Trajectory Clustering and Pattern Detection in Fishing Vessel Networks

Student: Colin Bradley (cb4102)
Advisor: Dr. Stanislav Sobolevsky

This work was enabled, in part, using the resources provided by the CUSP Data Facility.

TABLE OF CONTENTS

List of Figures and Tables

Abstract

The global fishing industry operates at a large scale and can have adverse environmental impacts on habitat and species, with an estimated 69% of global fishing stocks estimated to be fully or over exploited according to the Food and Agriculture Organization of the United Nations[1] . Yet detailed knowledge where and how fishing vessels travel and harvest is limited, particularly in international waters. Using location and trajectory data from Global Fishing Watch, this research applies trajectory clustering and pattern detection and trajectory classification to model the movements of the international fishing industry and understand repeated patterns of behavior. The research has implications for understanding what precise areas of the planet are being overfished and how patterns in fishing vessel trajectories can inform international environmental stewardship efforts.

Introduction

Global fisheries are an important source of nutrition and sustenance as well as economic opportunity for much of the world's population. In 2016, the Food and Agriculture Organization of the United Nations (FAO) estimated global production at 170 million tonnes, of which 90.9 million tonnes was harvested from marine areas (79.3 million from ocean fisheries and 11.6 million from inland lakes and rivers)[2]. The FAO estimates there are 4.6 million fishing vessels globally and that 59.1 million people derive income and employment from the industry.

However, global fishery population are on the verge of collapse; Worm *et al.* (2006) estimate that by 2050 most species populations will have collapsed (defined as over 90% of a species being depleted), driven by a combination of climate change and overfishing. Overfishing may be an even larger problem than previously estimated, with Daniel and Zeller (2016) estimating that the FAO may underestimate fishing extraction levels by up to 50%. In order to address this global crisis, governance and regulation will need to be reformed, including improved data and analysis on the impact of overfishing. One aspect of improved data and analysis is in understanding the behavior of fishing vessels, including the underlying patterns and trajectories for how and where vessels move around the ocean.

This research proposes applying spatio-temporal trajectory clustering and pattern detection modeling to high frequency GPS point data (approximately 64 million GPS points) generated from 985 fishing vessels with location information from 2012-2016[3]. The goal of the research is to gain a more granular understanding of what precise areas of the planet are being overfished and how patterns in fishing vessel trajectories can inform international environmental stewardship efforts. Specifically, the scope of the research includes three analytical approaches:

[1] FAO (2011)
[2] FAO (2016)
[3] Provided by Global Fishing Watch: https://globalfishingwatch.org/

- First, hot-spot clustering is conducted on GPS data using a DBSCAN (Density-based spatial clustering of applications with noise) algorithm to identify potential areas of overfishing.
- Second, vessel trajectories are segmented into trips and trajectory clustering is performed using K-means and DBSCAN algorithms to discover sequential patterns in the trajectories.
- Third, using features extracted from segmented trip trajectories a Random Forest Classification model is trained to test whether vessel types can be predicted based on their trajectory behavior.

The remainder of the paper is structured as follows: The following section contains a literature review on applicable methods; section three provides an overview of the fishing vessel data; section 4 details data pre-processing steps; section five provides a methodological overview of the analysis; the final section contains modeling results and discussion as well as conclusions.

Literature Review

The increasing availability of high-frequency GPS data as well as other technologies[4] has led to the proliferation spatio-temporal point data that can be used to model the trajectories of objects[5]. A trajectory can generally be thought of as a set of location points that are ordered chronologically, typically a latitude and longitude pair with an associated time stamp[6]. Trajectory clustering is considered one of the most important sub-fields of trajectory analysis[7]. Trajectory mining of spatio-temporal data can be grouped into three broad categories: Classification, clustering, and pattern mining and has been used for a wide variety of applications in the social and earth sciences[8]. The following review focuses on the literature most relevant for this research, specifically: trajectory pattern mining, trajectory clustering, and trajectory classification.

Trajectory Pattern Mining:

Pattern mining (PM) focuses on identifying patterns of movement that are hidden within trajectories, either due to noise or sparsity in data. There are two categories of PM relevant to this research study: togetherness patterns and sequential pattern discovery[9].

Togetherness patterns are focused on identifying patterns of common movement of multiple objects. The primary concern for togetherness pattern detection is how to characterize the spatial and temporal togetherness of objects (i.e. what is the proximity of objects such that we

[4] E.g. Global System for Mobile Communication, Radio Frequency Identification, WiFi, etc.
[5] Mazimpaka and Timpf (2016)
[6] Zheng (2015)
[7] Yao *et al*. (2017)
[8] Mazimpaka and Timpf (2016)
[9] Zheng (2015)

can consider them together and how many temporal periods do the object move together). There are a number of characterizations of togetherness patterns such as “flock”[10], which defines spatial togetherness based on a predefined space for a defined length of time. However, the “flock” approach has limitations such as rigidly defined shapes for determining togetherness (i.e. the predefined space). The “convoy” method attempts to overcome these limitations by using a density-based approach to measuring togetherness that can be applied arbitrarily to any shape of objects[11]. The swarm method builds on the convoy approach but does not require consecutive time intervals[12], which is relevant for identifying hotspot patterns in fishing vessel networks.

Sequential pattern discovery is concerned with finding common patterns from multiple object trajectories which travel in similar locations. Calculating sequential patterns in GPS trajectory data can be challenging since two sets of GPS coordinates, though they could be closely located, are not repeated exactly (dependent on the spatial granularity of the observations). One set of approaches for addressing this problem are clustering-based techniques. For example, GPS points (or groups of points) on trajectories are clustered and assigned a cluster id, such that a trajectory can be represented by a series of cluster id[13], which allows for dimensionality reduction of the trajectories.

Trajectory Clustering:

Trajectory clustering analysis groups objects into a finite number of categories based on a set of features or characteristics related to their spatial and temporal behavior. In the case of trajectories, this is primarily movement patterns, but may also consider things such as start/end location or speed. Like traditional clustering approaches, trajectory clustering algorithms rely on a concept of distance / dissimilarity to determine which cluster an objects should be placed in. A range of clustering algorithms have been proposed for trajectory analysis[14], and though it can be difficult to categorize trajectory clustering algorithms but many fall into traditional categories such as hierarchical (e.g. agglomerative), partitioning (e.g. K-means) or density-based (e.g. DBSCAN)[15].

Typically, trajectory clustering methods are an augmentation or application of a traditional method. For example, the Gaussian Mixture based approach was original proposed by Gaffney and Smyth (1999), while Birant and Kut (2007) developed the spatio-temporal DBSCAN (ST-DBSCAN) method to deal with spatial and temporal features in the DBSCAN model. The Tra-Clus algorithm developed by Lee *et al*. (2007) uses a partition and grouping approach to cluster sections of a trajectory into sub-trajectories. Zhang et al. (2018) build on the Tra-Clus

[10] Gudmundsson *et al*. (2006)
[11] Jeung et al. (2008)
[12] Zheng (2015)
[13] Zheng (2015)
[14] Yuan *et al*. (2017)
[15] Mazimpaka and Timpf (2016)

model and propose a hierarchical clustering approach for trajectory clustering using single-link clustering that considers speed, direction, and time. Ossama *et al.* (2011) augment K-means to cluster moving objects by introducing the object's directionality as a core clustering feature.

There are also more novel approaches and non-traditional approaches to trajectory clustering. For example, Yao *et al.* (2017) use a recurrent neural network (RNN) approach to extract fixed length trajectories from irregular GPS readings using a sliding window and sequence-to-sequence autoencoder. In another study, Dewan *et al.* (2017) propose a SOM-TC (Self-Organizing Map Based Trajectory Clustering) model, which pre-processes GPS trajectories using a neural network and captures uncertainty in the underlying location readings.

Trajectory clustering has been used in a wide range of applications, and is widely applied in fields such as traffic management, activity / movement modeling, and meteorology[16]. Palma *et al.* (2008) use an approach that helps characterize points on a single trajectory to identify important stops on a route. Chimwayi and Anuradha (2018) apply ST-DBSCAN to epidemiology, while Shaw and Gopalan (2014) use k-means to cluster frequent patterns in hurricane trajectories. A range of approaches have been used in transportation such as Kianfar and Edara (2013), who compared the efficacy of K-means, Hierarchical Clustering, and Gaussian Mixture Models (GMM) in identifying traffic congestion or Liu *et al* (2010) who apply agglomerative clustering methods (called Mobility-Based Clustering) to detect taxi traffic.

Trajectory Classification:

The third area of trajectory analysis relevant to this research is trajectory classification. Trajectory classification is primarily concerned with using trajectory data to classify objects. Typically, trajectories are segmented and features are extracted from the GPS data, and those features are used for classification modeling[17]. Trajectory classification models , with a major area of research concerned with classifying travel mode type[18] or classifying the status of an object on a trajectory, such as Zhu *et al*. (2012) who classify taxis as either occupied or not-occupied using processed trajectories and locations of interest (e.g. museums). Classification techniques have been proposed for fishing vessel networks, for example, DeSouza *et al*. (2016) use a hidden markov model and other machine learning techniques to mine patterns in vessel trajectories and classify types of vessels based on their speed and trajectory patterns.

[16] Bian *et al* (2018)
[17] Zheng (2015)
[18] Timothy *et al.* (2006)

Data Overview

The raw GPS location data is provided by Global Fishing Watch[19]. The data consists of Automatic Identification System (AIS) GPS coordinate information for 985 unique vessels (each denoted by a unique Maritime Mobile Service Identity number) from 2012-2016, where each latitude and longitude coordinates contain a timestamp. In total there are 64.6 million raw GPS points with a latitude and longitude range that spans the globe and an average of 65.5 thousand points for each boat[20]. There are a few noteworthy limitations in the data that will need to be addressed during data pre-processing:

- Not every boat has data for the entire time period so the length of the movement vector for each boat is different (e.g. A vessel may have 6,500 location data points while another vessel may have 100,000, as illustrated in *Fig 1*)
- Location data are not in even intervals for a given boat (as an illustrative example there can be 2 minutes between location data points, or 1 hour between two other location data points for the same vessel at different times in the trajectory)

Fig 1: Frequency of GPS Pings per Vessel

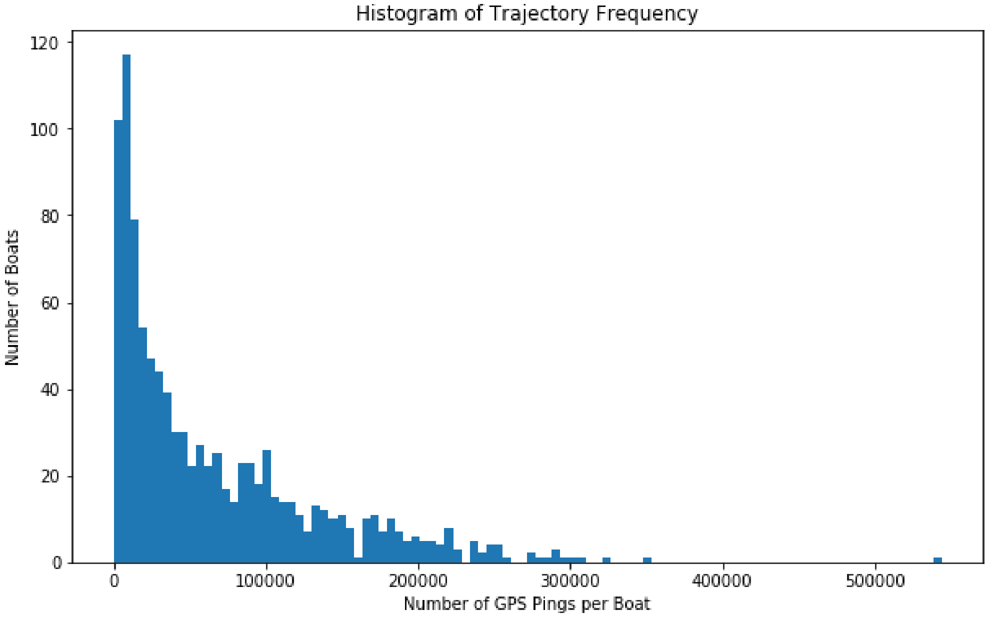


Figure 1 shows the frequency distribution of GPS pings per vessel.

In addition to a range of GPS frequencies, there is a wide range of trajectory behavior across vessels. Within a given vessel's trajectory, there can be macro behavior (overall route behavior)

[19] Raw data can be found at the following GitHub repository: https://github.com/GlobalFishingWatch/training-data

[20] An illustrative sample of the data can be found in Item 1 of the Appendix.

as well as highly localized patterns. Figure 2 shows a number of examples at different geographic granularity, to illustrate the variety and variation in the trajectories.

Fig 2: Vessel Trajectory Examples: There is a High Variation in Trajectory and Locations for Vessels in the Data Set

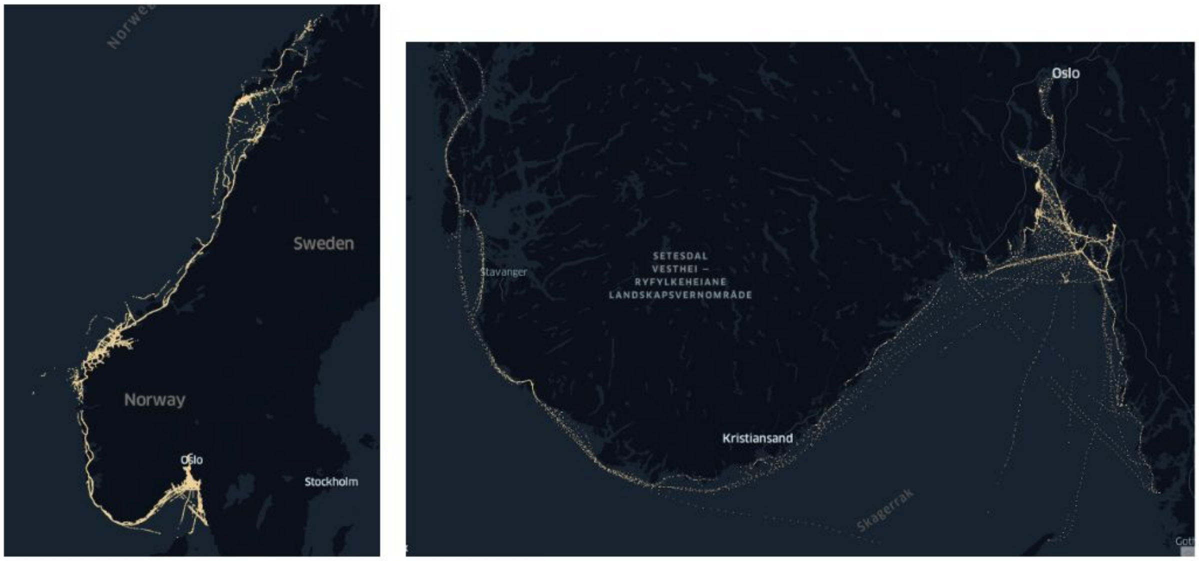


Fig 2A: Norwegian Vessel (number of locations: 97,984) that ranges from Oslo to Northern Norway, with a home port/base in Oslo, Norway. Fig 2A shows the extent of the range of the trajectory, mapped as points. The second figure shows trajectory detail around Oslo.

