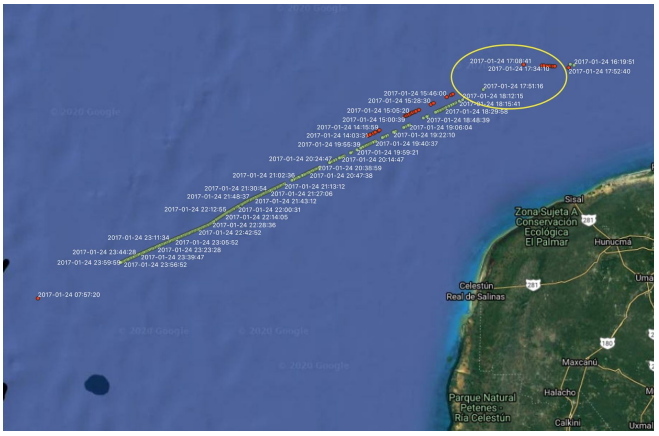


Fig 4.7: Pair of Suspicious Vessel Trajectories

We see that there is an intersection in longitude and latitude of both vessels between 2017-01-24 14:24:00 to 2017-01-24 19:12:00 using linear interpolation. We observe that while the 2 lines intersect, there are missing points for both vessels. The unusualness of both trajectories lies in the SOG. We see that there is missing SOG data points for the tanker being bunkered while there is a recorded data point for the

bunker vessel that dips to 5-7.5 knots. This could suggest an illegal bunkering activity during the time in which it slowed down since the sending of AIS signals could have been stopped deliberately. While we note that there is less than 2 hours during the time that there was a dip in SOG or missing AIS data, there might not be sufficient time for bunkering to take place, nonetheless this could imply a suspicious activity such as illegal trades.

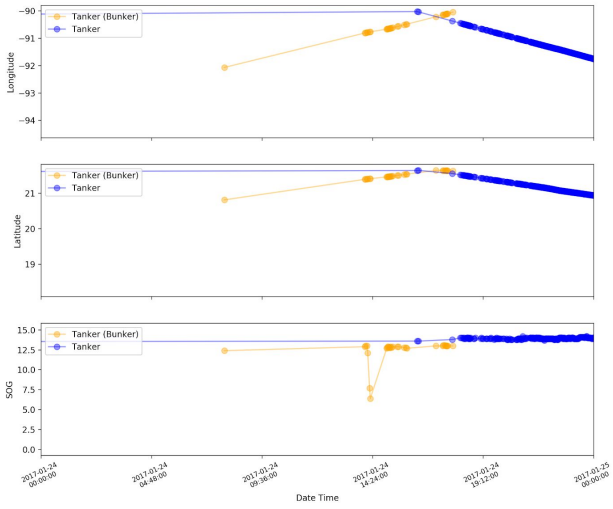


Fig 4.8: Vessel Movement of Potential Suspicious Activity

Limitations

During the process of finding anomalous ships, we faced several issues when using the DBSCAN algorithm. The parameters in DSBCAN has to be set manually and tuned to get optimal results. We identify this to be a limitation because if we were to convert our work into a system to automatically detect suspicious behaviour, the parameters eps and minimum points in a cluster have to be set before the clustering can take place. In consideration of future work, we hope that we can devise a method to automatically set the parameters.

Furthermore, while we were able to find outliers automatically, the analysis of the outliers had to be done manually through visualisation. We hope to be able to overcome this by locating neighbouring trajectories to anomalous ones via methods that consider space, time and direction of the vessel trajectories.

Recommendations

Upon analysing our results, we have decided a few approaches by which we can utilise our findings to drive impact as well as to improve our current observations.

Early Identification for Investigation

Firstly, using our current findings, we are able to identify certain ship types that have appeared to have anomalous movement - compared to the substantial amount of ships that were originally in the dataset, we are now able to single out certain ships that could have probable shipping activities.

Scoping down to these ships, further investigation can be done at an early stage to allow for prevention rather than late intervention. For example, regulations can be imposed to ensure that ships are reporting their ship capacities at specific intervals so as to better track possible suspicious activity. Failure to comply may result in a fine, or possibly a retraction of ship licensing.

