An Analysis of Maritime Navigational Warnings

Spatio-Temporal Data Analysis and Data Mining

Final Project, 2019

UCL

Authors: Katherine Jamieson, John R. Hoopes and Kristian Lunow Nielsen

All code available: https://github.com/robisoniv/navwarning-analysis

Word Count (excluding headers, captions, tables etc.): 6660

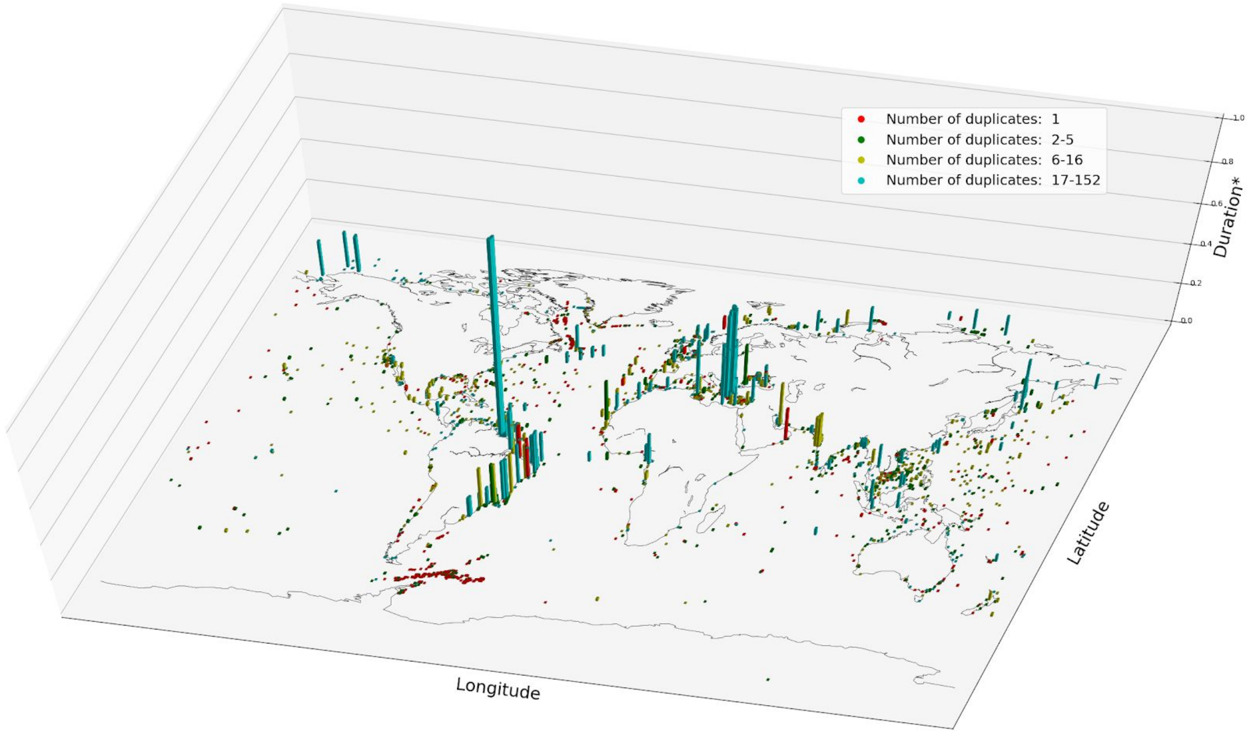
Introduction and data description

Mariners have braved the unknown for millennia, putting to sea to explore, trade, harvest, conquer and travel. The maritime environment presents a complex set of interacting and emergent risks including rocks, shoals and shallows, debris adrift, human threats such as pirates and attackers, extreme weather, earth events such as tsunamis. Historically constraints to scientists' understanding of the causes of such events, along with the lack of communication technology, meant that mariners putting to sea had to submit to the unknown and respond to threats as they arose. Needless to say, many ships have been lost, and many people have died, at sea.

Advancements in forecasting and communications technologies have done much to mitigate this risk. Captains now receive communications at sea regarding nearby risks, enabling them to respond well before a risk could be detected by those on board. Further, the World-Wide Navigational Warning System (WWNWS) disseminates maritime security information to ship captains at sea to warn of immediate or on-going global waters. (Human Environment and Transport Inspectorate 2019). These navigational warnings (navwarnings) present information about natural and human hazards in the maritime space and are of vital importance for shipping, military or any nautical based organizations that navigate the oceans, see figure-1 for the extent of the dataset.

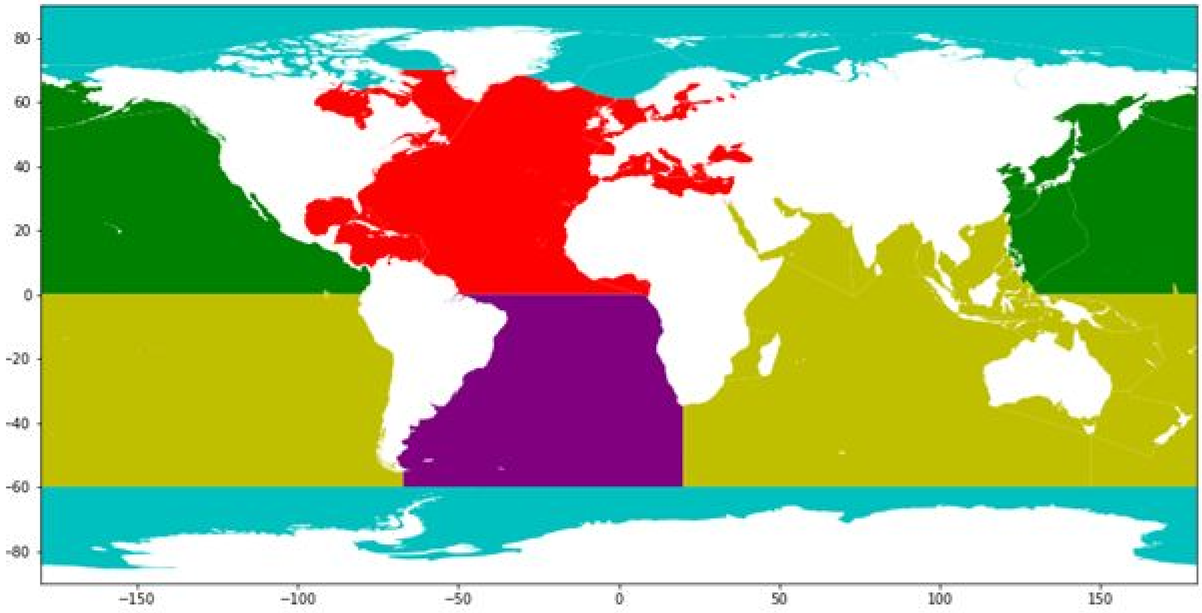
Our analysis will focus on utilizing nautical navigational warnings to conduct a risk assessment of the oceans through methodologies learned in the course. The dataset chosen for analysis was collected from the National Geospatial-Intelligence Agency (2019).

Figure-1: ***Extent of the dataset***



*: Duration is relative to the maximum duration of 834 days

Figure-2: ***Navigation Warning Limits***



The navigational warnings are divided into 5 regional areas across the globe. They are labelled as below:

1. navareaIV (in red)
2. navareaXII (in green)
3. hydroLant (in purple)
4. hydroPac (in yellow)
5. hydroArc (in blue)

The methodologies used for this analysis include; Natural Language Processing, Clustering using DBSCAN and Kernel Density Estimations. Instead of a direct comparison between the three methods, we will conduct a thorough investigation of the data with a combination of the three, each building off of each other.

Exploratory spatio-temporal data analysis

The dataset forming the basis of this investigation was collected on 76 dates between 1 June 2018 to 5 October 2018. Over this 126 day period 53,956 observations representing individual navigational warnings were issued for mariners by NAVAREA Coordinators serving one of 21 geographic NAVAREAs (Mukhurhee 2017). To inform our analytical approach we calculated summary statistics globally and across regional groupings of NAVAREAs, and measured temporal and spatial autocorrelation.

Summary Statistics

Navigational warnings are issued daily to inform mariners at sea of current hazards. Some warnings span a number of days, meaning duplicate observations exist within the dataset. For temporal statistics we included the entire dataset as issued by the data provider, though for the purposes of calculating spatial summary statistics we opted to exclude duplicate observations, defined as warnings with identical centroid coordinates. This decision was made largely because duplicate points would affect the spatial analysis methods employed including clustering and kernel density estimation.

Temporal

Between 610 and 782 warnings were issued on each of the 76 dates represented in the dataset, with an observed mean of 709.95 warnings per day, and standard deviation of 45.534. Dates with no observations were excluded from our analysis; navwarnings are issued daily, and gaps were due to errors in data collection.