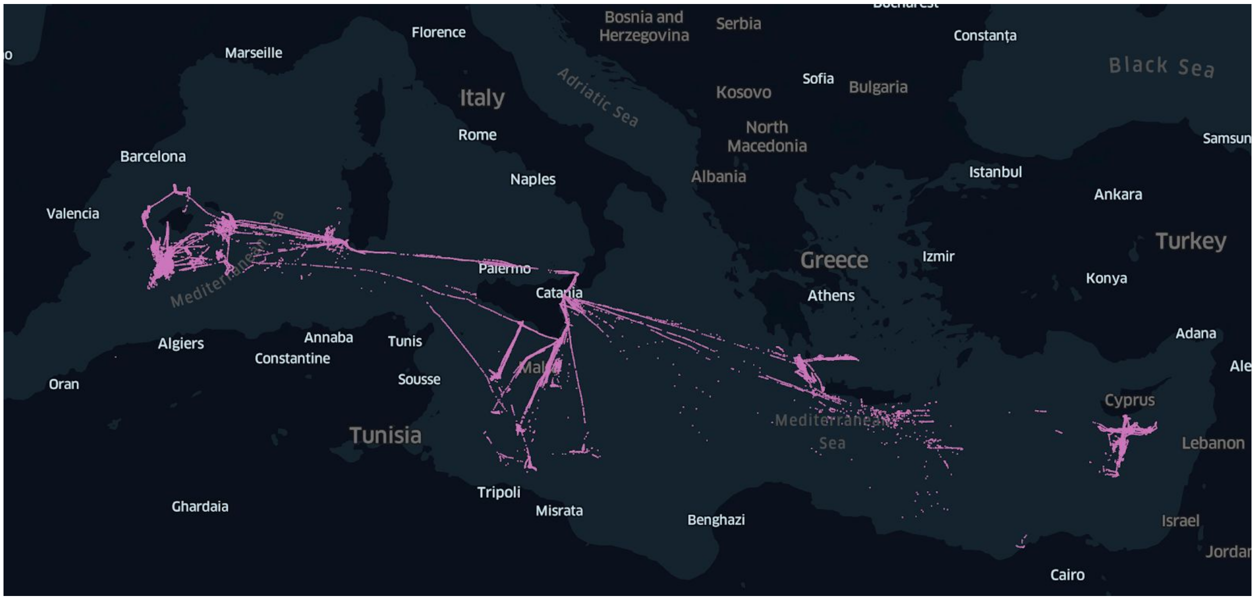


Fig 2B: Mediterrean fishing vessel (number of locations = 33,153)

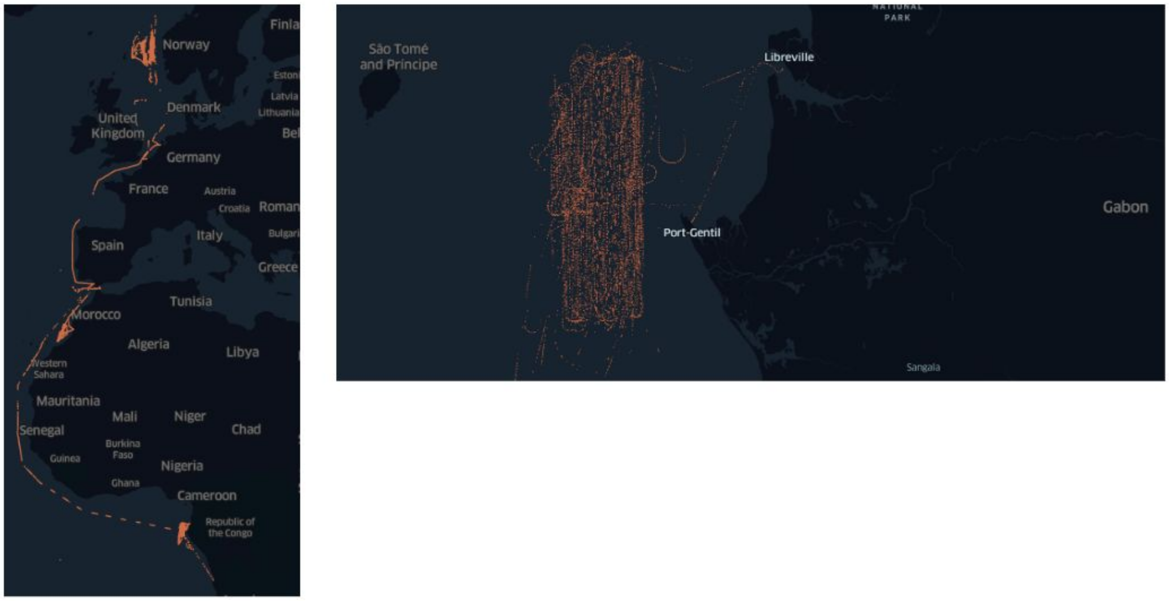


Fig 2D: International Fishing Vessel (number of locations = 32,504). Large range vessel which travels from Norway to South Western African. Fishing activity can be observed in clusters off the coast of Norway, Morocco, and Gabon. Detailed photo shows location activity off the coast of Gabon and is indicative of a fishing trawler.

Trajectory Preprocessing:

In order to conduct trajectory clustering there are a number of important data preprocessing steps in order to transform the raw location data into a structure suitable for cluster modeling, specifically resampling trajectories so that the time intervals are uniform and segmenting full vessel trajectories into more meaningful trips.

<u>Resampling:</u>

The vessel trajectory data presents two main issues that need to be resolved with resampling: non-uniform time intervals and inconsistent time ranges for each vessel. To overcome these issues, a uniform resampling approach was used, organizing the data into 60 minute time ranges time period 2012-2016. When a vessel has more than one data point in a 60 minute window, the last recorded point is used. When a vessel does not have a data point in timestamp window, a linear interpolation technique is used between the previously available and next available data point in the time series.

The drawback of resampling is that you are inevitably losing information contained in the raw data. Data is lost when interpolated averages are used, particularly if a vessel has data at a frequency higher than once every hour. In addition, the linear interpolation introduces additional error to the trajectory, particularly if there are long periods of time between points.

Trajectory Segmentation and Stay Point Detection:

Once the trajectories are resampled the concept of a "trip" is introduced in the data, in order to divide activity between time at port, idle time, and time traveling or harvesting. To do this, stay point detection is conducted. First, the speed at which the vessel is moving between two sequential location points is calculated. When the speed is under a certain threshold (<0.3 km/hour) the vessel is flagged as idle during that time interval. When a vessel has been idle for 4 hours a trip is considered complete (this is typically when a boat is at a port), and once the vessel begins moving again a new trip id is assigned to the boat. This splits a given boat's full trajectory into a series of segmented trips. This approach also has the added benefit of trajectory compression, reducing the size of the trajectory as stationary observations can be discarded. After segmentation, there are a total of 234,769 trips composed of 41,750,598 resampled GPS points. Figure 3 shows a sample of trajectories after they are processed into trips, demonstrating the high spatial granularity in the data.

Fig 3: Sample of Processed Trajectories

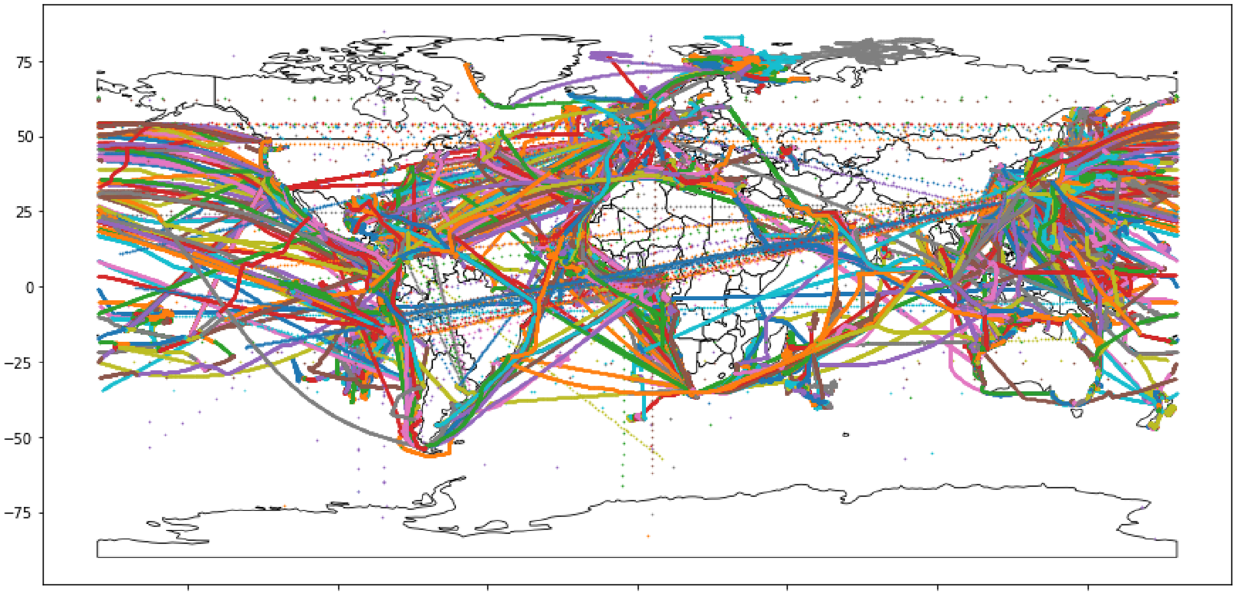


Fig 3 shows a sample of 33,442 segmented trips from the trajectories of 125 vessels. The figure demonstrates the high spatial granularity of the data, which contains 5,731,447 gps coordinate pairs in this sample.

To illustrate the context gained from segmenting trajectories into individual trips, Figure 4 provides examples of two vessels and shows their full vessel trajectories, segmented trip trajectories, and a map for geographic context.

Figure 4: Examples of Segmented Trips for an Individual Vessel

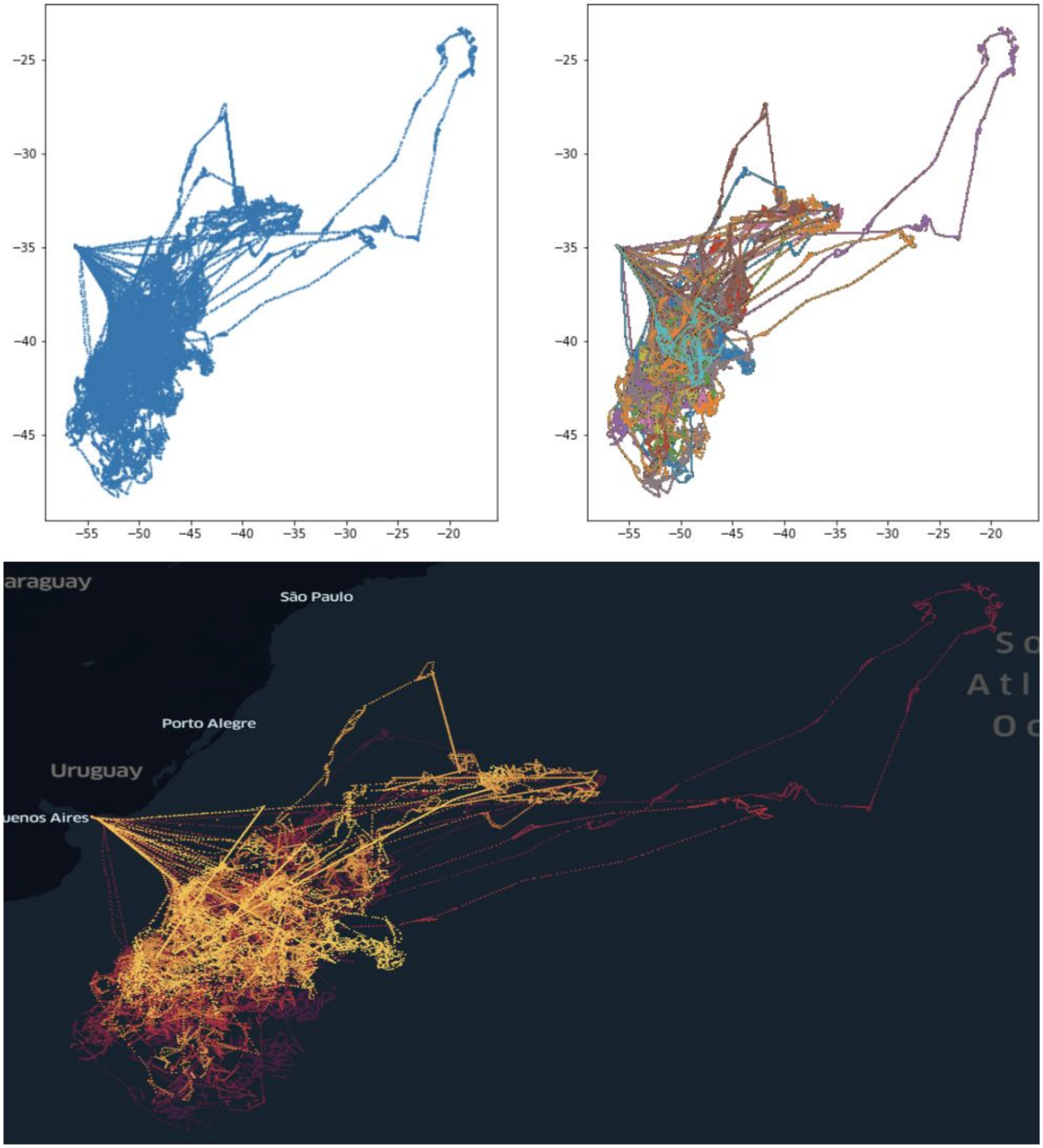


Fig 4A: High density GPS point data (top left) from a vessel, prior to segmentation. This vessel had 90,814 raw GPS points prior to data pre-processing and 36,510 GPS points post-processing. The total trajectory has been segmented into 47 unique trips (top right). The trajectory is provided on a map for context, colored based on the individual segmented trip, shows a regional vessel based on Buenos Aires (bottom).