Automation of Fraud Predictions

Rather than finding the vessels near anomalous points by visualization, there are options available for future work. We would like to suggest a system that can automatically churn out the conjugates to the anomalous points or trajectories. For example, we could consider graph networks built from data in the database. Graph databases are particularly useful in exploring highly connected data and would be useful to help us to find neighbouring vessels to outliers in a quick and efficient manner.

In that particular vein, automated dashboards may provide stakeholders with a better view of current bunkering conditions and allow for early intervention. Since dashboards are often comprehensive and provide a good overview in terms of detail and impact, the increased monitoring of current ship movements and activities will help Stakeholders easily notice a spike or dip in real time. As demonstrated earlier in the various Vessel Movement charts, some of the detailed metrics that should be displayed on the dashboard could include: (i) Average Bunkering Volume (ii) Average Time taken for different types of ships from port to port (iii) Average time taken for AIS signals to be received - So as to be able to track any

outliers who may be going off the grid for suspicious activities.

Increasing Robustness of Current Models

Based on our current models, some of the algorithms may be computationally intensive and require modifications to become more efficient. Given enough resources to be able to figure out the missing SOGs values in the dataset will allow more material for the algorithms to learn from, and subsequently, predict better predictions of potentially fraudulent bunkering activity.

Due to time constraints, we are unable to tune many of the prediction algorithms to find the optimal prediction parameters. Given more time, one possible way to optimise the outlier algorithms such as DBSCan and CBLOF would be to run GridSearchCV. Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model. A list of values to choose from should be given to each hyper parameter of the model, then cross validation is performed in order to determine the hyper parameter value set which provides the best accuracy levels.

As the dataset is large, we might need additional computational power that could be realised given more resources. But at an industry level, in order to ensure better prediction rates, it might be more cost effective to invest in such resources, than to spend more on checking suspicious vessels.

Collaboration with Local Authorities and Stakeholders

With the increase in globalisation, fraud detection has become an important goal for many port authorities, ensuring that their port is safe, smart and efficient. Leveraging on the power of analytics as well as the maritime understanding and knowledge of local port authorities and stakeholders, it is then that we may optimise efficiency and predictive power of current algorithms. The wealth of information garnered will then be fed back into the algorithm, and the contextual expertise of users will work in tandem with data to provide more insight to the predictions that are generated.

In Singapore, such collaborations are already present and in place. Using the IBM Traffic Prediction Tool, predictive analytics will be applied to forecast vessel arrival timings and potential traffic congestion. The partnership between IBM and the Maritime Port Authority of Singapore will also uncover new methods for sense-making and aid in event monitoring to detect unusual behavior of vessels and prevent illegal bunkering through fusion analytics, anomaly detection and data mining. These digital capabilities are intended to improve port security and safety, and could be one learning point for stakeholders in the Bunker Fraud Detection Industry.

Conclusion

Using the various algorithms, we have obtained sufficient evidence that could prove useful in identifying suspicious vessel activity. In the long run with proper tuning, these algorithms should serve useful in the prevention of more illegal bunkering activities, and contribute to making our seas a safer medium for trade and logistical transportation.

References

1. K. Kowalska and L. Peel, "Maritime anomaly detection using Gaussian Process active learning," *2012 15th International Conference on Information Fusion*, Singapore, 2012, pp. 1164-1171.

2. Vries, Gerben Klaas Dirk & Someren, Maarten. (2013). Recognizing Vessel Movements from Historical Data. 10.1007/978-1-4614-6230-9_7.

3. B. Liu, E. N. de Souza, C. Hilliard and S. Matwin, "Ship movement anomaly detection using specialized distance measures," *2015 18th International Conference on Information Fusion (Fusion)*, Washington, DC, 2015, pp. 1113-1120.

4. He, Zengyou & Xu, Xiaofei & Deng, Shengchun. (2003). Discovering Cluster Based Local Outliers. Pattern Recognition Letters. 24. 1641-1650. 10.1016/S0167-8655(03)00003-5.

5. Sheng, P.; Yin, J. Extracting Shipping Route Patterns by Trajectory Clustering Model Based on Automatic Identification System Data. Sustainability 2018, 10, 2327.

6. Stan Salvador and Philip Chan. 2007. Toward accurate dynamic time warping in linear time and space. Intell. Data Anal. 11, 5 (October 2007), 561–580.

7. Wu, Daniel & Aarsnes, Marion. (2017). An Introduction to Assessing Bunkering Operations Through AIS Data. 10.13140/RG.2.2.21415.04009.