Figure-3: ***Frequency distribution - Daily warnings***

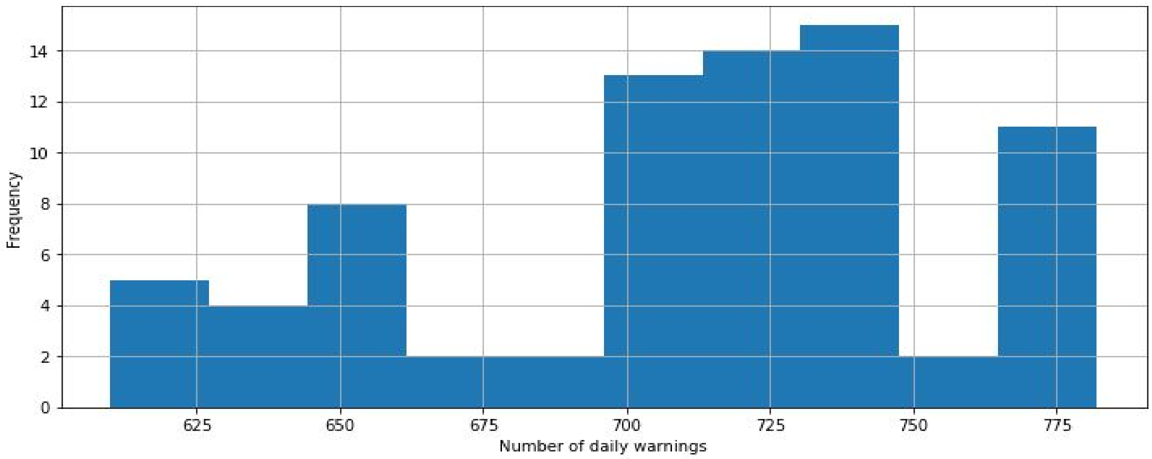


Figure-4: ***Daily warning counts by date***

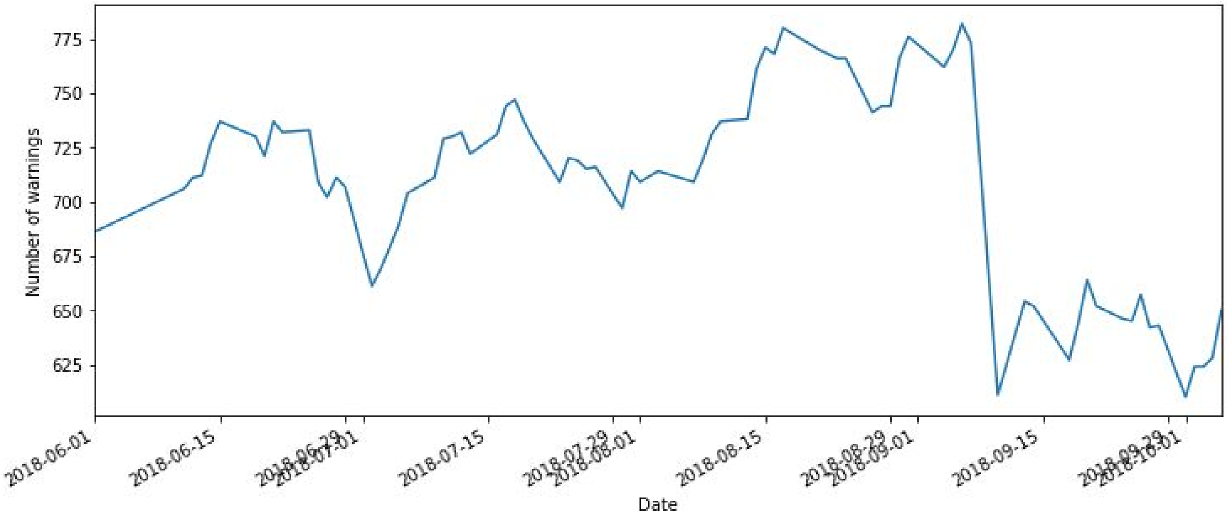


Figure-5: ***Autocorrelation Function - Navwarning daily frequencies***

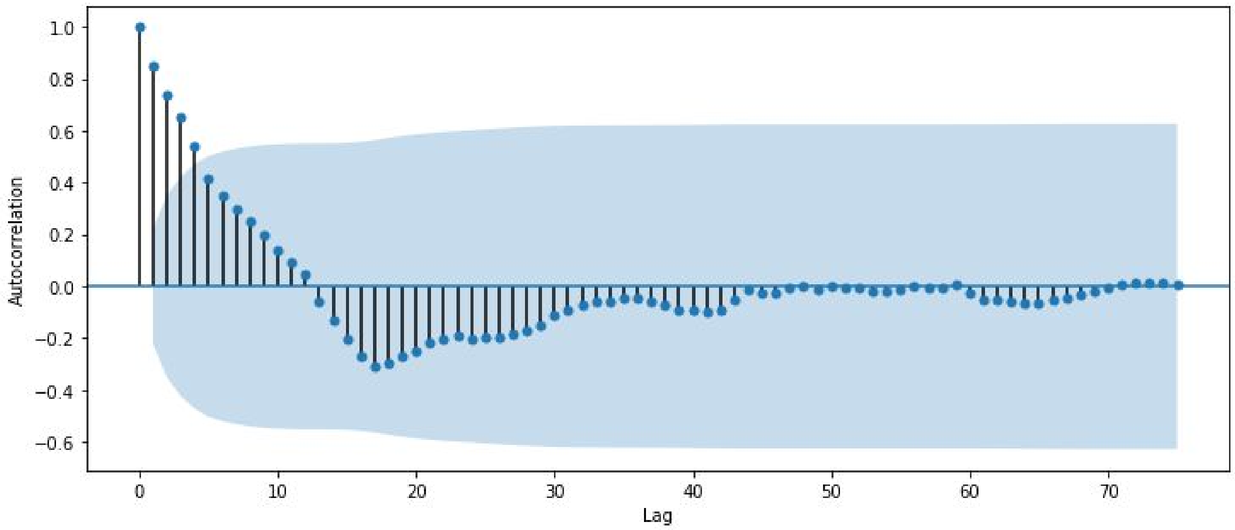


Figure-5 depicts the correlation between daily frequencies and frequencies at various lags, spanning the duration encompassed in the dataset. The ACF plot in Figure-5 shows that there are autocorrelation present at short duration, and that autocorrelation decays as lag increases.

Based on the methodology in (International Hydrographic Organization, 2014), it is stated that every message should be re-broadcasted with a specific frequency, which is likely what we see in the Figure-5. Without more temporal data at hand, it is decided to disregard to temporal aspect for now. However, this could be something for further investigation in the future.

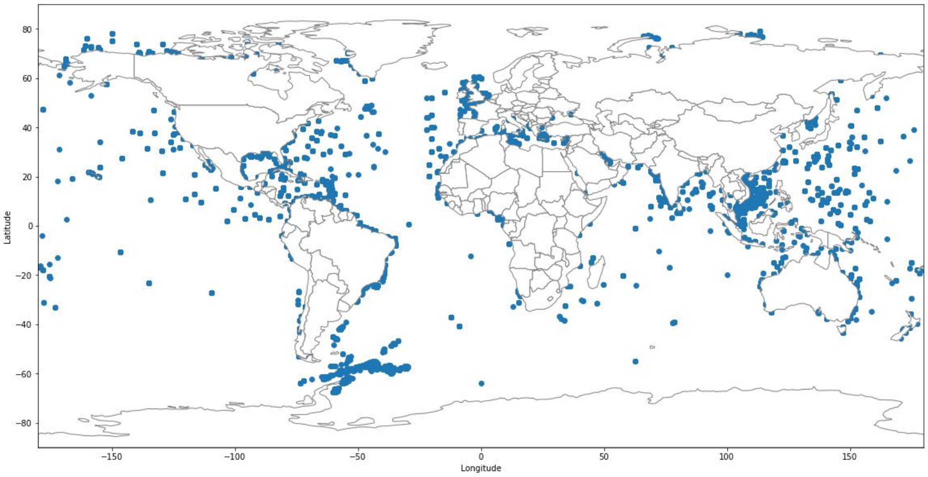
Spatial

Points, linestrings and polygons were included in the original dataset collected. Out of the 53,956 observations, 74.71% were points, 22.30% were linestrings and 2.99% were polygons.

Points

40,309 navwarnings were issued with point geometries. The longitudinal and latitudinal means were -13.44 and 22.05 respectively, though these values do not account for the fact that the points exist on a sphere, which is a topologically continuous study area. The set of longitude and latitude values had standard deviations of 89.68 and 30.72 respectively, indicating substantial spread.

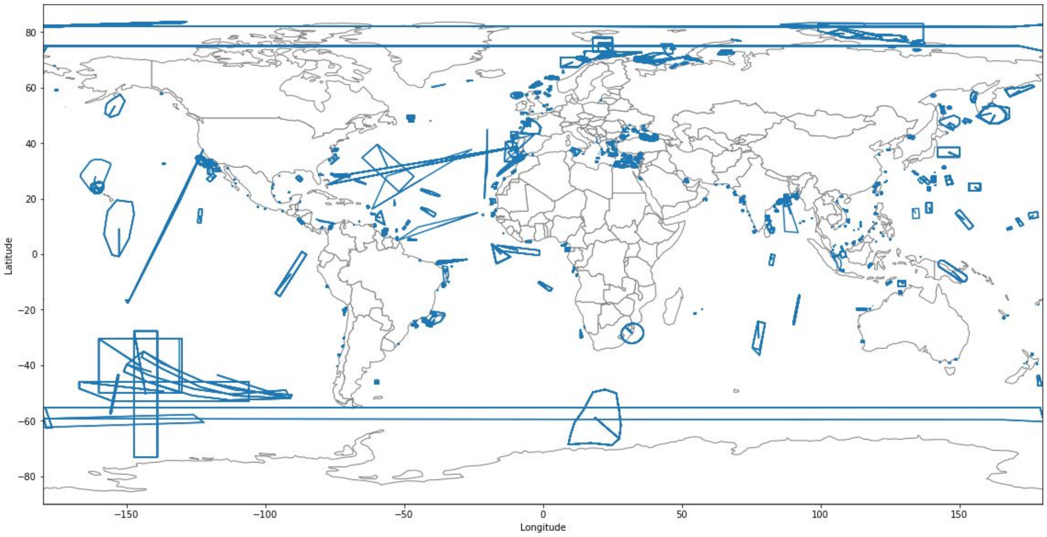
Figure-6: ***Point navwarning observations***



Linestrings

Associated with 12,034 navwarnings, linestrings often were connected to warnings related to survey operations, hazardous operations and military exercises, indicating that vessels should avoid those areas.