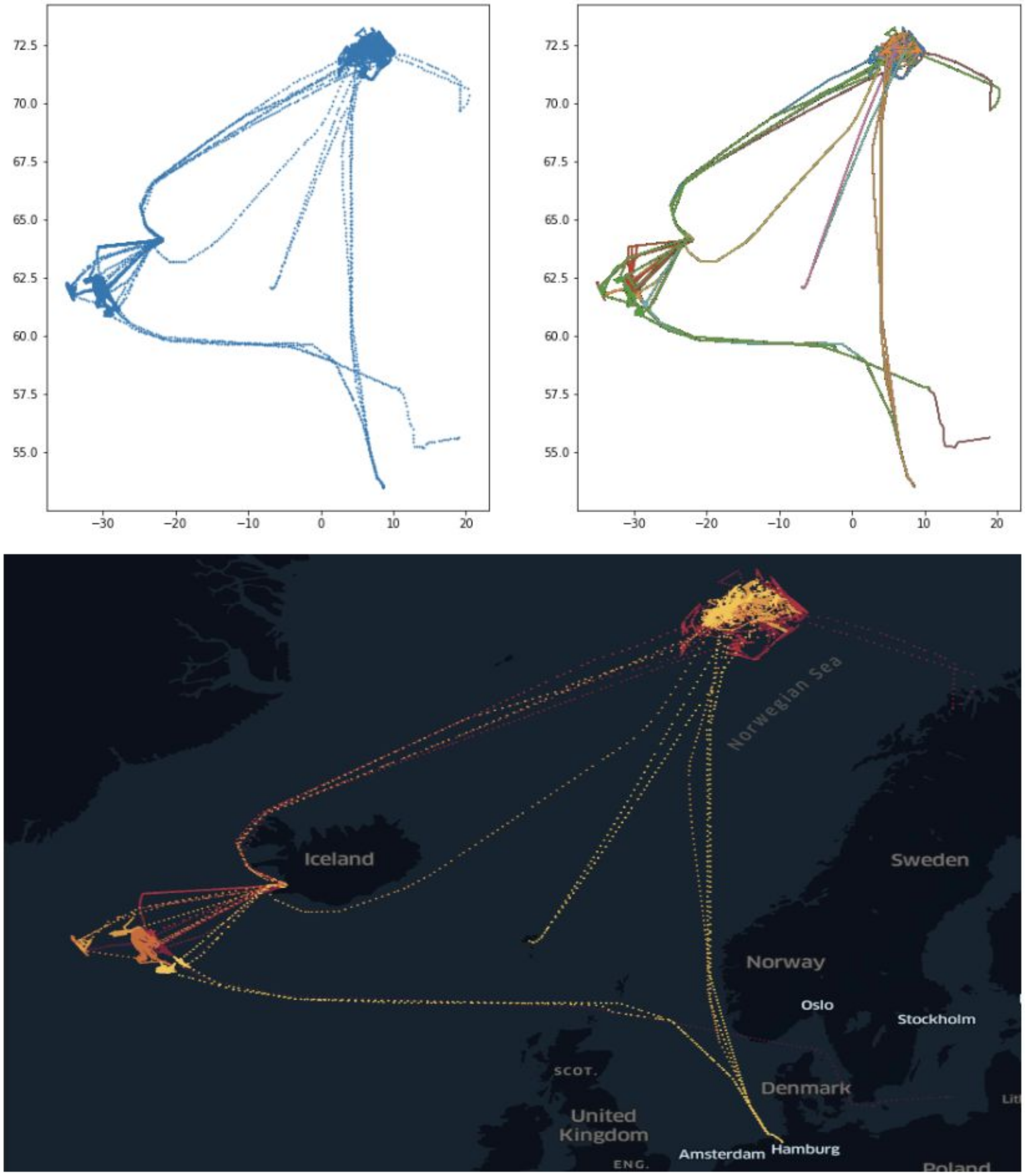


Fig 4B: High density GPS point data (left) from a vessel, prior to segmentation. This vessel had 69,574 raw GPS points prior to data pre-processing and 13,752 GPS points post-processing. The total trajectory has been segmented into 36 unique trips. The trajectory is provided on a map for context, colored based on the individual segmented trip show a fishing vessel in the North Sea, with a home port in Iceland.

Classification Model Trajectory Feature Extraction

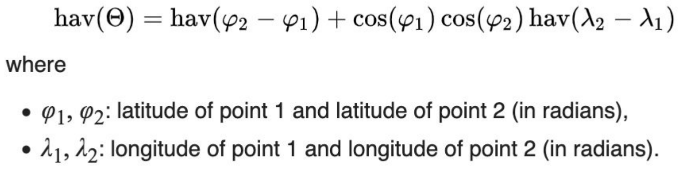
After the trajectories are segmented into individual trips, 8 features are extracted from the trips and used as input into the Random Forest Classification model. The features include: Distance traveled, average speed of trip, and geographic trip span (composed of minimum, maximum, and median latitude and longitude coordinates for the trip). In total there were 309,124 trips in the classification data set, reflecting a high sampling rate (20 minutes rather than 60 minutes) and the fact that not all vessels were labelled and so not all trips were included in the model.

Given that a supervised learning approach is used to classify the vessels, vessel type classifications were linked to the trajectory features based on MMSI. In total there are 12 classifications for vessels (e.g. Cargo, Trawlers, etc.), a full list of which can be found in Item 2 of the Appendix.

Modeling Methodology

Hot Spot Clustering

For hot-spot detection, a DBSCAN algorithm was employed[21] and tested at various levels for the hyper parameters epsilon and minimum number of neighbors. DBSCAN is a widely used clustering algorithm that works by identifying areas of similar density in the feature space and groups these densities into distinct clusters and then iterates to add similar trajectories to existing clusters, based on a trajectory's proximity to an existing cluster. The model takes an input epsilon (the radius measure with which to look for neighboring trajectories) and the minimum number of neighbors, however, the number of clusters does not have to be specified *ex ante*. Crucially, the model can also help with the identification of outliers and noise, which may arise from GPS reading errors, and also has the ability to generate cluster shapes that are irregular. For hot-spot clustering latitude and longitude coordinates were converted into radians and distance was calculated using the haversine distance function, defined as[22]:



$$\operatorname{hav}(\Theta) = \operatorname{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\operatorname{hav}(\lambda_2 - \lambda_1)$$

where

- φ_1, φ_2: latitude of point 1 and latitude of point 2 (in radians),
- λ_1, λ_2: longitude of point 1 and longitude of point 2 (in radians).

To improve computational run-time, latitudes and longitudes were rounded to two decimal places and the counts of GPS pings were preserved and used as weights in the DBSCAN algorithm.

[21] Ester *et al*. (1996)

[22] Wikipedia contributors "Haversine Formula"

Trajectory Clustering Pattern Detection:

Clustering is conducted on the segmented trip trajectories. The following section details distance measures used in the modeling and the modeling approaches.

Distance Measure:

Clustering algorithms require the definition of a distance metric that the algorithm can use to assess the similarity / dissimilarity between two objects. Traditional clustering methods, such as K-Means, use Euclidean distance between individual points, however, trajectory clustering requires measuring distance between two sets rather than just two points. There have been a number of proposed distance measures in the trajectory clustering literature such as Dynamic Time Warping, Edit Distance on Real Sequences, and Longest Common Subsequence[23]

In this work, Bidirectional Hausdorff Distance (BHD) is used[24]. Hausdorff distance works by calculating the distance between two sets and taking the maximum of the minimum distance between points in those sets[25]. For example, given sets A and B, BHD will calculate the distance from set A to set B and the distance from set B to set A (since these distances are not always symmetric)[26]. The equation is given by:

$H(A, B) = \max(d(A, B), d(B, A))$ where $d(A, B) = \max a \in A \min b \in B \|a - b\|$

Since we are analyzing spatial data (latitude and longitude), haversine distance is used as the distance metric between coordinate points. This distance calculation is used to generate a NxN dissimilarity matrix in Euclidean space, which includes the distances between a given trajectory and all other trajectories in the data set.

Clustering Model Approaches:

Once the distance/dissimilarity matrix is calculated, a k-means clustering algorithm[27] is used to cluster the trajectories. K-means takes as input the number of clusters (k) for data to be grouped into. The algorithm then initializes n = k number of centroids and every trajectory is assigned to a cluster (based on which centroid the trajectory is closest to). The algorithm then iterates through a process of moving the centroids to reduce total distance and re-assigning points to their closest center until the model converges on a solution when further movement of centroids and reassignment does not reduce the average distance of all points to their cluster center.

[23] See Zheng (2015) for an overview of trajectory clustering distance metrics.

[24] Chen *et al.* (2011)

[25] Bian *et al*. (2018)

[26] Chen *et al.* (2014)

[27] Lloyd (1982)

A DBSCAN algorithm, as described above, is also applied to the segmented trip trajectories and compared with the k-means approach.

Vessel Classification From Trajectory Features

To perform vessel classification, a Random Forest Classifier (RFC)[28] was trained on a set of 8 features extracted from the trajectories of each vessel's trips to classify vessels into one of 12 groups. In total there were 309,124 segmented trips where a vessel had a classification that could be used for model training. The model was trained and tested by splitting the data into training, test, and validation sets and a GridSearch cross validation was performed to tune hyperparameters[29]. Once the model was trained and tested it is evaluated overall and at the class level on precision, recall, and accuracy.

An RFC model was implemented because it is fairly robust to feature multicollinearity and generally performs well on multi-class classification tasks. Since an RFC is an ensemble method, the risk of overfitting the data is mitigated. As a non-parametric modelling approach, RFC is a useful method because it makes no underlying assumptions about the nature of the data. In addition, RFC Gini importance offers a natural method for assessing feature importance in modelling.

Results and Discussion

Hot Spot Clustering Analysis

Hot spot clustering using the DBSCAN algorithm was tested with three epsilon values (10km, 50km, and 100km) with the aim of reducing the dimensionality and noise in the trajectory data and to identify outliers and hotspots that are likely to be areas with overfishing activity. The inputs into the model are GPS locations and are weighted based on the number of pings in that area so that cluster intensity can be determined post-clustering. Epsilon = 50km produced the most promising results (see Fig. 5)[30], identifying 548 clusters and detecting 420,457 outlier points. Figure 5B shows the top ten cluster locations by intensity, identifying core fishing activity in the South Pacific off the coast of Chile, West Africa, and coastal Japan. Though the model successfully identifies a number of hotspots, the approach may not be robust to different levels of spatial granularity. In addition, the hotspots are only indicative of fishing effort but don't provide any context about what type of boats or the trajectory patterns in hotspot areas, suggesting that trajectory clustering on segmented trips may provide more context about fishing vessel patterns.

[28] Breiman (2001)

[29] Hyperparameters evaluated in the Grid Search CV include number of estimators, max tree depth, minimum sample leaf, and minimum sample split.

[30] For results for epsilon = 10km and epsilon = 100km see Item 3 in the Appendix.

Fig 5: DBSCAN Clustering Results for Epsilon = 50km

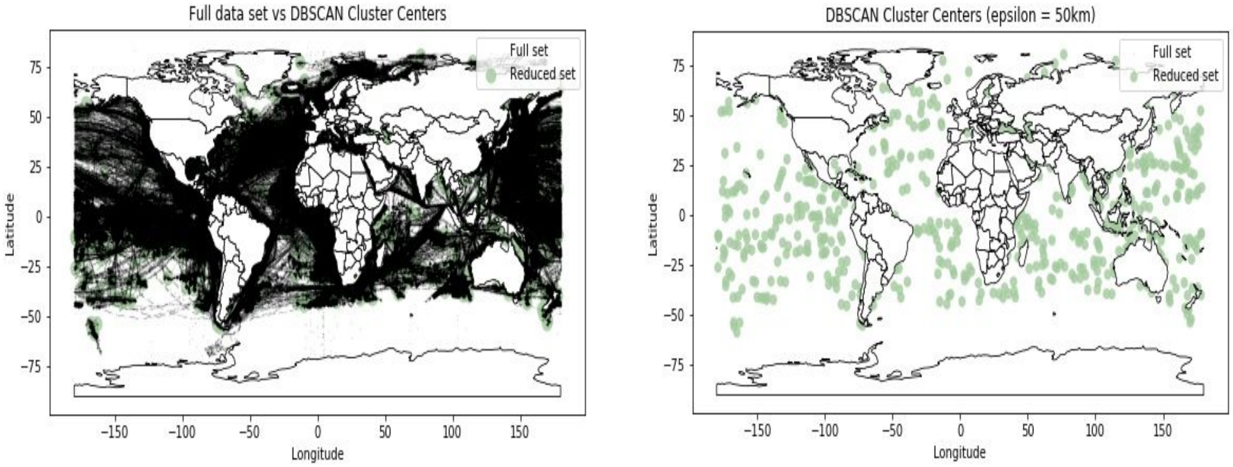


Fig 5A: GPS coordinate data is mapped (left) showing a high degree of noise in the data. The DBSCAN model successfully filters the noisy GPS data and identifies 420,457 outliers and 548 clusters. The cluster centers are identified (right)

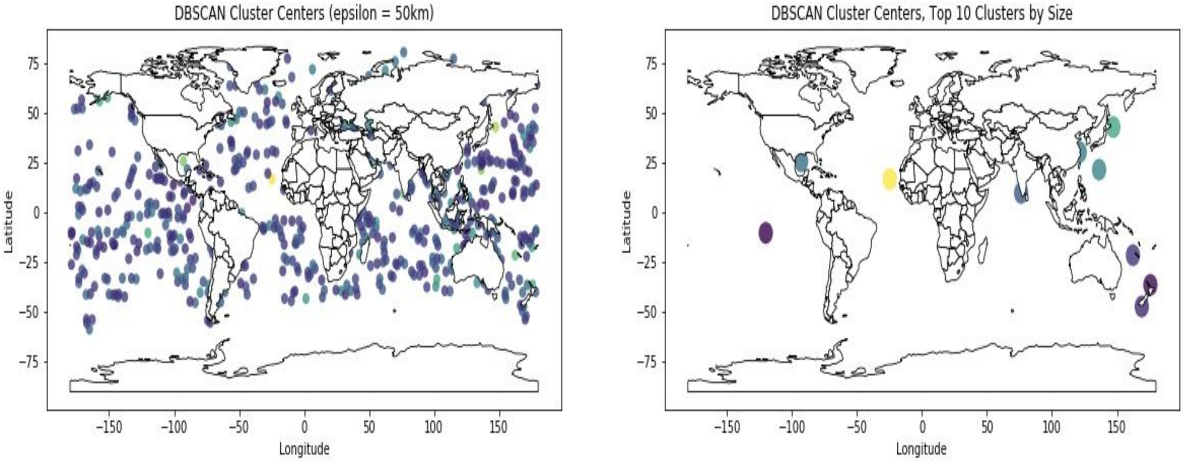

DBSCAN Cluster Centers (epsilon = 50km)
Latitude
Longitude
75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150
DBSCAN Cluster Centers, Top 10 Clusters by Size
Latitude
Longitude

Fig 5B: Cluster are colored based on intensity / size (left). On the right, the largest clusters, which indicate the highest concentration of fishing vessel activity. The locations are consistent with domain knowledge about where significant global fishing activity occurs.

Trajectory Clustering Pattern Detection

In this section trajectory clustering results are reported for a sample of 125 boats, containing 21,470 unique trip trajectories. A sample of the full data set was used to improve processing time and to reduce noise in the visualizations.

DBCAN Results:

The DBSCAN trajectory clustering model results demonstrate the strengths and weaknesses associated with this approach. The model was highly sensitive to the specification of epsilon, often generating one large cluster that included all trajectories if epsilon was too high and generating too many clusters when epsilon was too small and categorizing a significant portion of the data as outlier data points. As can be seen in Figure 6, DBSCAN generally categorized trans-oceanic trips as outliers, and formed clusters in coastal areas with high levels of trip traffic.

Figure 6: DBSCAN Trajectory Clusters, Global View

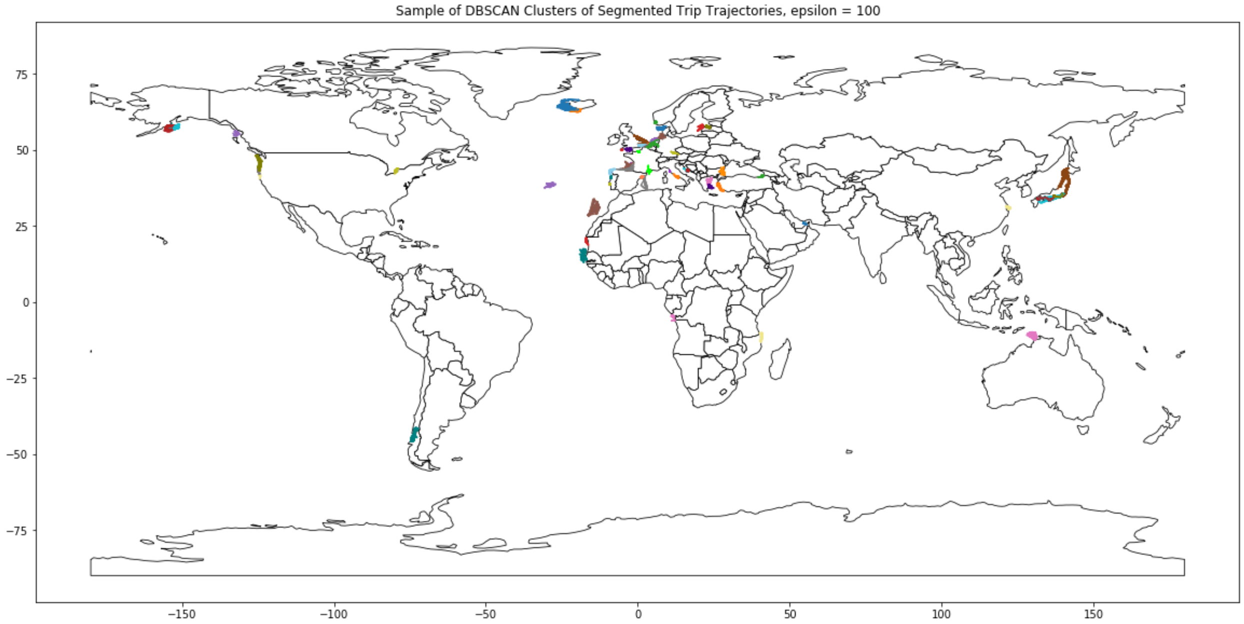

Sample of DBSCAN Clusters of Segmented Trip Trajectories, epsilon = 100
75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150

Fig 6 shows the results from DBSCAN trajectory clustering at a global level. The DBSCAN method deals well with trajectories clusters that are closely located (usually around coastal areas), however, much of the activity in international waters is not captured in clusters and shows up as noise.

Figure 7 shows the same clusters, but at a regional level for Europe. As can be seen, DBSCAN is able to identify unique clusters of trip activity, but the clusters are highly localized and tend to correspond only with coastal fishing trips, which is intuitive since there are a large number of these trips and they occur in a small geographic area. At high levels of spatial granularity the DBSCAN model is effective at identifying unique trip trajectory activity, such as separating unique trajectory activity around Japan (illustrated in Figure 8)[31]. The results from the DBSCAN model suggest that it may be a good model choice at a small spatial scale, but not as effective at clustering trajectories at a global level.

Figure 7. DBSCAN Trajectory Clustering Results, Regional View

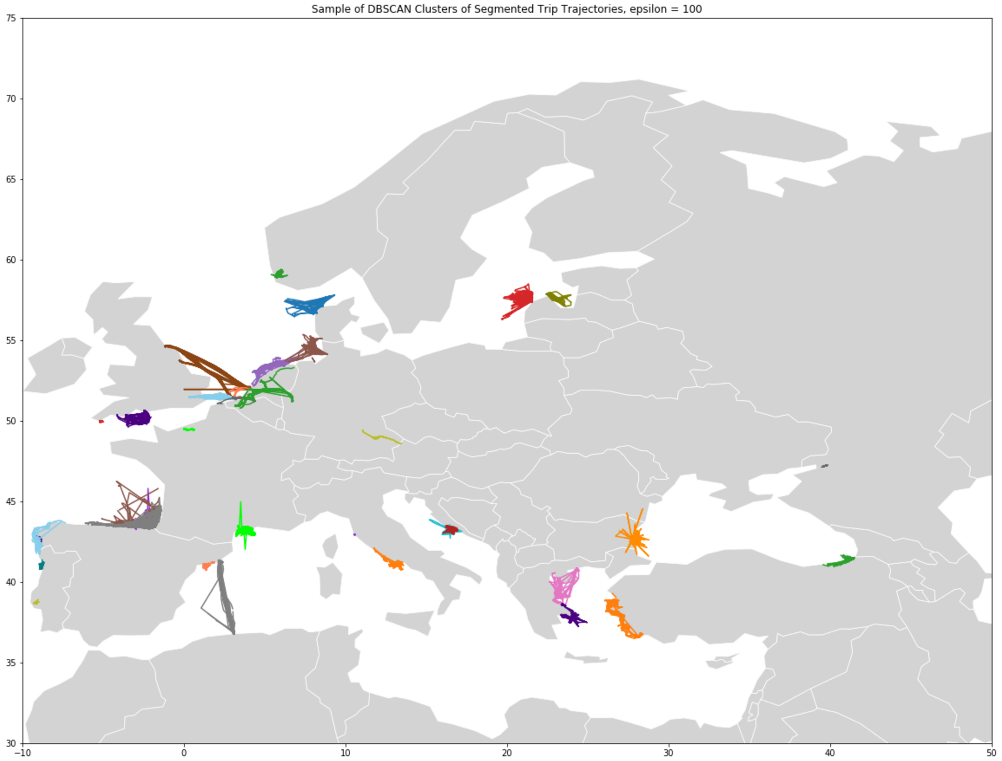

Sample of DBSCAN Clusters of Segmented Trip Trajectories, epsilon = 100
75
70
65
60
55
50
45
40
35
30
−10
0
10
20
30
40
50

Fig 7 Shows DBSCAN Trajectory Clustering Results at Regional level. This figure shows the strengths of DBSCAN at generating clusters with granular differences in trip activity, however, it shows the limitations, where the model fails to segment longer trips and categorizes longer trips as outliers.

[31] For another example see Item 4 in the Appendix.

Figure 8: DBSCAN Clustering Results, Detailed View

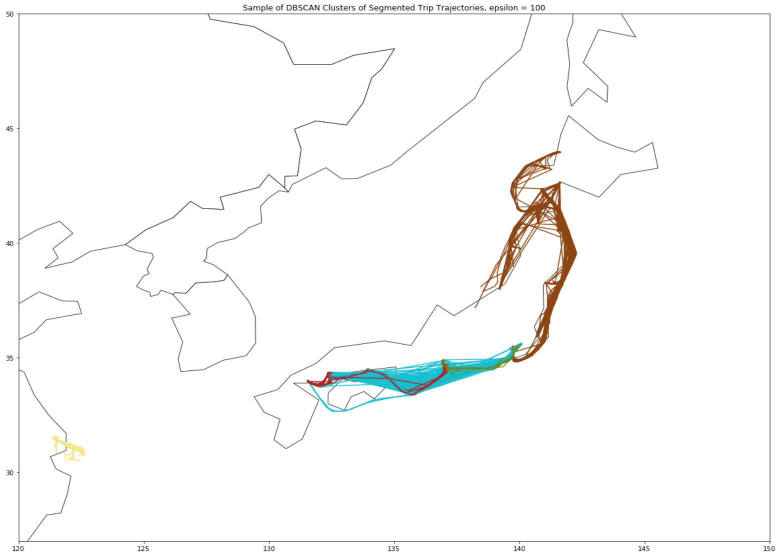

Sample of DBSCAN Clusters of Segmented Trip Trajectories, epsilon = 100
50
45
40
35
30
120
125
130
135
140
145
150

Fig. 8 shows DBSCAN results around Japan. As can be seen in the figure, DBSCAN method does well at high levels of spatial granularity and is able to cluster unique trajectories.

K-Means Results:

To select the optimal number of clusters for the K-means model, a silhouette score analysis[32] was performed (figure 9) and the number of clusters was set at 400, since gains in the average silhouette score after this level are marginal. The results for K-means suggest the model is robust to both local and global trajectory patterns.compared to the DBSCAN results.

The global results are displayed in Figure 10 and illustrate the ability of the model to distinguish between a wide range of complex trajectory patterns. An additional feature of the K-means model is that it was able to effectively identify outlier trajectories, which was used to remove noise and trajectories where there were errors due to interpolation or GPS sensor error, allowing the model to focus on repeat patterns and large clusters but retaining the ability to perform outlier detection if needed[33].

[32] Silhouette score measures how similar objects are to their own cluster vs. the dissimilarity to other cluster groups.

[33] Note, small clusters (fewer than 2 trips) were excluded from graphs in the results section for the purposes of simplifying the visualization

Figure 9: Silhouette Score Plot for K-Means Trajectory Clustering

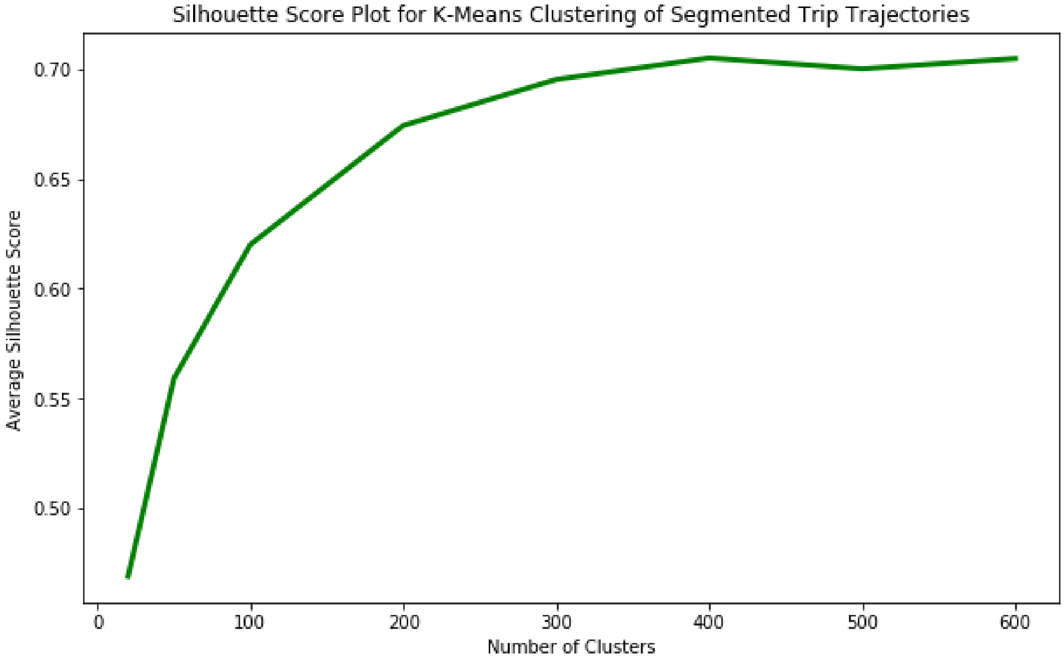


Figure 9 shows the average silhouette score vs. the number of clusters in the model. The silhouette scores are relatively high, and k = 400 is chosen based on the results of the testing.