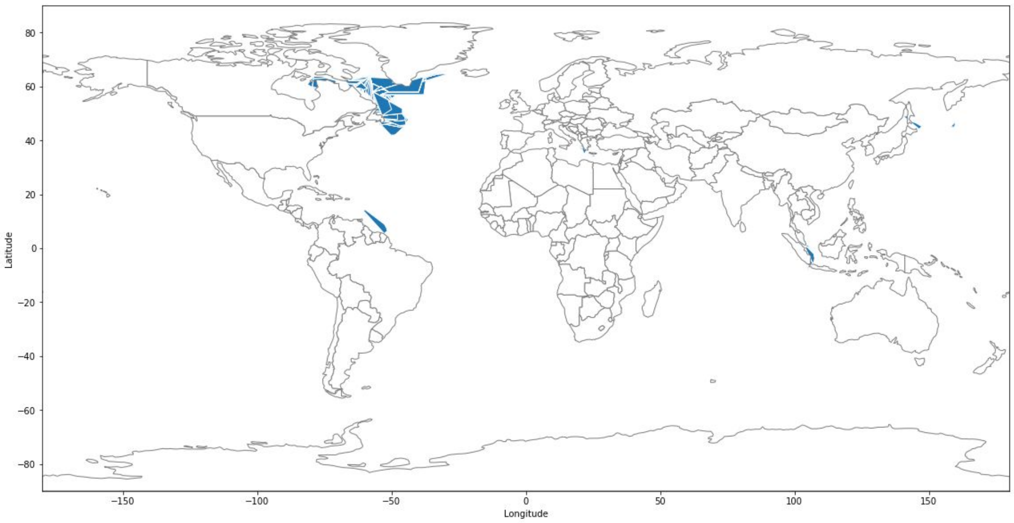
Figure-7: ***Linestring navwarning observations***



Polygons

Polygonal geometries were associated with 1,613 navigational warnings in the dataset analyzed. Polygons were often associated with warnings related to underwater operations, iceberg bulletins and cable operations.

Figure-8: ***Polygon navwarning observations***



Centroids

For the purposes of our analyses, we reduced polygon and linestring geometries into [longitude, latitude] coordinate pairs by calculating their centroids. While this did result in a loss of some information, it enabled cross-comparison of all navigational warnings.

Figure-9: ***Centroids of navwarning observations***

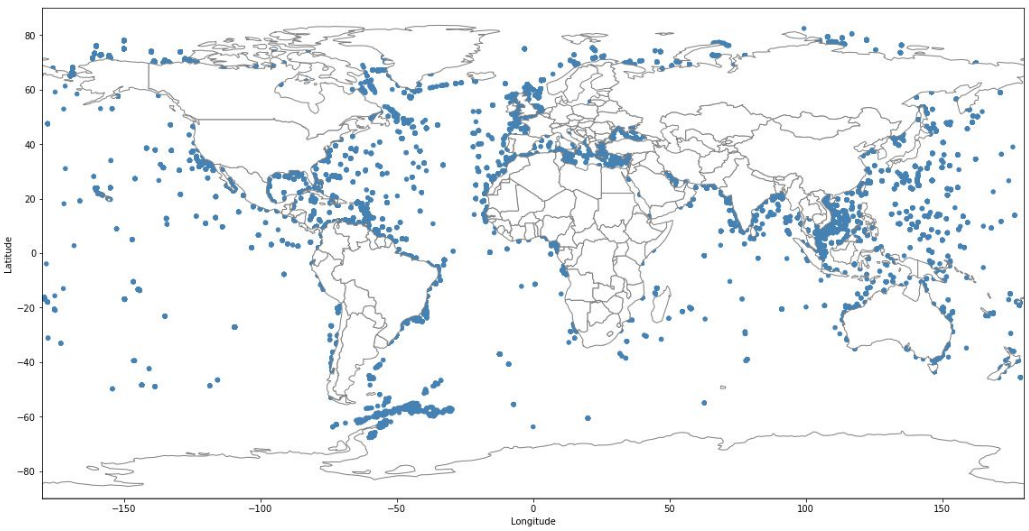


Figure-10 shows that kernel density estimates of point frequencies were highest around latitudes 15 - 20° N. These latitudes encompass the eastern Caribbean and South China Sea, two of the areas of highest navwarning point densities. A local maximum near 70°N is explained by a large number of observations related to scientific and cable operations occurring along the Arctic coasts of Russia, Alaska and Canada.

Figure-11 depicts the kernel density estimate of centroid longitudes. The maximum estimate occurs near -60° (i.e. 60°W), where numerous observations exist near the Drake Passage between South America and Antarctica, as well as in the eastern Caribbean, western Atlantic

Ocean and eastern US seaboard. A local maximum near 0° corresponds to high densities of warnings in the western European maritime space, especially around Great Britain.

Figure-10: ***Kernel density estimation of centroid latitudes***

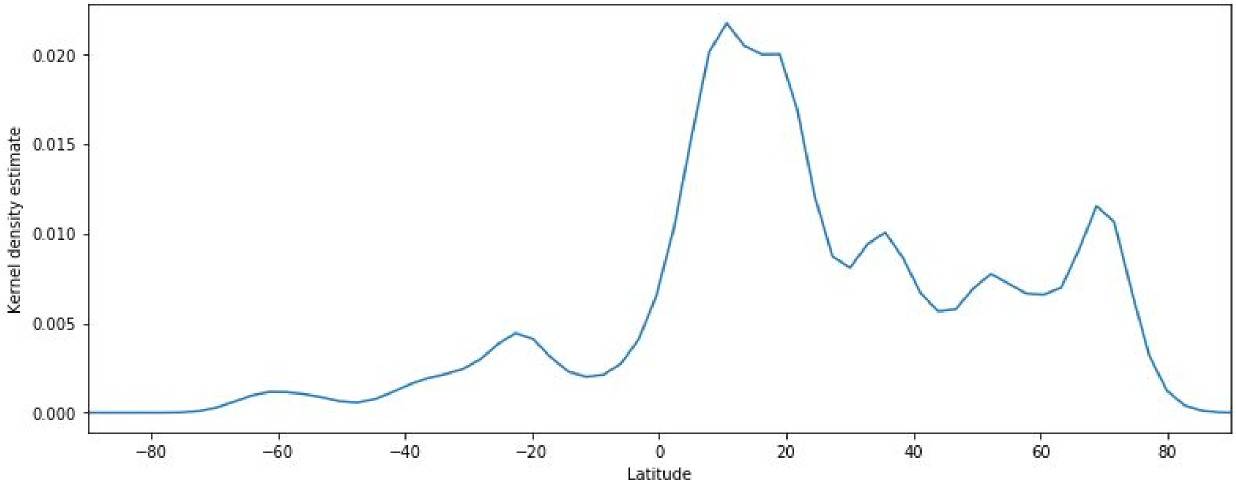
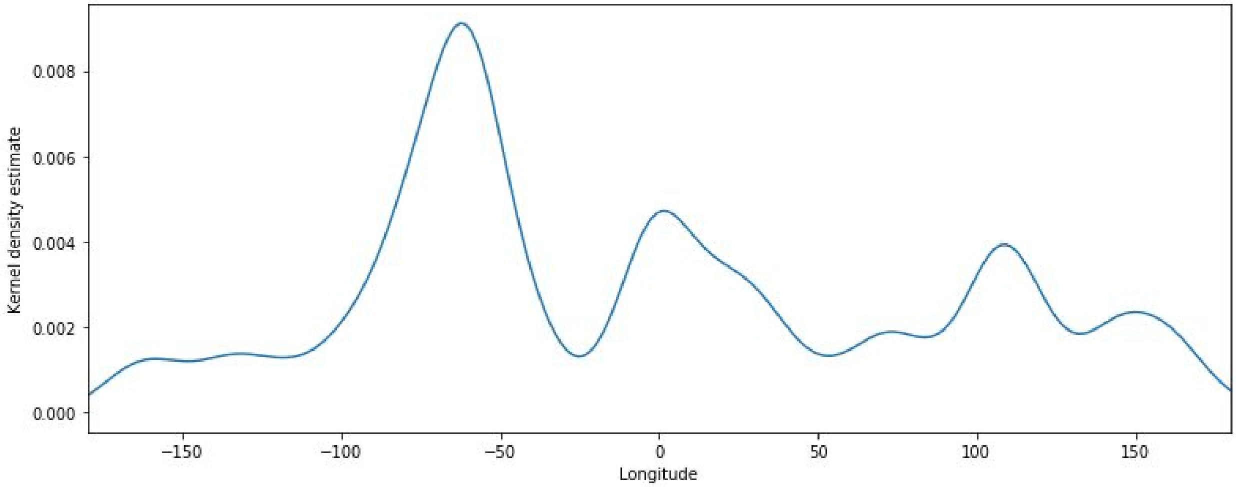