Figure 10: K-means Trajectory Clustering Results

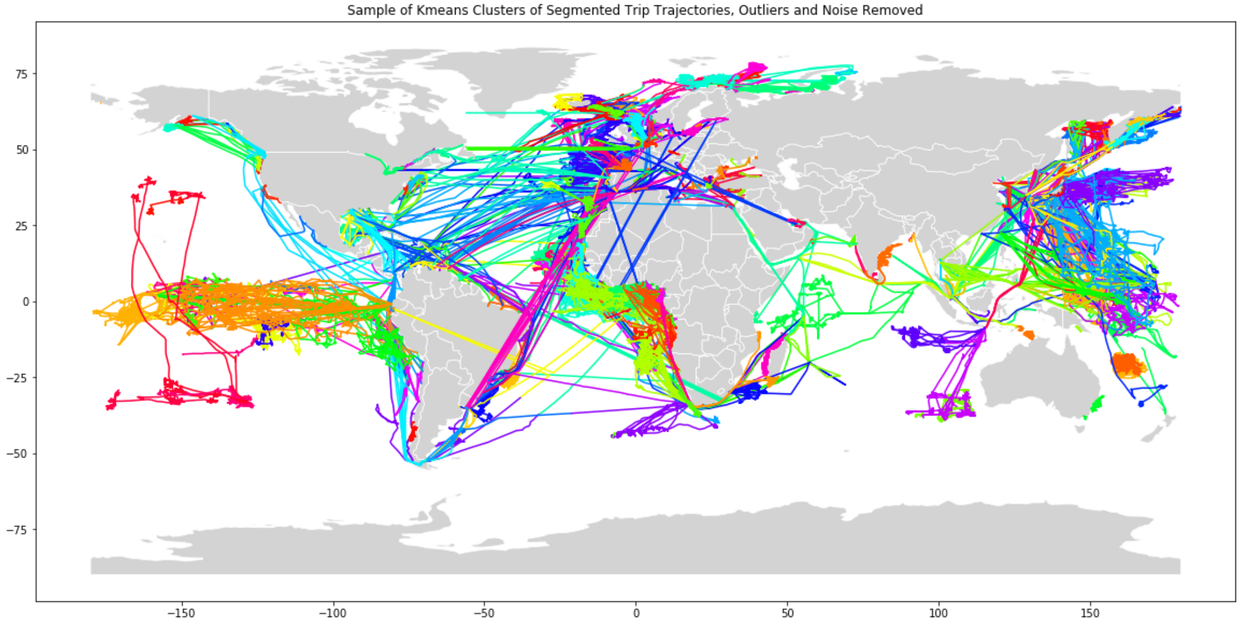

Sample of Kmeans Clusters of Segmented Trip Trajectories, Outliers and Noise Removed
75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150

Fig 10 shows the results for K-means trajectory clustering where k=400. The model effectively identified outliers and grouped them into small cluster of size 1 or 2, which are excluded from this graph

At a global level the K-means results allow us to detect repeat patterns in vessel trajectories that can be used to identify areas of significant fishing effort (Figure 11). The K-means results go beyond the hotspot clustering analysis by identifying specific trajectory movement patterns, rather than just hotspot zones.

Figure 11: K-Means Results: Largest Clusters by Number of Trips

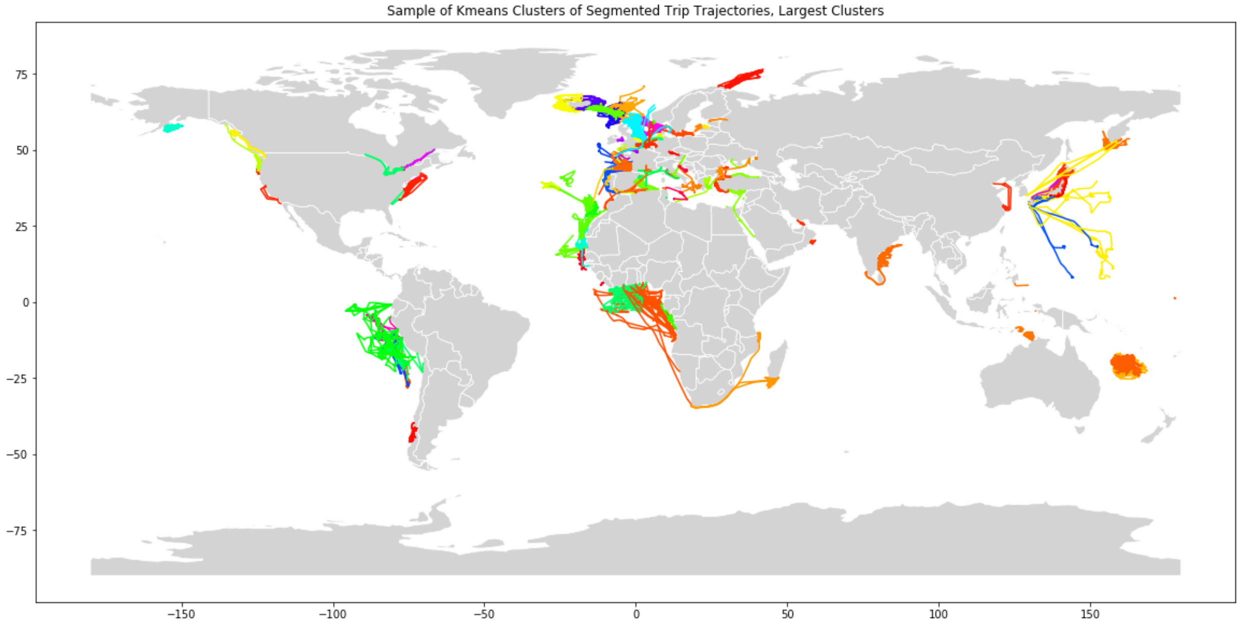

Sample of Kmeans Clusters of Segmented Trip Trajectories, Largest Clusters
75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150

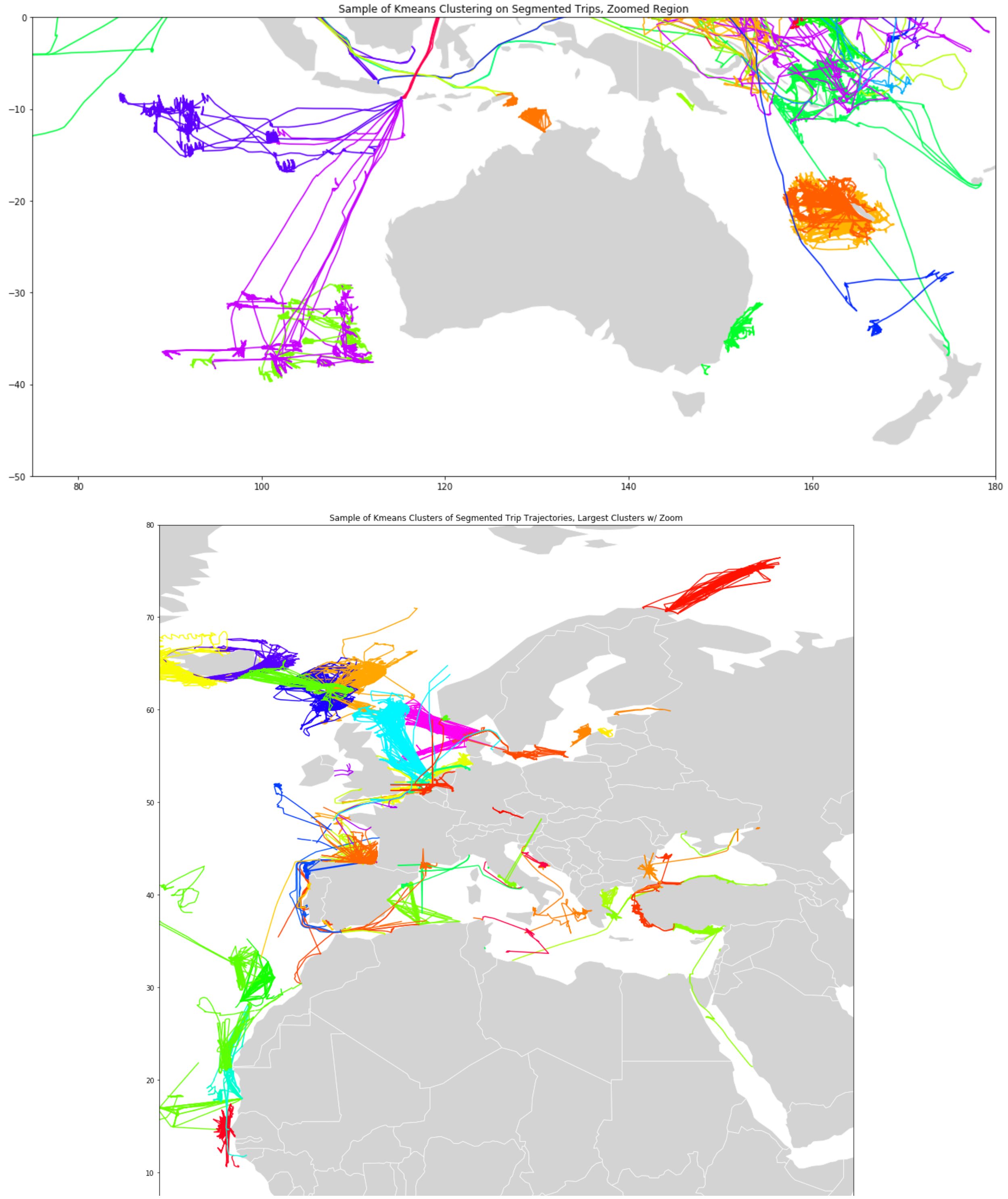
Fig 11 shows the location of the 75 largest clusters identified by the K-Means model.

At a more localized level, the K-means model is able to segment complex trajectory differences of varying shapes and paths (as illustrated in Figure 12). The model is able to separate trajectories that may cross over one another (and therefore share some similarities in terms of location) but are distinct from one another. For example, a trajectory that crosses over another at a perpendicular angle or a splitting two trajectories into separate clusters when they may share a similar start but bifurcate part of the way through the trajectory.

The primary limitation of the K-means model stem from two issues: 1) Processing time complexity is not favorable in very high dimensional space, and can be as bad as O(n^(k+2/p)) where *n* is the number of samples and *p* is the number of features[34]. 2) K-means results depend on the interpolation method chosen and when linear interpolation is not sufficient to represent a trajectory accurately, k-means is not able to correct or control for this which may create noisy clusters. The first issue in particular, means this approach may not be practical for handling larger (i.e. a full AIS data set) without significant downsampling and pre-processing.

[34] Arthur and Vassilvitskii (2006)

Figure 12: K-Means Results: Regional Views



Sample of Kmeans Clustering on Segmented Trips, Zoomed Region

Sample of Kmeans Clusters of Segmented Trip Trajectories, Largest Clusters w/ Zoom

Figure 12 shows K-Means results for a zoomed region. The results show the ability of the model to distinguish complex differences in trajectory patterns and shows the approach is robust to trajectories of different shapes and sizes.

Figure 13: Examples of Trips Grouped into Clusters

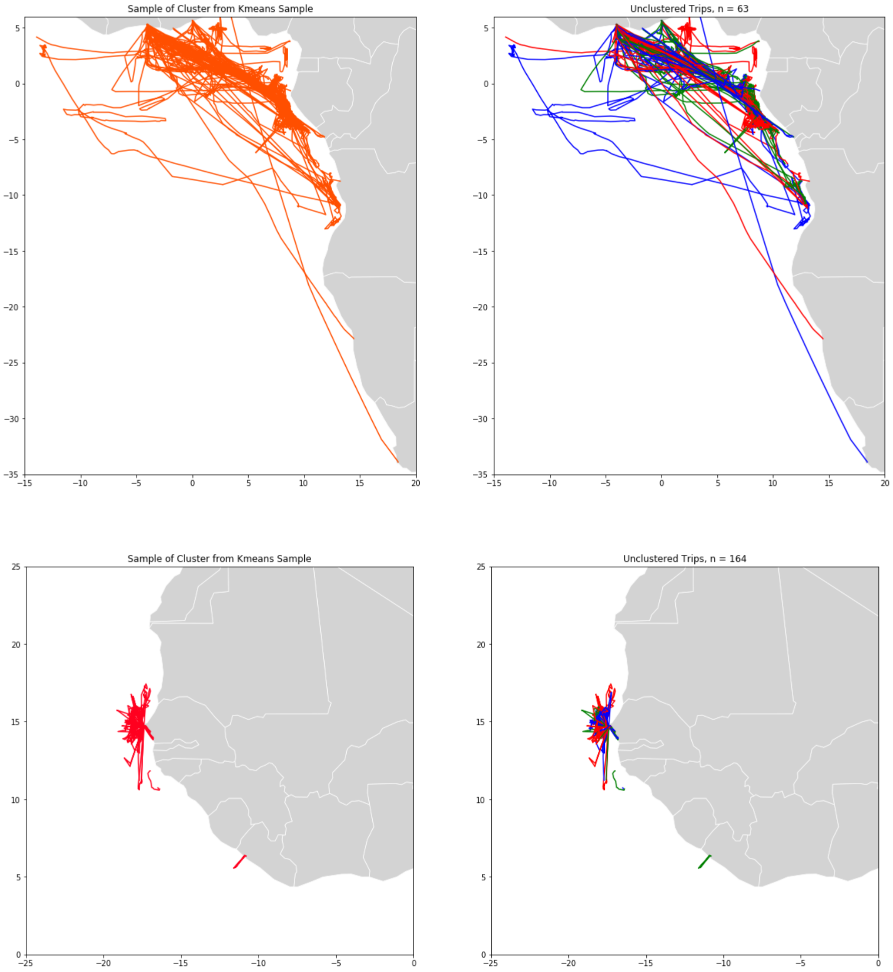

Sample of Cluster from Kmeans Sample
5
0
−5
−10
−15
−20
−25
−30
−35
−15
−10
−5
0
5
10
15
20
Unclustered Trips, n = 63
Sample of Cluster from Kmeans Sample
25
20
15
10
5
0
−25
−20
−15
−10
−5
0
Unclustered Trips, n = 164

Figure 13 illustrates the grouping of segmented trip trajectories into clusters.

Vessel Classification From Trajectory Features

The Random Forest trajectory classification model builds on the trajectory clustering analysis by predicting the type of vessel based on a trips' trajectory features. There are 12 classes of vessels to predict, with relatively balanced class size (see Fig 14).

Hyperparameters were tuned, starting with the number of estimators for the ensemble method. The cross validation suggested that 145 estimators achieved the highest balanced accuracy, after which there are only marginal increases in accuracy. Testing at 145 estimators suggested their might be overfitting, with training set accuracy of approximately 0.99 and test accuracy around 0.9. To control for overfitting risk, an additional cross validation was performed on max tree depth, the minimum number of samples to split on, and the minimum sample leaf. These parameters have the effect of mitigating overfitting and so overall model accuracy is lower but result in a more robust model and comparable accuracy results across training, test, and validation sets (see Table 1).