Figure-11: ***Kernel density estimation of centroid longitudes***



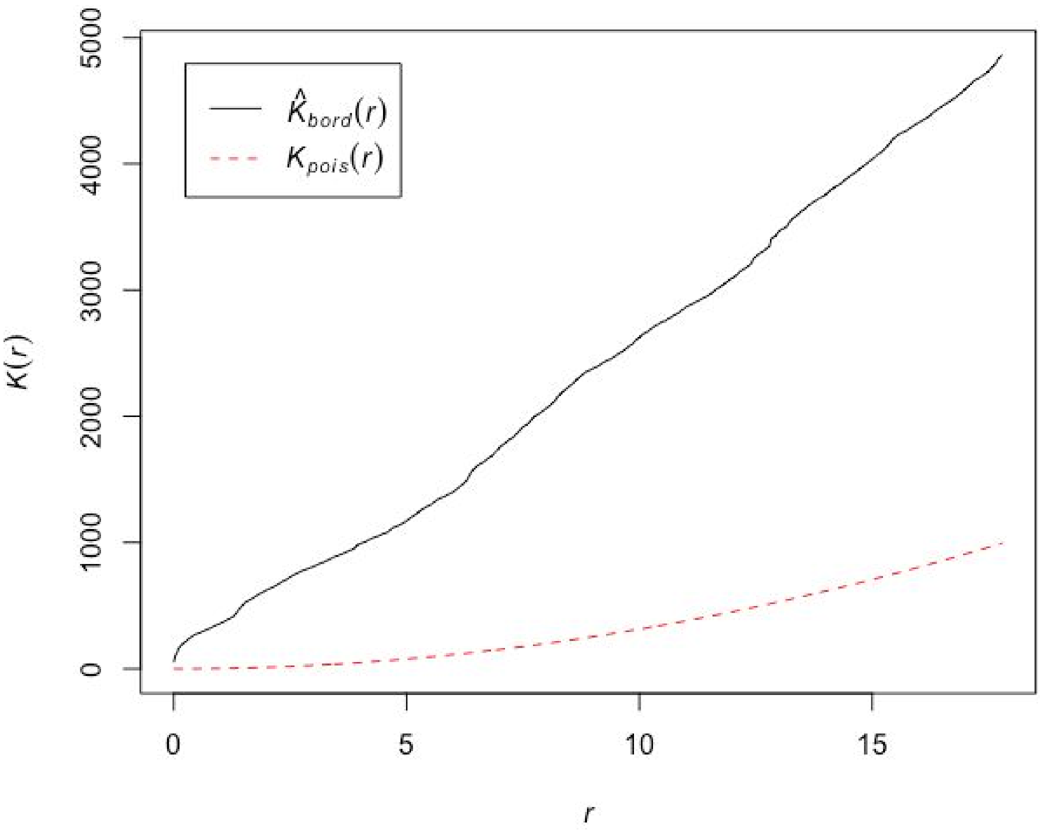
Spatial Autocorrelation

We calculated a Ripley's K function to empirically assess spatial clustering of navigational warning centroids. The primary advantage offered by the K-function employed is that it "provides a continuous evaluation of clustering over a range of distances, and hence is claimed to be free from MAUP [modifiable areal unit problem]" (Lu 2009). The function iterates over a number of distances, r, quantifying whether clustering, randomness or dispersion is observed at that scale $(\widehat{K}_{bord}(r))$. By comparing this with values expected in a pattern of complete spatial randomness ($\mathrm{K}_{\mathrm{pois}}(\mathrm{r})$) we discern the point pattern at that r.

Table-1: ***Interpretation of Ripley's K***

Relationship	**Meaning**
$\widehat{K}_{bord}(r) > K_{pois}(r)$	Spatial clustering
$\widehat{K}_{bord}(r) = K_{pois}(r)$	Complete spatial randomness
$\widehat{K}_{bord}(r) < K_{pois}(r)$	Spatial dispersion

Figure-12: ***K-function for navwarning point centroid pattern***



As evident in Figure-12, spatial clustering is observed at all distances in the point pattern considered. This is interpreted to mean that navigational warnings are more likely to be issued in certain areas of the world's oceans. Based on this evidence, we will perform further analysis, to identify in which areas of the maritime space are mariners more likely to encounter navwarnings.

Section 1: Topic Modelling and Class Labelling

Author: Kristian Lunow Nielsen

Topic modelling is a task within the field of Natural Language Processing (NLP), and topic modelling is frequently used when performing text classification. This part of the project is concerned with the messages contained in the warnings, and topic modelling of the content. Topic modelling of the content is suitable because the warnings by nature contains different topics. There are many ways to conduct topic modelling, and this project uses Latent Dirichlet Allocation (LDA), as it is the most popular choice (Cai et al,2016). The idea surrounding LDA is that documents, in our case the navigational warnings, are produced from a set of topics. It is the probability distribution of the topics that is believed to generate the words in the documents, and LDA tries to determine the topics from the collection of individual words obtained from the documents.

There are two interesting aspects, namely the spatial extent of the warnings and the fact that the warnings do not have class label as such. The spatial extent is interesting in terms of evaluating the stability of the topics detected on the entire dataset. Class labelling of the messages could be useful in many contexts, where different class-labelled warnings carry different weights, because of the different risk they represent or to evaluate cluster detected. An example of the former could be the risk-score-tool in the third individual project within this series, and an example of the latter could be the cluster analysis in the second project in this series.

The aim of the project is to understand the topic contained in the messages and based on these, assign class labels to warnings, in order to gain a broader understanding of the nature of warnings.

This individual project is structured in the following way; topics for the entire dataset is detected. The robustness of the topics is examined by examining the topics detected in the individual sub areas, mentioned in section 1. Lastly, the topics are used to assign class labels to each warning in the dataset.

The class labels are assigned through a repeated learning algorithm (RLACL), that uses initial beliefs about the most significant words within a class, to detect the correct class label for each warning. RLACL is proposed and engineered by the author, and a detailed outline of the algorithm is found in algorithm-1. The general idea in RLACL is that the probabilistic insights from LDA are utilised in a repeated fashion to classify the messages, i.e. the warnings. RLACL operates in an unsupervised fashion, as the warnings do not come with a class label a priori. By nature, that makes assessment of the performance difficult. However, K-Nearest Neighbour (K-NN) can with low additional effort be applied to cluster the tokens (words) in the documents, which should give an idea about the performance of RLACL. LDA and K-NN have certain similarities, and with the outcome of RLACL reflecting the results obtained through LDA, it seems like a valid way of accessing the performance of RLACL.

Algorithm-1: *Repeated Learning Algorithm for class labelling (RLACL)*

```
Initialise beliefs
For each topic: Define dictionary to keep track of word frequency.
Define slack or maximum number of words used for classification (maxNum).
cleanClassifications = 0

While(cleanClassifications < slack):
   For message in messages:
      Clean the message.
      Allocate a message to a topic, based on the beliefs.

      If a message is associated with more than one topic:
         Flag the message, to determine the correct topic later.
      Else:
         Assign the message to the associated topic
         Update word-frequency-dictionary for the topic.
         cleanClassifications += 1

   # Time to check the multiple-topic-associated-messages
   For message in flaggedMessages:
      Calculate the likelihood of the message belong to each of the topics.
      Allocate the message to the topic with the highest likelihood.
      cleanClassifications += 1

   For each topic: Update the beliefs with the most frequent word within the topic

   If maxNum is defined:
      If number of words in the individual beliefs equals maxNum:
         Break While-loop
```

Evaluating the likelihood of a message belonging to a certain class, taking place in line 15-18 in Algorithm-1, is based on one of the fundamental assumptions in multiclass-classification; namely that one observation can only belong to one class.
The reason why this seems appropriate, when topics normally aren't regarded as mutually exclusive for documents, is because it's warnings that are the documents. Warnings needs to be concise to avoid misunderstandings.

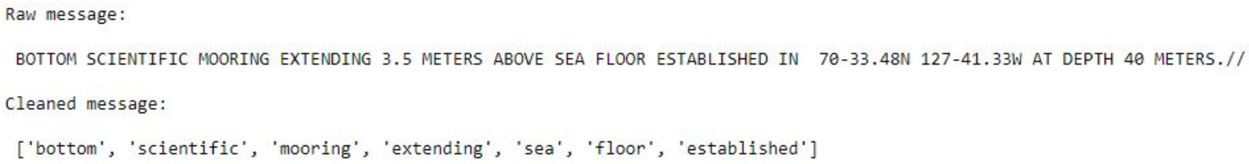
Topic Modelling

The messages are *cleaned* a prior to the topic modelling, to filter away as much noise, and thereby extract as much meaning, as possible. The cleaning implies the following steps:

- Convert to lowercase
- Remove stop words, punctuation, numeric characters, months, words not found in the English vocabulary and *words* with less than three characters
- Remove a bag of additional words.

The bag-of-additional words are elaborated, as the remaining cleaning is standard in the literature. The additional words removed are words that as such do not contribute any meaning, and here are words such as . See insights-1 for an example.
The cleaning implies that the vocabulary of the messages shrinks from +16000 words to just over 1000 words.

Insights-1 – ***An example of the pre-processing done on the messages.***



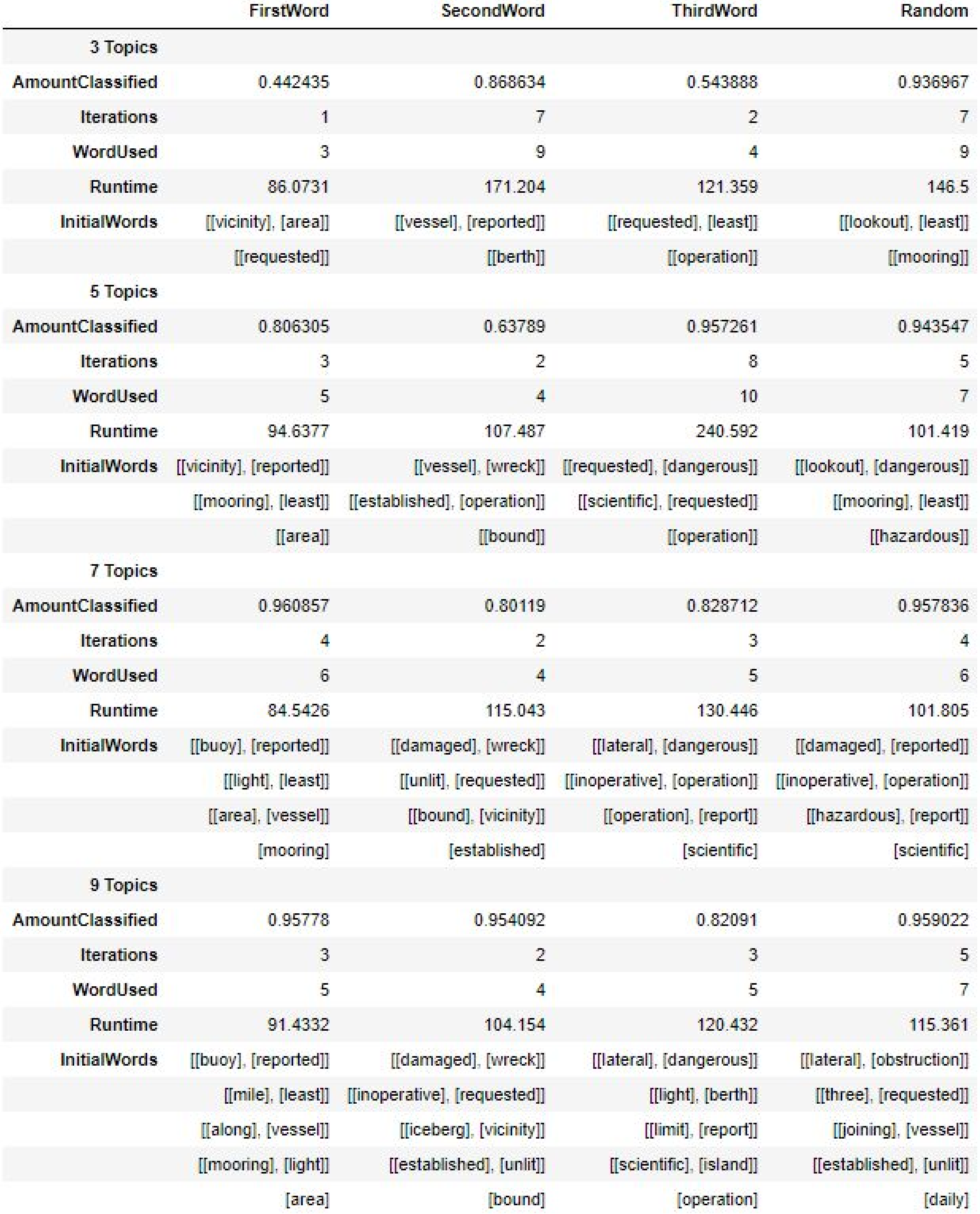
```
Raw message:
 BOTTOM SCIENTIFIC MOORING EXTENDING 3.5 METERS ABOVE SEA FLOOR ESTABLISHED IN  70-33.48N 127-41.33W AT DEPTH 40 METERS.//
Cleaned message:
 ['bottom', 'scientific', 'mooring', 'extending', 'sea', 'floor', 'established']
```

LDA needs the number of topics pre-specified, and the aim of the following discussion is to let the data try to reveal the number of topics. Due to the nature of how these warnings are structured, outlined in (International Hydrographic Organization, 2014), a range of categories emerge, see appendix. The categories can be used to indicate the appropriate number of topics, instead of arbitrarily choosing several topics. There is a trade-off when deciding on the number of topics, which is that more topics implies increasing chance of a word being labelled as being relevant in multiple topics, and a word is only allowed to belong to one group in RLACL to make the labelling unambiguous. Furthermore, more topics makes it increasingly difficult to interpret the topics clearly.

The performance of a set of possible numbers of topics are examined, and a final number of topics is chosen based on the performance. The performance metrics are the classified amount and number of words used classification. The range of possible numbers of topics are chosen based on the categories from (International Hydrographic Organization, 2014); grouping of similar categories in proposed subcategories, to ease interpretation of the topics produced, see appendix. This favours the number of topics being around 4-6. So, the range to be examined is . For each number of topics, a matrix, having the topics as rows and the top three words as columns, is generated. In addition, a fourth column is added, containing randomly chosen words for each topic. The classification is performed with each column as *initial beliefs*, and the results are shown in table-2. The purpose is to evaluate different sets of initial beliefs, generated by LDA. It is interesting to have in mind the implications of how the random combination is chosen; LDA aims as locating a few words that describe a certain topic and assign close-to-zero probability (A probability of the word being meaning full in describing that topic) to all other words. The implication is choosing random words within the words located by LDA has a high probability of resulting in a good separation, and thereby class labelling. Differently, choosing the random selection amongst all words in the vocabulary should result in a poor separation with high probability, as a result of many words having a close-to-zero probability within a certain topic. The former way of choosing the random selection is the chosen here.

It is expected that the difference in the classification results, within a set of topics, between the most significant words and the two less significant combinations is minor, as the most significant words are expected to be added quickly to the bag-of-words used to classify.

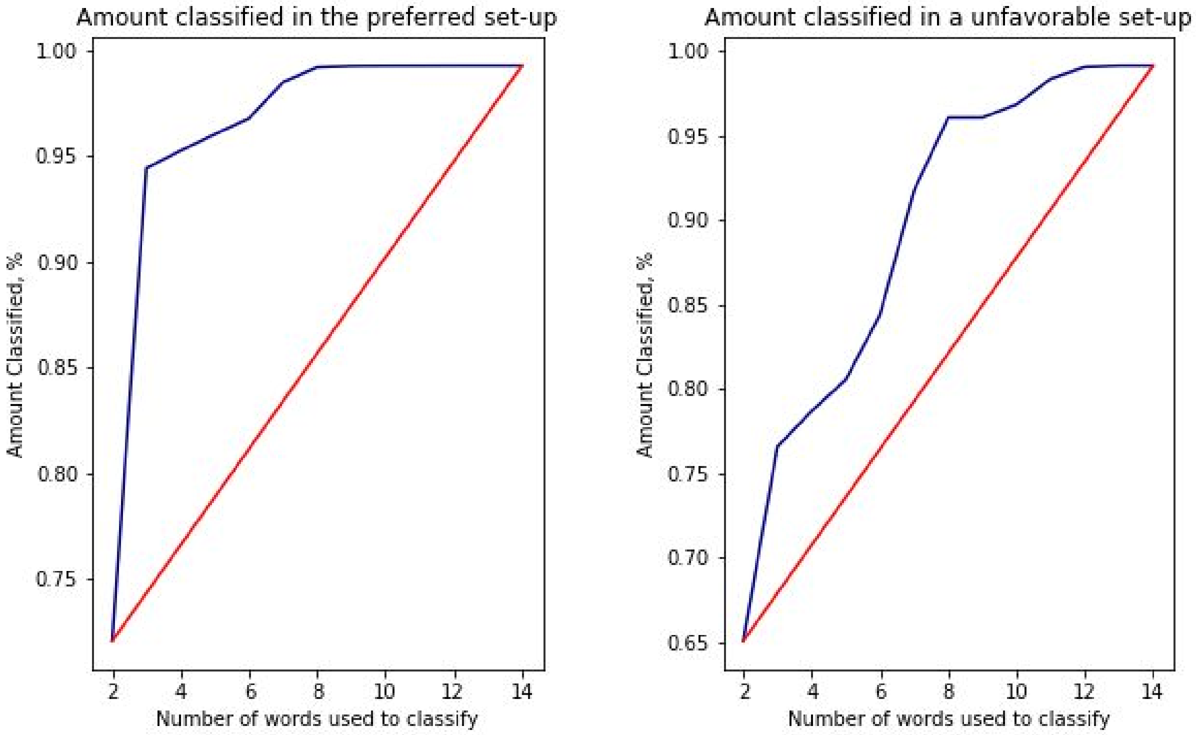
Table-2: ***Results from RLACL***



	FirstWord	SecondWord	ThirdWord	Random
3 Topics				
AmountClassified	0.442435	0.868634	0.543888	0.936967
Iterations	1	7	2	7
WordUsed	3	9	4	9
Runtime	86.0731	171.204	121.359	146.5
InitialWords	[[vicinity], [area]]	[[vessel], [reported]]	[[requested], [least]]	[[lookout], [least]]
	[[requested]]	[[berth]]	[[operation]]	[[mooring]]
5 Topics				
AmountClassified	0.806305	0.63789	0.957261	0.943547
Iterations	3	2	8	5
WordUsed	5	4	10	7
Runtime	94.6377	107.487	240.592	101.419
InitialWords	[[vicinity], [reported]]	[[vessel], [wreck]]	[[requested], [dangerous]]	[[lookout], [dangerous]]
	[[mooring], [least]]	[[established], [operation]]	[[scientific], [requested]]	[[mooring], [least]]
	[[area]]	[[bound]]	[[operation]]	[[hazardous]]
7 Topics				
AmountClassified	0.960857	0.80119	0.828712	0.957836
Iterations	4	2	3	4
WordUsed	6	4	5	6
Runtime	84.5426	115.043	130.446	101.805
InitialWords	[[buoy], [reported]]	[[damaged], [wreck]]	[[lateral], [dangerous]]	[[damaged], [reported]]
	[[light], [least]]	[[unlit], [requested]]	[[inoperative], [operation]]	[[inoperative], [operation]]
	[[area], [vessel]]	[[bound], [vicinity]]	[[operation], [report]]	[[hazardous], [report]]
	[mooring]	[established]	[scientific]	[scientific]
9 Topics				
AmountClassified	0.95778	0.954092	0.82091	0.959022
Iterations	3	2	3	5
WordUsed	5	4	5	7
Runtime	91.4332	104.154	120.432	115.361
InitialWords	[[buoy], [reported]]	[[damaged], [wreck]]	[[lateral], [dangerous]]	[[lateral], [obstruction]]
	[[mile], [least]]	[[inoperative], [requested]]	[[light], [berth]]	[[three], [requested]]
	[[along], [vessel]]	[[iceberg], [vicinity]]	[[limit], [report]]	[[joining], [vessel]]
	[[mooring], [light]]	[[established], [unlit]]	[[scientific], [island]]	[[established], [unlit]]
	[area]	[bound]	[operation]	[daily]