Figure 14: Counts of Trips by Vessel Class

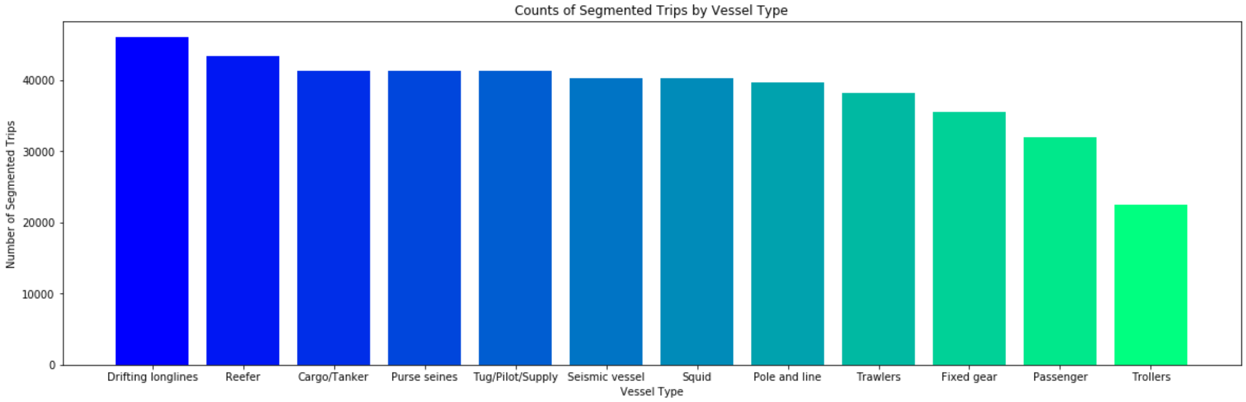


Figure 15: Balanced Accuracy vs. Number of Estimators

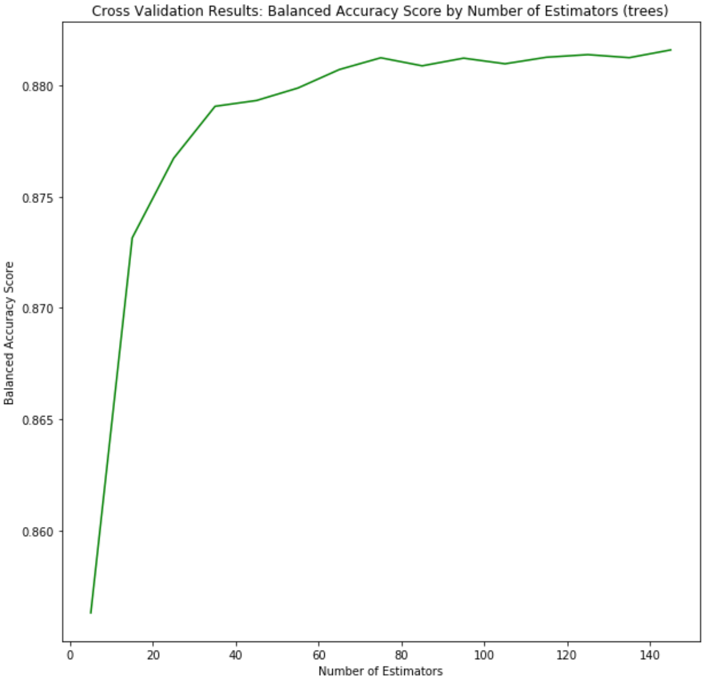
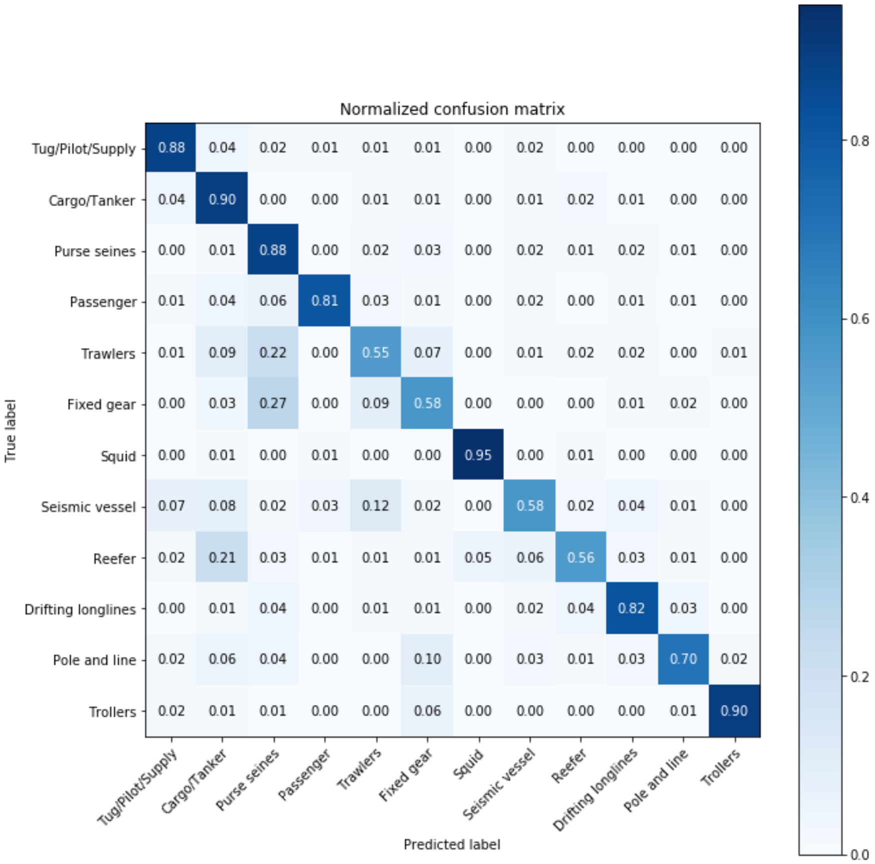


Table 1: Aggregate Accuracy Scores for Vessel Classification

	Weighted Avg. Accuracy (F1-score)	Precision	Recall
Training Data Set	0.794	0.800	0.796
Test Data Set	0.790	0.796	0.792
Validation Data Set	0.789	0.796	0.791

Overall the model achieves weighted average accuracy of 0.79 across all classes and has a similar performance for both precision (true positives divided by true positive and false positives) and recall (true positive divided by true positive and false negative). At a class level, the model performs with higher accuracy on some classes (e.g. Tug, Squid, and Trollers) than others (e.g. Trawlers, Reefers) suggesting that there may be more contextual features to extract from trajectories to improve model performance (see Fig 16 for a complete confusion matrix of the validation set)[35].

Figure 16. Confusion Matrix Example (Validation set)



	Tug/Pilot/Supply	Cargo/Tanker	Purse seines	Passenger	Trawlers	Fixed gear	Squid	Seismic vessel	Reefer	Drifting longlines	Pole and line	Trollers
Tug/Pilot/Supply	0.88	0.04	0.02	0.01	0.01	0.01	0.00	0.02	0.00	0.00	0.00	0.00
Cargo/Tanker	0.04	0.90	0.00	0.00	0.01	0.01	0.00	0.01	0.02	0.01	0.00	0.00
Purse seines	0.00	0.01	0.88	0.00	0.02	0.03	0.00	0.02	0.01	0.02	0.01	0.00
Passenger	0.01	0.04	0.06	0.81	0.03	0.01	0.00	0.02	0.00	0.01	0.01	0.00
Trawlers	0.01	0.09	0.22	0.00	0.55	0.07	0.00	0.01	0.02	0.02	0.00	0.01
Fixed gear	0.00	0.03	0.27	0.00	0.09	0.58	0.00	0.00	0.00	0.01	0.02	0.00
Squid	0.00	0.01	0.00	0.01	0.00	0.00	0.95	0.00	0.01	0.00	0.00	0.00
Seismic vessel	0.07	0.08	0.02	0.03	0.12	0.02	0.00	0.58	0.02	0.04	0.01	0.00
Reefer	0.02	0.21	0.03	0.01	0.01	0.01	0.05	0.06	0.56	0.03	0.01	0.00
Drifting longlines	0.00	0.01	0.04	0.00	0.01	0.01	0.00	0.02	0.04	0.82	0.03	0.00
Pole and line	0.02	0.06	0.04	0.00	0.00	0.10	0.00	0.03	0.01	0.03	0.70	0.02
Trollers	0.02	0.01	0.01	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.01	0.90

A benefit of Random Forest Classifiers is the feature importance information that can be extracted from the model (Fig 16). The trip span measures (e.g. the latitude and longitude features) proved more important than either trip distance or speed. A limitation of the Random Forest model is that because it is an ensemble method, the interpretability of the feature importance is limited.

[35] See Appendix Item 5 for detailed class-level precision, accuracy, and recall metrics as well as confusion matrices for training, test, and validation sets.

Figure 17: Random Forest Classifier Results: Feature Importance

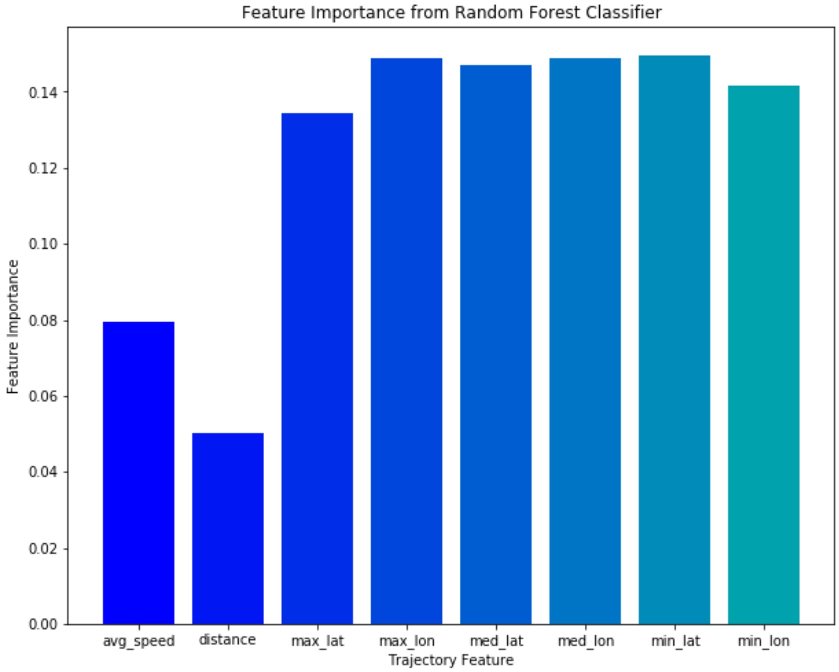


Fig 17 shows feature importance measures for the trajectory feature model inputs

Though the classifier performed well across major classification model metrics, further steps could be taken to extract more complex trajectory features to improve the classification model's performance. A clear next step for this research would be classifying the activity of the vessel at different points in the trajectory, specifically identifying when a boat is harvesting vs. traveling.

Conclusion

This research demonstrated the application of trajectory clustering methods on high frequency fishing vessel GPS data in an effort to identify repeat patterns of overfishing activity. The analysis focused on identifying hotspots of fishing activity using DBSCAN methods, trajectory clustering on segmented trips trajectories, and classification of boat type based on trajectory features.

Understanding fishing vessel behavior and patterns will continue to be an important for environmental stewardship and enforcement of trade and labor standards. Future work in this area should focus on improved interpolation techniques, given that large gaps in GPS readings can be common, especially on the high seas. As mentioned above, research on classification should focus on identifying vessel behavior so that precise locations and patterns of harvesting activity can be identified.

References

Akasapu, A. K., Sharma, L. K., and Ramakrishna, G. (2010). Efficient trajectory pattern mining for both sparse and dense dataset. *International Journal of Computer Applications*. 9(5), 45-48. https://pdfs.semanticscholar.org/de46/d74ebbaf6dac3576e7315e92b1bbe00c92c6.pdf

Alon, J., Sclaroff, S., Kollios, G., and Pavlovic, V. (2003). Discovering clusters in motion time series data. *In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 375–381. DOI:10.1109/cvpr.2003.1211378.

Arthur, D. and Vassilvitskii, S. (2006). How slow is the k-means method? *In SCG'06 Proceedings of the twenty-second annual symposium on computational geometry*, *ACM,* 144-153. DOI: 10.1145/1137856.1137880

Besse, P. C., B. Guillouet, J.-M. Loubes, and F. Royer. (2016) Review and perspective for distance-based clustering of vehicle trajectories. *IEEE Transactions on Intelligent Transportation Systems*, 17(11), 3306–3317.

Bian, J., Tian, D., Tang, Y., and Tao, D. (2018). A survey on trajectory clustering analysis. arXiv:1802.06971.

Birant, D. and Kut, A. (2007). St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1), 208-221. DOI:10.1016/j.datak.2006.01.013

Breiman, L. (2001). Random forests, *Machine Learning*, 45(1), 5-32.

Chen, Z., Yan, Y., and Ellis, T. (2014). Lane detection by trajectory clustering in urban environments. *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on. IEEE*.

Chen, J., Wang, R., Lui, L., and Song, J. (2011). Clustering of trajectories based on hausdorff distance. *Electronics, Communications and Control (ICECC), 2011 International Conference on. IEEE.* DOI: 10.1109/ICECC.2011.6066483

Chimwayi, K.B. and Anuradha, J. (2018). Clustering West Nile Virus spatio-temporal data using ST-DBSCAN. *Procedia Computer Science*, 132, 1218-1287. DOI: https://doi.org/10.1016/j.procs.2018.05.037

De Souza, E., Boerder, K., Matwin, S., and Worm, B. (2016). Improving Fishing Pattern Detection from Satellite AIS Using Data Mining and Machine Learning. *PLOS ONE. PLOS ONE*, 11(7): e0158248DOI: https://doi.org/10.1371/journal.pone.0158248

Dewan, P., Ganti R., and Srivatsa, M. (2017). SOM-TC: Self-organizing map for hierarchical trajectory clustering, *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, *IEEE*, 1042–1052. DOI: 10.1109/ICDCS.2017.244

Dobrkovic, A., Iacob M-E., Hillegersberg, J.V. (2018). Maritime Pattern Extraction and Route Reconstruction from Incomplete AIS Data. *International Journal of Data Science and Analytics*, 5(2-3), 111-136. DOI: https://doi.org/10.1007/s41060-017-0092-8

Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, *AAAI Press*, 226-231. https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Food and Agriculture Organization of the United Nations (FAO). (2016). *FAO Yearbook of Fishery and Aquaculture Statistics*. http://www.fao.org/fishery/static/Yearbook/YB2016_USBcard/booklet/web_i9942t.pdf

Gaffney, S. and Smyth, P. (1999). Trajectory clustering with mixture of regression models. *KDD' 99: Proc. Fifth International Conference on Knowledge Discovery and Data Mining (1999), ACM Press*, 63–72. DOI:10.1145/312129.312198.