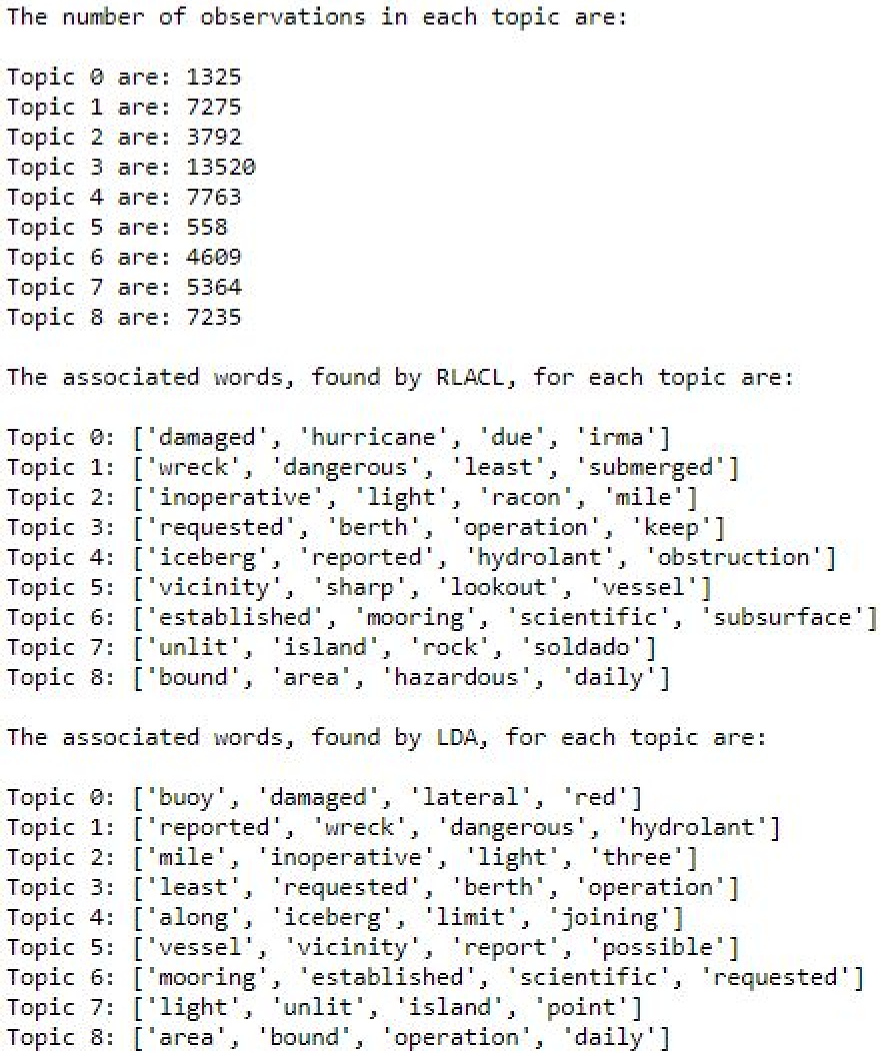
The results are seen in table-2, and before evaluating, it is worth thinking about how the results should be evaluated. A high amount of observations classified is without doubt of great importance, but fewer words used is preferred. The amount classified is an increasing function in the number of words used, and fewer words used, all else being equal, should imply greater generalisation properties. The algorithm is terminated if the change in the classified amount is less than one percentage point, between two restrictions on the number of words used, which favours set-ups with greater generalisation properties. See figure-13 for illustration. The function in figure-13 can take on many shapes, which is visible by the set of obtainable performances in table-2. The variety in the set of obtainable performances shows that different set-ups reach very different local optimums, and the preferred set-up is the set-up that maximises the area between the blue and the red curve in figure-13.

Figure-13: ***Amounts classified in two set-ups***



Notice from table-2 that all set of possible numbers of topics *can* produce decent results, i.e. around 95% classification rate, within this set-up. However, for the low number-of-topic sets more words are needed (9 for three-topics and 7 for five-topics). The third column for the five-topic case is not considered, because the same word appears in two different topics in the initial beliefs. The best performing set-up is nine topics and the second most significant words, which is able to classify 95.4% of the observations, with the use of ~3.5% of the words in the vocabulary (36 words in total and there are 1032 words in the vocabulary).

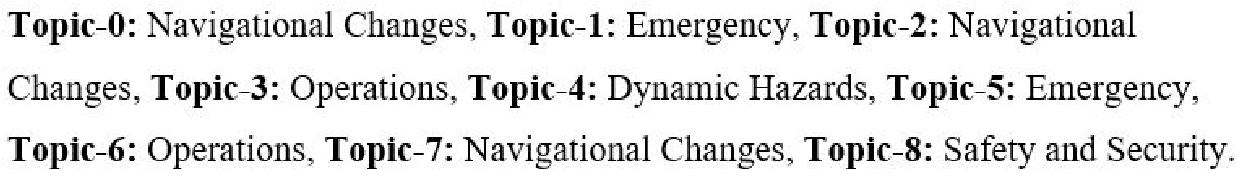
Insights-2 – ***Extended results for the preferred model***



```
The number of observations in each topic are:

Topic 0 are: 1325
Topic 1 are: 7275
Topic 2 are: 3792
Topic 3 are: 13520
Topic 4 are: 7763
Topic 5 are: 558
Topic 6 are: 4609
Topic 7 are: 5364
Topic 8 are: 7235

The associated words, found by RLACL, for each topic are:

Topic 0: ['damaged', 'hurricane', 'due', 'irma']
Topic 1: ['wreck', 'dangerous', 'least', 'submerged']
Topic 2: ['inoperative', 'light', 'racon', 'mile']
Topic 3: ['requested', 'berth', 'operation', 'keep']
Topic 4: ['iceberg', 'reported', 'hydrolant', 'obstruction']
Topic 5: ['vicinity', 'sharp', 'lookout', 'vessel']
Topic 6: ['established', 'mooring', 'scientific', 'subsurface']
Topic 7: ['unlit', 'island', 'rock', 'soldado']
Topic 8: ['bound', 'area', 'hazardous', 'daily']

The associated words, found by LDA, for each topic are:

Topic 0: ['buoy', 'damaged', 'lateral', 'red']
Topic 1: ['reported', 'wreck', 'dangerous', 'hydrolant']
Topic 2: ['mile', 'inoperative', 'light', 'three']
Topic 3: ['least', 'requested', 'berth', 'operation']
Topic 4: ['along', 'iceberg', 'limit', 'joining']
Topic 5: ['vessel', 'vicinity', 'report', 'possible']
Topic 6: ['mooring', 'established', 'scientific', 'requested']
Topic 7: ['light', 'unlit', 'island', 'point']
Topic 8: ['area', 'bound', 'operation', 'daily']
```

LDA is often criticised for producing topics that can be difficult to interpret but here the themes appear rather clear. Some topics appears to have the same theme, and each topic could be interpreted as follows:

Table-3 – *Interpretation of the topics*

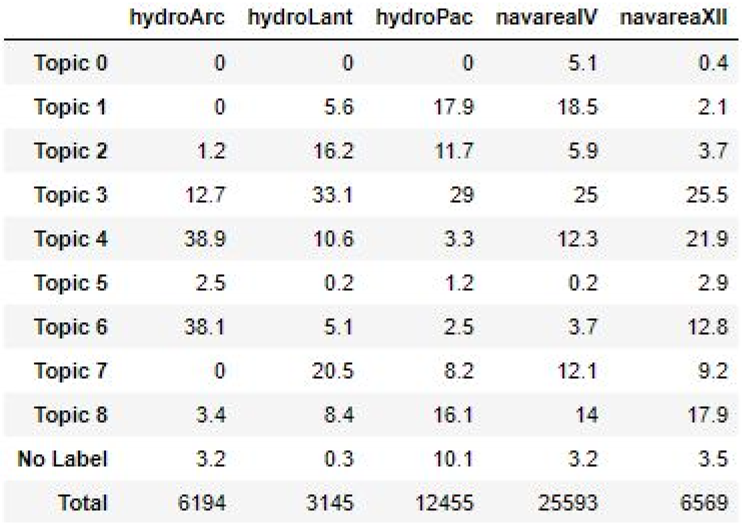


Topic-0: Navigational Changes, **Topic-1:** Emergency, **Topic-2:** Navigational Changes, **Topic-3:** Operations, **Topic-4:** Dynamic Hazards, **Topic-5:** Emergency, **Topic-6:** Operations, **Topic-7:** Navigational Changes, **Topic-8:** Safety and Security.

Furthermore, there are similarities in the words found by RLACL and LDA in most of the topics, and typically it's 2-3 words out of four, which is as expected.

With the number of topics located, it is time to consider the stability over the five subareas, which is done by examining the proportion of each topic in each area, see table-3, and the topics suggested for each subarea (Appendix,table-12). Table-4 quantifies the insights from table-12, which is that different areas are dominated by different topics. With *hydroArc* dominated by *dynamic hazards* (topic-4, likely icebergs) in the south and *operations* (topic-6) in the north, and *hydroLant* dominated by *operations* (topic-3) and *navigational changes* (topic-7), as examples.

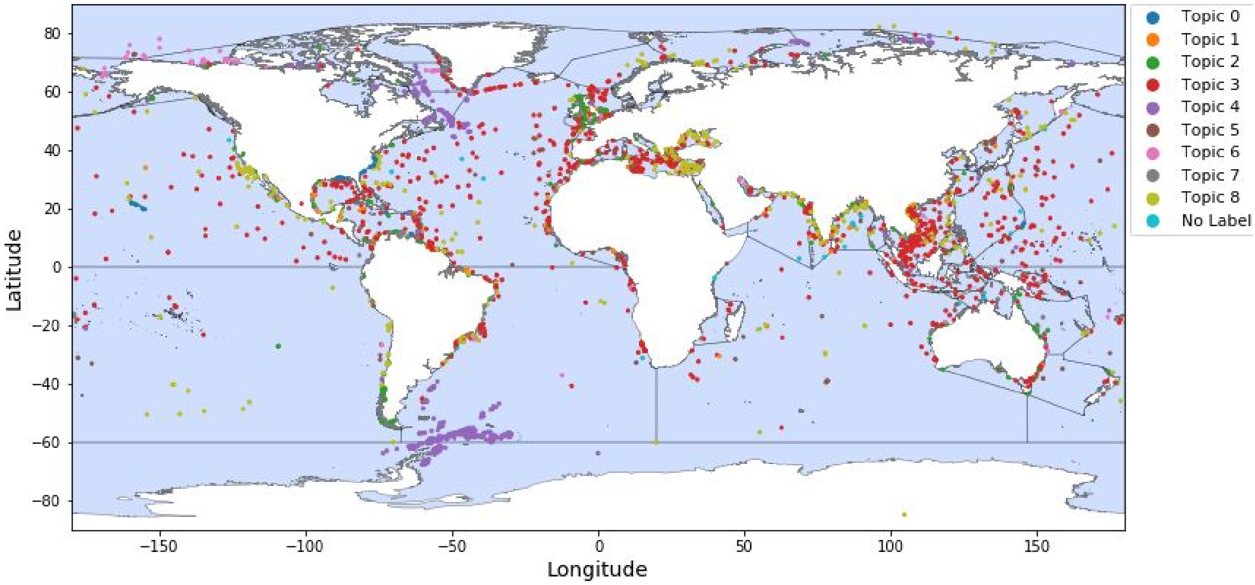
Table-4 – *Frequency of each topic in each area*



	hydroArc	hydroLant	hydroPac	navarealV	navareaXII
Topic 0	0	0	0	5.1	0.4
Topic 1	0	5.6	17.9	18.5	2.1
Topic 2	1.2	16.2	11.7	5.9	3.7
Topic 3	12.7	33.1	29	25	25.5
Topic 4	38.9	10.6	3.3	12.3	21.9
Topic 5	2.5	0.2	1.2	0.2	2.9
Topic 6	38.1	5.1	2.5	3.7	12.8
Topic 7	0	20.5	8.2	12.1	9.2
Topic 8	3.4	8.4	16.1	14	17.9
No Label	3.2	0.3	10.1	3.2	3.5
Total	6194	3145	12455	25593	6569

Figure-14 shows all observations along with the topic they have been labelled. Figure-14 highlights an insight, which is not directly visible from table-4, namely that topic-3 is the dominating topic across all observations.

Figure-14: ***Class labelled observations***

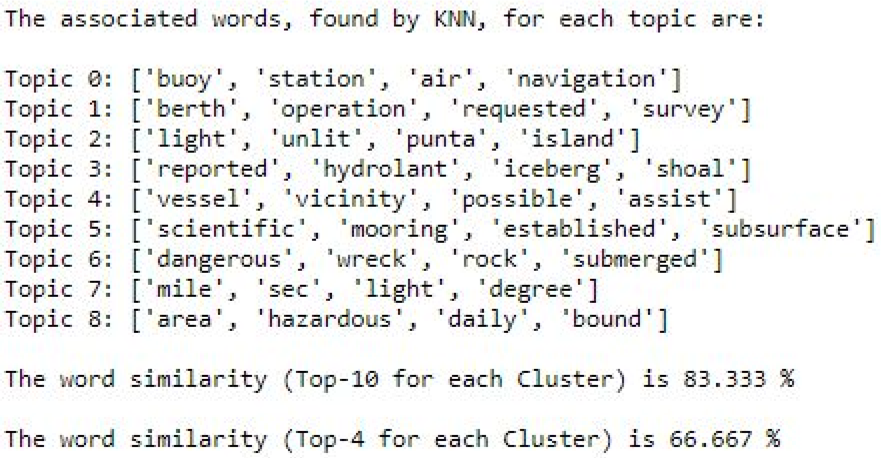


K-Nearest Neighbour

As mentioned in the introduction, the warnings do not per se come with a class label, which makes it difficult to access the performance of the labelling. However, K-NN is frequently used as a baseline for unsupervised text classification (He et al., 2004) (Dasarathy, 1991), because of its simplicity. One way to assess how the labelling went, is to compare the words associated with each cluster generated by K-NN and the words found by RLACL, see insights-2.

Pre-processing is needed for K-NN to extract insights from the documents, and for this is Term-Frequency-Inverse-Document-Frequency (TF-IDF) used, which is a commonly used pre-processing method in text mining. TF-IDF basically measures how relevant a term is in a document and is an equivalent pre-processing method to frequency-counting, as used in RLACL. The words associated with the clusters proposed by K-NN, along with the word similarity between the words proposed by RLACL and K-NN, is seen from Insights-3.

Insights-3: ***Comparison between results from K-NN and RLACL***



```
The associated words, found by KNN, for each topic are:

Topic 0: ['buoy', 'station', 'air', 'navigation']
Topic 1: ['berth', 'operation', 'requested', 'survey']
Topic 2: ['light', 'unlit', 'punta', 'island']
Topic 3: ['reported', 'hydrolant', 'iceberg', 'shoal']
Topic 4: ['vessel', 'vicinity', 'possible', 'assist']
Topic 5: ['scientific', 'mooring', 'established', 'subsurface']
Topic 6: ['dangerous', 'wreck', 'rock', 'submerged']
Topic 7: ['mile', 'sec', 'light', 'degree']
Topic 8: ['area', 'hazardous', 'daily', 'bound']

The word similarity (Top-10 for each Cluster) is 83.333 %

The word similarity (Top-4 for each Cluster) is 66.667 %
```

Insights-3 reveals that there a substantial similarity between the results produced by K-NN and RLACL, especially when the words located by RLACL is compared with the top-10 words in each cluster found by K-NN.

Section 2: DBSCAN Analysis of Maritime Navigational Warnings

Author: Katherine Jamieson

Introduction and Methodology

An appropriate methodology for analysing the geographical coordinate data, such as navigational warnings, is clustering. Clustering is a process that groups points together based on similarities. This is relevant to nautical route forecasting which benefits from understanding where areas of warnings are likely to be found. Due to the vast spatial distance of the data, a density based clustering approach is suitable for this analysis. The Density Based Spatial Clustering Applications with Noise (DBSCAN) algorithm was first introduced in 1996 (Ester et. al).

The DBSCAN was chosen as the best clustering for the dataset due to three key attributes. These are; the number of clusters does not need to be known in advance, the clusters can form in arbitrary shapes, and all points do not need to be allocated to clusters as the DBSCAN by design allows for noise (Ester et. al, 1996). As the dataset is large, it would be difficult to assume the number of clusters that will be formed. The DBSCAN clustering method is also essential to determining where global patterns lie in an arbitrary shape which is an important feature for a dataset of this size and scope (Ester et. al, 1996).

There are two key parameters that need to be set in order to run the DBSCAN algorithm; epsilon and min_samples. The epsilon defines the radius that is explored around each point while the min_samples is the minimum amount of points that need to be contained within the cluster (Ester et. al, 1996). The min_samples can be set to 1 if it is desirable to have no noise as all points will then be included in a cluster (though a cluster may be constituted of a single point).

The epsilon must be chosen carefully for this dataset as the global scale of the data results in large distances between points. Therefore, a large epsilon value should be chosen in order to achieve meaningful clusters in the data. The third parameter is the metric used for calculating distance between points. This is commonly either chosen as Euclidean or Haversine. For this analysis the haversine distance metric will be used as it requires data to be in the longitude and latitude format but the metric must be in radians (scikit-learn, 2018a).

The DBSCAN results can be validated with a silhouette score. A silhouette score indicates how far the sample is away from the neighbouring clusters which a value of 1 being the optimal score and a negative 1 score being the least optimal, indicating that some data points may have been mis-classified into the wrong clusters (scikit-learn, 2018b).