Gudmundsson, J. and Kreveld, M. V. (2006). Computing longest duration flocks in trajectory data. *In Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems. ACM*, 35–42. DOI: 10.1145/1183471.1183479

Jeung, H., Shen, H., and Zhou, X. (2008). Convoy queries in spatio-temporal databases. *In Proceedings of the 24th IEEE International Conference on Data Engineering. IEEE*, 1457–1459. DOI: 10.1109/ICDE.2008.4497588

Kianfar, J. and Edara, P. (2013). A data mining approach to creating fundamental traffic flow diagram. *Procedia - Social and Behavioral Sciences* 104(2), 430-439. DOI: https://doi.org/10.1016/j.sbspro.2013.11.136

Lee, J.G., Han, J., and Whang, K. Y. (2007). Trajectory clustering: A partition-and group framework. *In Proc. 2007 ACM SIGMOD International Conference on Management of Data (2007), ACM Press*, 593–604. DOI:10.1145/1247480.1247546.

Li, Z., Ding, B, Jan, J., Kays, R., and Nye, P. (2010). Mining periodic behaviors for moving objects. I*n KDD '10: Proc. 16th ACM SIGKDD international conference on Knowledge discovery and data mining (2010), ACM Press*, 1099–1108. doi:10.1145/1835804.1835942.

Liu, S., Liu, Y., Ni, L. M., Fan, J., and Li, M. (2010). Towards mobility-based clustering. I*n Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10*, *ACM Press*, 919–928. DOI: 10.1145/1835804.1835920

Lloyd, S. P. (1982). Least squares quantization in PCM. *Information Theory, IEEE Transactions on 28.2*: 129-137.

Marine and Inland Fisheries Service, Fisheries and Aquaculture Resources Use and Conservation Division. FAO Fisheries and Aquaculture Department. (2011). Review of the state of world marine fishery resources. *FAO FISHERIES AND AQUACULTURE TECHNICAL PAPER. No. 569. Rome, FAO*.

Mazimpaka, J. D., and Timpf, S. (2016). Trajectory data mining: A review of methods and Applications. *Journal of Spatial Information Science,* 13, 61-99. DOI: 10.5311/JOSIS.2016.13.263

Orellana, D., Bregt, A. K., Ligtenberg, A., and Wachowicz, M. (2012). *Exploring visitor movement patterns in natural recreational areas. Tourism Management* 33(3), 672–682. DOI:10.1016/j.tourman.2011.07.010.

Ossama, O., Moktar, H., and El-Sharkawi, M. (2011). An extended k-means technique for clustering moving objects. *Egyptian Informatics Journal*, 12(1), 45-51. DOI: https://doi.org/10.1016/j.eij.2011.02.007

Palma, A. T., , Bogorny, V., Kuijpers, B., AND Alvares, L. O. (2008). A clustering based approach for discovering interesting places in trajectories. *In SAC '08: Proc. 2008 ACM symposium on Applied computing*, *ACM*, 863–868. DOI:10.1145/1363686.1363886.

Pauly, D., and Zeller, D. (2016). Catch reconstructions reveal that global marine fisheries catches are higher than reported and declining. *Nature Communications,* 7, Article number: 10244.

Pokorny, F. T., Goldberg, K., and Kragic, D. (2016). Topological trajectory clustering with relative persistent homology. *In 2016 IEEE International Conference on Robotics and Automation (ICRA), IEEE*, 16–23. DOI: 10.1109/ICRA.2016.7487092

Shaw, A. A., and Gopalan, N.P. (2014). Finding frequent trajectories by clustering and sequential pattern mining. *Journal of Traffic and Transportation Engineering*, 1(6), 393-403.

Sung, C., Feldman, D., and Rus, D. (2012). Trajectory Clustering for Motion Prediction. *In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE*. DOI: 10.1109/IROS.2012.6386017

Timothy, A. Varshavsky, A. Lamarca, Chen, M. Y., and Chounhury, T. (2006). Mobility detection using everyday GSM traces. *In Proceedings of the 8th International Conference on Ubiquitous Computing. ACM*, 212–224. DOI: https://doi.org/10.1007/11853565_13

Vincenty, T. (1975). "Direct and Inverse Solutions of Geodesics on the Ellipsoid with application of nested equations". *Survey Review. XXIII*, 176, 88–93.

Wikipedia contributors. "Haversine Formula." *Wikipedia, The Free Encyclopedia*, July, 2019. https://en.wikipedia.org/wiki/Haversine_formula

Worm, B., Barbier, E. B., Beaumont, N., Duffy, J.E, Folke, C., Halpern, B.S., Jackson, J.B.C., Lotze, H.K., Micheli, F., Paumbi, S.R., Sala, E., Selkoe, K.A., Stachowicz, J.J., and Watson, R.

(2006). Impacts of biodiversity loss on ocean ecosystem services. *Science*, 314(5800), 787-790. DOI: 10.1126/science.1132294

Yao, D., Zhang, C., Zhu, Z., Huang, J. and Bi, J. (2017) .Trajectory clustering via deep representation learning, *In 2017 International Joint Conference on Neural Networks (IJCNN), IEEE,* 3880–3887. DOI: 10.1109/IJCNN.2017.7966345

Yuan, G., Sun, P., Zhao, J., Li, D., and Wang, C. (2017). A review of moving object trajectory clustering algorithms. *Artificial Intelligence Review*, 47(1), 123–144. DOI: https://doi.org/10.1007/s10462-016-9477-7

Zhang, D., Lee, K., and Lee, I. (2018). Hierarchical trajectory clustering for spatio-temporal periodic pattern mining. *Expert Systems with Applications, 9*2: 1-11. DOI: https://doi.org/10.1016/j.eswa.2017.09.040

Zheng, Y. (2015). Trajectory Data Mining: An Overview. *ACM Trans. On Intelligent Systems and Technology*, 6(3), 1-41. DOI: http://dx.doi.org/10.1145/2743025

Zhu, Y., Zheng, Y., Zhang, L., Santani, D., Xie, X. and Yang, Q. (2012). Inferring Taxi Status using GPS Trajectories. Technical Report MSR-TR-2011-144. https://arxiv.org/abs/1205.4378v2

Programming References

Jones, E., Oliphant, T., Peterson, P., & others. (2001). *SciPy: Open source scientific tools for Python*. Retrieved from "http://www.scipy.org/"

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Oliphant, T. E. (2006). *A guide to NumPy* (Vol. 1). Trelgol Publishing USA.

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Appendix

Item 1: AIS GPS Data Sample

Table 2: AIS GPS Data Sample

MMSI	Timestamp	Latitude	Longitude
170052967564863	2012-01-04 00:37:06	59.092514	10.914706
170052967564863	2012-01-04 01:00:57	59.092552	10.914706

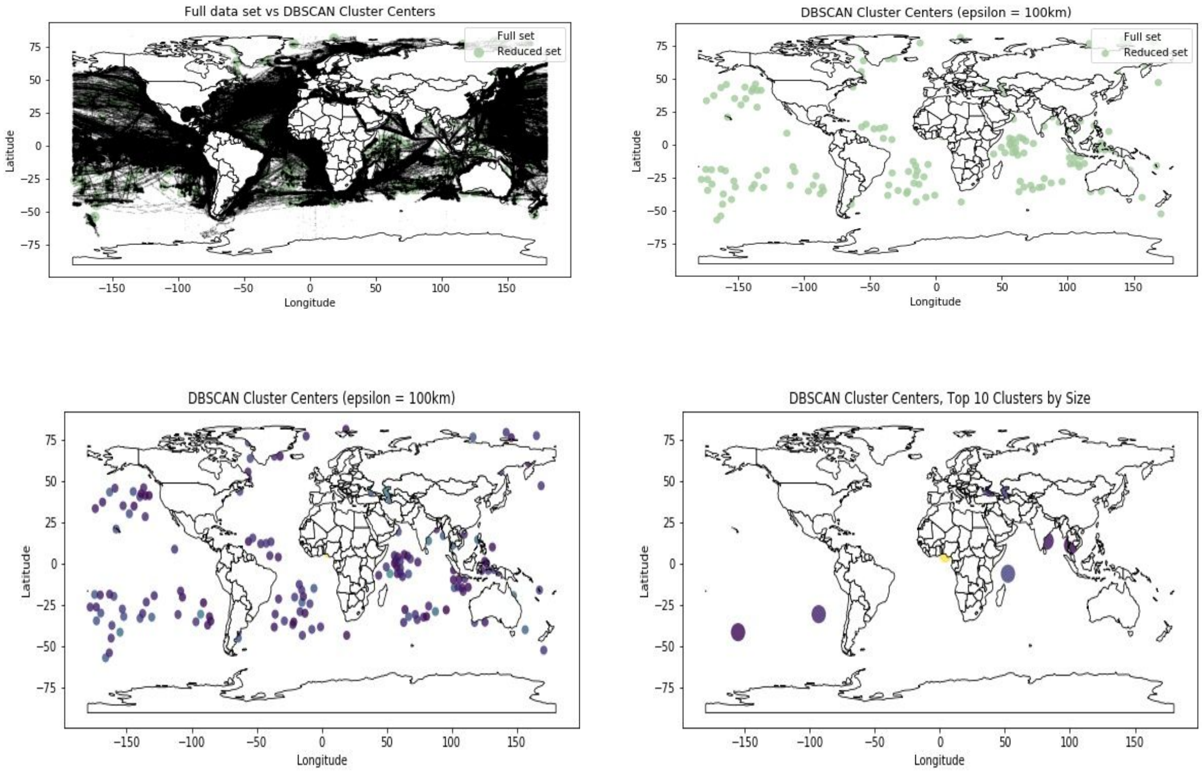
Item 2: Vessel Classes and Labels Used in Modeling

Table 3: Vessel Types and Class Labels

Class	Class Label
Tug/Pilot/Supply	**1.0**
Cargo/Tanker	**2.0**
Purse seines	**3.0**
Passenger	**4.0**
Trawlers	**5.0**
Fixed gear	**6.0**
Squid	**7.0**
Seismic vessel	**8.0**
Reefer	**9.0**
Drifting longlines	**10.0**
Pole and line	**11.0**
Trollers	**12.0**

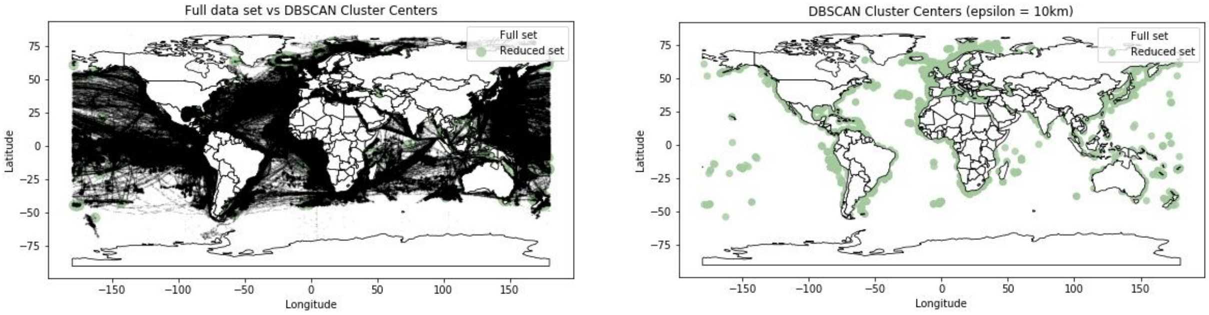
Item 3: Additional DBSCAN Hotspot Clustering Results

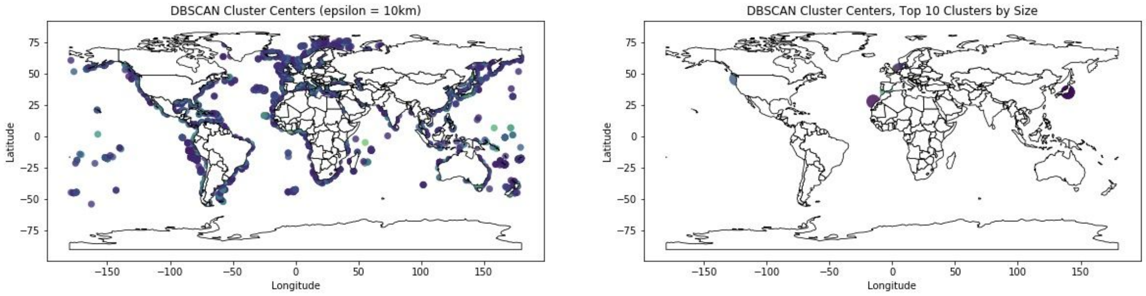
Figure 18: DBSCAN Clustering Results for Epsilon = 100km


Full data set vs DBSCAN Cluster Centers
Full set
Reduced set
DBSCAN Cluster Centers (epsilon = 100km)
Full set
Reduced set
DBSCAN Cluster Centers (epsilon = 100km)
DBSCAN Cluster Centers, Top 10 Clusters by Size
Latitude
Longitude
75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150

Epsilon = 100km, # of clusters = 158; Outliers: 68,865

Figure 19: DBSCAN Results for Epsilon = 10km




DBSCAN Cluster Centers (epsilon = 10km)
Latitude
Longitude
75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150
DBSCAN Cluster Centers, Top 10 Clusters by Size
Latitude
Longitude

# of clusters = 5,761, Identified outliers = 3,498,162

Item 4: DBSCAN Trajectory Clustering Example:

Figure 20: Example of Clustered Trajectories and Individual Trips

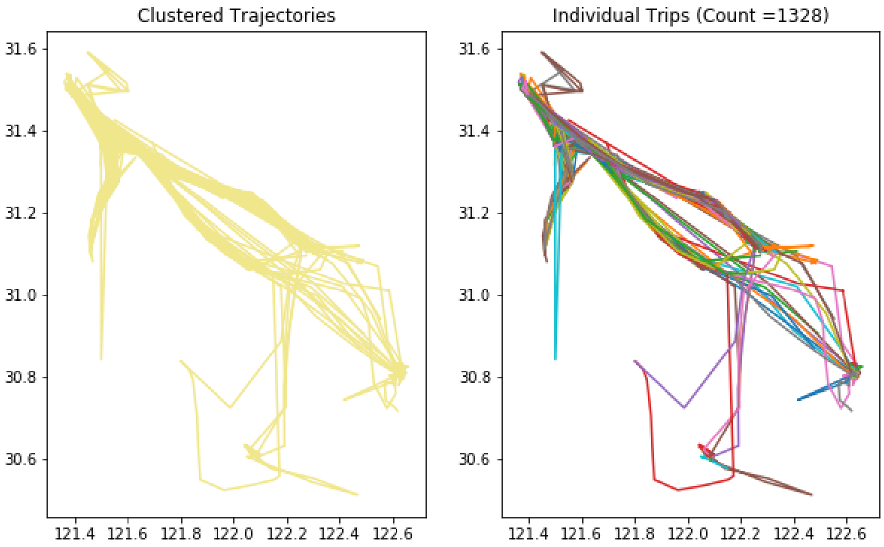
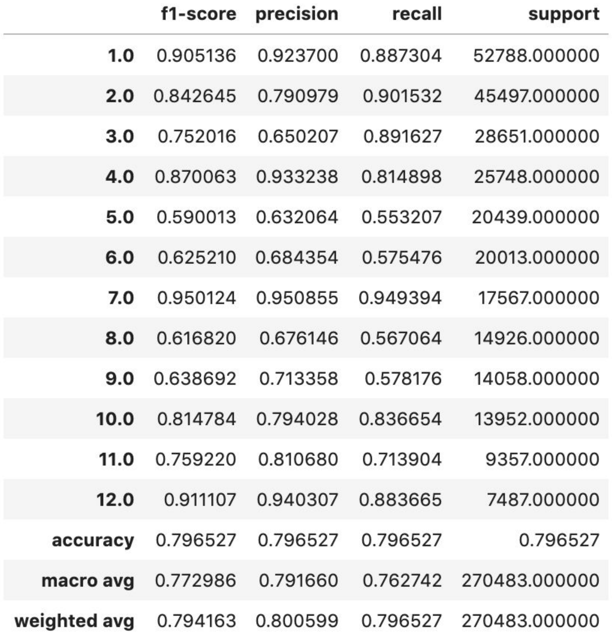


Figure 19 illustrates an example of a clustered created by the DBSCAN model (left) and the underlying trips composed in the cluster (right)

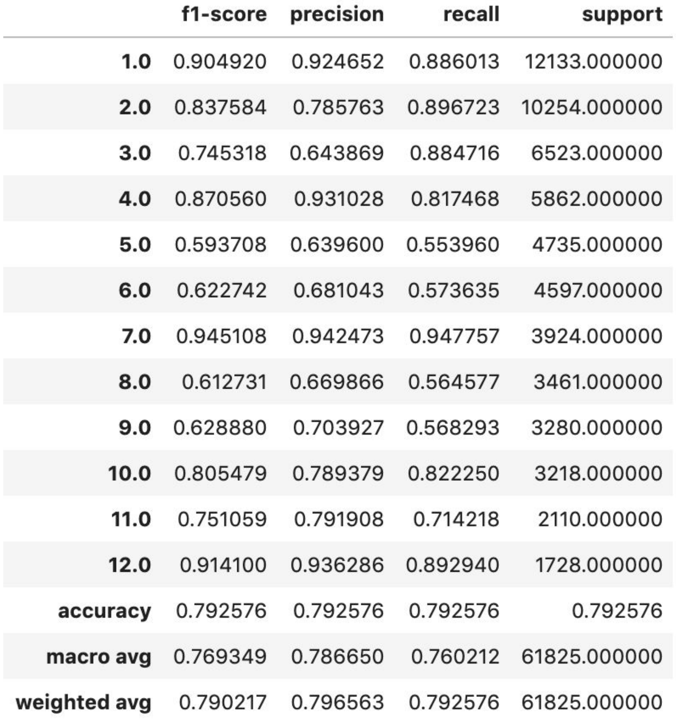
Item 5: Classification Model Detailed Tables

Table 4: Training Data Class-Level Evaluation Metrics



	f1-score	precision	recall	support
1.0	0.905136	0.923700	0.887304	52788.000000
2.0	0.842645	0.790979	0.901532	45497.000000
3.0	0.752016	0.650207	0.891627	28651.000000
4.0	0.870063	0.933238	0.814898	25748.000000
5.0	0.590013	0.632064	0.553207	20439.000000
6.0	0.625210	0.684354	0.575476	20013.000000
7.0	0.950124	0.950855	0.949394	17567.000000
8.0	0.616820	0.676146	0.567064	14926.000000
9.0	0.638692	0.713358	0.578176	14058.000000
10.0	0.814784	0.794028	0.836654	13952.000000
11.0	0.759220	0.810680	0.713904	9357.000000
12.0	0.911107	0.940307	0.883665	7487.000000
accuracy	0.796527	0.796527	0.796527	0.796527
macro avg	0.772986	0.791660	0.762742	270483.000000
weighted avg	0.794163	0.800599	0.796527	270483.000000

Table 5: Test Data Class-Level Evaluation Metrics

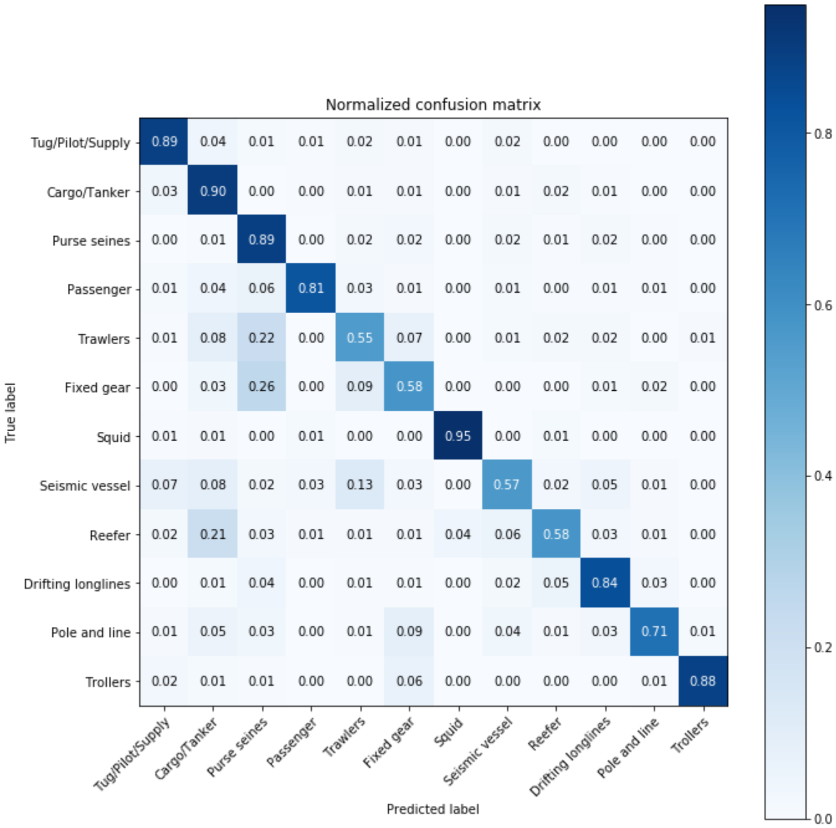


	f1-score	precision	recall	support
1.0	0.904920	0.924652	0.886013	12133.000000
2.0	0.837584	0.785763	0.896723	10254.000000
3.0	0.745318	0.643869	0.884716	6523.000000
4.0	0.870560	0.931028	0.817468	5862.000000
5.0	0.593708	0.639600	0.553960	4735.000000
6.0	0.622742	0.681043	0.573635	4597.000000
7.0	0.945108	0.942473	0.947757	3924.000000
8.0	0.612731	0.669866	0.564577	3461.000000
9.0	0.628880	0.703927	0.568293	3280.000000
10.0	0.805479	0.789379	0.822250	3218.000000
11.0	0.751059	0.791908	0.714218	2110.000000
12.0	0.914100	0.936286	0.892940	1728.000000
accuracy	0.792576	0.792576	0.792576	0.792576
macro avg	0.769349	0.786650	0.760212	61825.000000
weighted avg	0.790217	0.796563	0.792576	61825.000000

Table 6: Validation Data Class-Level Evaluation Metrics

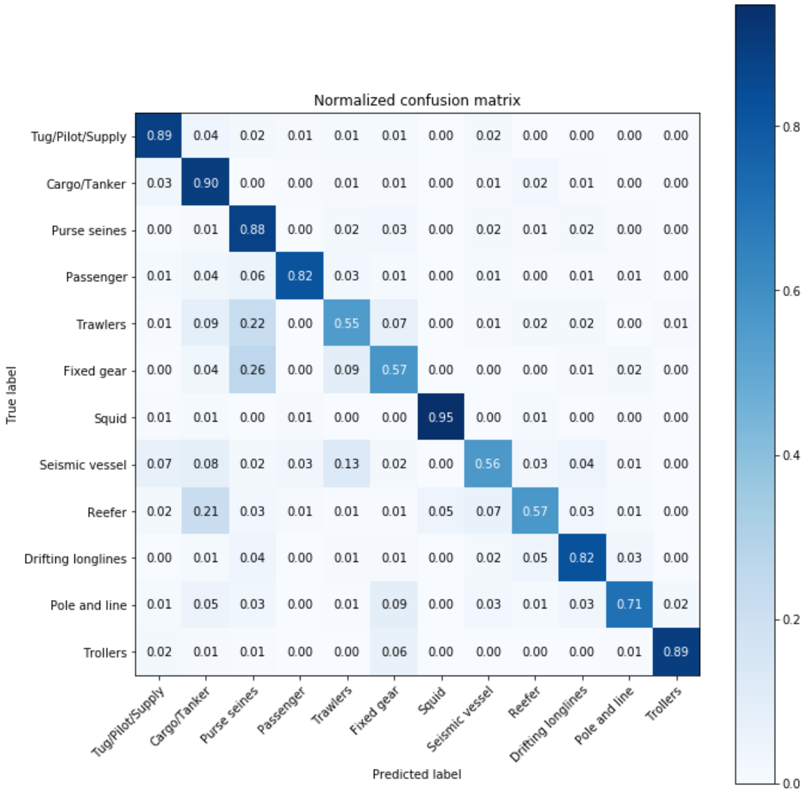
	f1-score	precision	recall	support
1.0	0.902285	0.921482	0.883871	7595.000000
2.0	0.836625	0.784877	0.895678	6432.000000
3.0	0.739971	0.637427	0.881834	4079.000000
4.0	0.867099	0.930470	0.811809	3709.000000
5.0	0.593939	0.641667	0.552821	2925.000000
6.0	0.625785	0.686386	0.575017	2946.000000
7.0	0.948017	0.945104	0.950948	2426.000000
8.0	0.621247	0.675261	0.575234	2140.000000
9.0	0.624618	0.704702	0.560878	2004.000000
10.0	0.806072	0.791727	0.820948	2005.000000
11.0	0.741619	0.782571	0.704740	1287.000000
12.0	0.912821	0.930608	0.895700	1093.000000
accuracy	0.791569	0.791569	0.791569	0.791569
macro avg	0.768341	0.786024	0.759123	38641.000000
weighted avg	0.789284	0.796019	0.791569	38641.000000

Figure 21: Training Data Confusion Matrix



	Tug/Pilot/Supply	Cargo/Tanker	Purse seines	Passenger	Trawlers	Fixed gear	Squid	Seismic vessel	Reefer	Drifting longlines	Pole and line	Trollers
Tug/Pilot/Supply	0.89	0.04	0.01	0.01	0.02	0.01	0.00	0.02	0.00	0.00	0.00	0.00
Cargo/Tanker	0.03	0.90	0.00	0.00	0.01	0.01	0.00	0.01	0.02	0.01	0.00	0.00
Purse seines	0.00	0.01	0.89	0.00	0.02	0.02	0.00	0.02	0.01	0.02	0.00	0.00
Passenger	0.01	0.04	0.06	0.81	0.03	0.01	0.00	0.01	0.00	0.01	0.01	0.00
Trawlers	0.01	0.08	0.22	0.00	0.55	0.07	0.00	0.01	0.02	0.02	0.00	0.01
Fixed gear	0.00	0.03	0.26	0.00	0.09	0.58	0.00	0.00	0.00	0.01	0.02	0.00
Squid	0.01	0.01	0.00	0.01	0.00	0.00	0.95	0.00	0.01	0.00	0.00	0.00
Seismic vessel	0.07	0.08	0.02	0.03	0.13	0.03	0.00	0.57	0.02	0.05	0.01	0.00
Reefer	0.02	0.21	0.03	0.01	0.01	0.01	0.04	0.06	0.58	0.03	0.01	0.00
Drifting longlines	0.00	0.01	0.04	0.00	0.01	0.01	0.00	0.02	0.05	0.84	0.03	0.00
Pole and line	0.01	0.05	0.03	0.00	0.01	0.09	0.00	0.04	0.01	0.03	0.71	0.01
Trollers	0.02	0.01	0.01	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.01	0.88

Figure 22: Test Data Confusion Matrix



	Tug/Pilot/Supply	Cargo/Tanker	Purse seines	Passenger	Trawlers	Fixed gear	Squid	Seismic vessel	Reefer	Drifting longlines	Pole and line	Trollers
Tug/Pilot/Supply	0.89	0.04	0.02	0.01	0.01	0.01	0.00	0.02	0.00	0.00	0.00	0.00
Cargo/Tanker	0.03	0.90	0.00	0.00	0.01	0.01	0.00	0.01	0.02	0.01	0.00	0.00
Purse seines	0.00	0.01	0.88	0.00	0.02	0.03	0.00	0.02	0.01	0.02	0.00	0.00
Passenger	0.01	0.04	0.06	0.82	0.03	0.01	0.00	0.01	0.00	0.01	0.01	0.00
Trawlers	0.01	0.09	0.22	0.00	0.55	0.07	0.00	0.01	0.02	0.02	0.00	0.01
Fixed gear	0.00	0.04	0.26	0.00	0.09	0.57	0.00	0.00	0.00	0.01	0.02	0.00
Squid	0.01	0.01	0.00	0.01	0.00	0.00	0.95	0.00	0.01	0.00	0.00	0.00
Seismic vessel	0.07	0.08	0.02	0.03	0.13	0.02	0.00	0.56	0.03	0.04	0.01	0.00
Reefer	0.02	0.21	0.03	0.01	0.01	0.01	0.05	0.07	0.57	0.03	0.01	0.00
Drifting longlines	0.00	0.01	0.04	0.00	0.01	0.01	0.00	0.02	0.05	0.82	0.03	0.00
Pole and line	0.01	0.05	0.03	0.00	0.01	0.09	0.00	0.03	0.01	0.03	0.71	0.02
Trollers	0.02	0.01	0.01	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.01	0.89