As there are the 5 region areas which are used to divide the nautical warnings, it makes sense to only look at clusters within these subsets of the data rather than the entire dataset which spans the globe. These areas are; hydroArc (Area 0), hydroLant (Area 1), hydroPac (Area 2), navareaIV (Area3) and navareaXXI (Area 4) as introduced at the beginning of the report. The total area of each region is different so the same epsilon value could not be used across each algorithm.

Building off the previous analysis on topic modelling of the navigational warnings done in the previous section of the report, we can look to see if there are any relationships that exist between the clustering that were formed and the classification (topics). The class labels were merged with the dataset in a new column so that once the clusters were created and each point was given a cluster label, they could be compared to the ‘ground truth’ of the class labels. The relationship between the two can evaluated by using Homogeneity, Completeness and Adjusted Rand scores. Rosenberg and Hirschberg (2007) define the following two desirable objectives for any cluster assignment: homogeneity which identifies if each cluster contains only members of a single class and completeness which measures if all members of a given class are assigned to the same cluster.

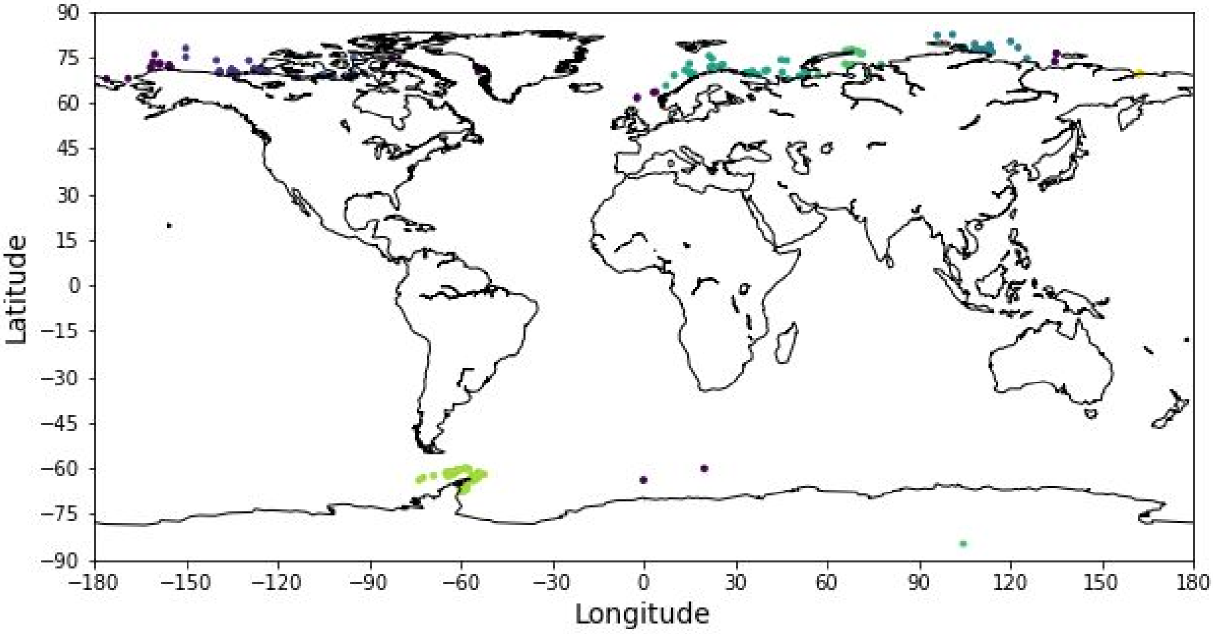
The Rand Index Adjusted score calculates the similarity between two clusters and it adjusted for chance. A score of 0.0 indicates that there is random labelling independent of the number of clusters and samples while a score of 1.0 indicates that clusters are identical (scikit-learn, 2018c)

The analysis was undertaken using the scikit-learn library in Python. The sklearn.cluster DBSCAN library and the sklearn.metrics were run to measure the Silhouette Coefficient, Homogeneity, Completeness and Adjusted Rand Index scores.

Results and Validation

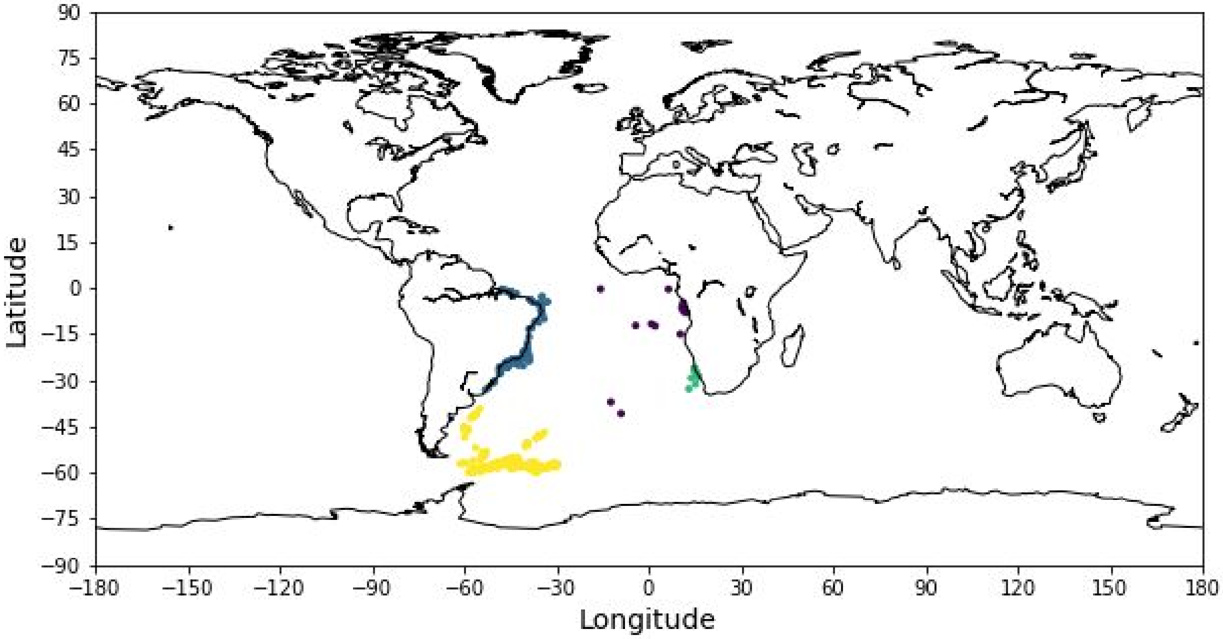
The DBSCAN algorithm was run across the 5 key navigational areas with good results.

Figure-16: ***Cluster in HydroArc (Area 0)***



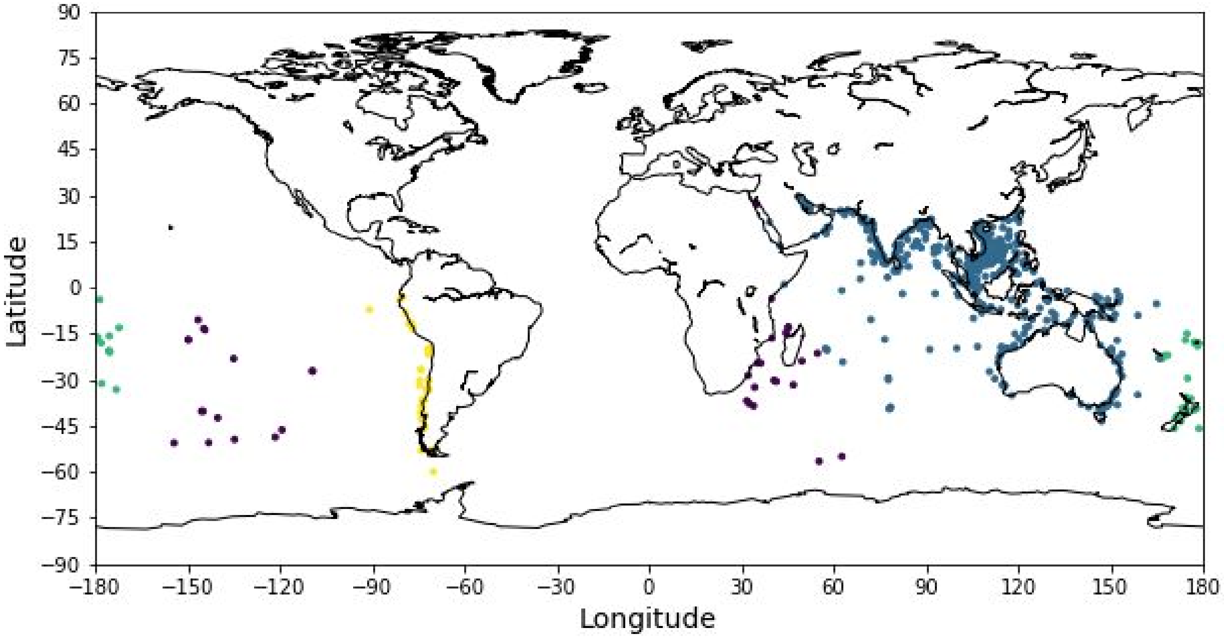
The hydroArc region encompasses the Arctic and Southern Oceans (North and South poles) which results in points which appear far on a linear plane but on the global sphere are closer. The epsilon chosen for this area was 0.2 with 15 min_samples resulting in a Silhouette Coefficient of 0.73.

Figure-17: ***Clusters in HydroLant (Area 1)***



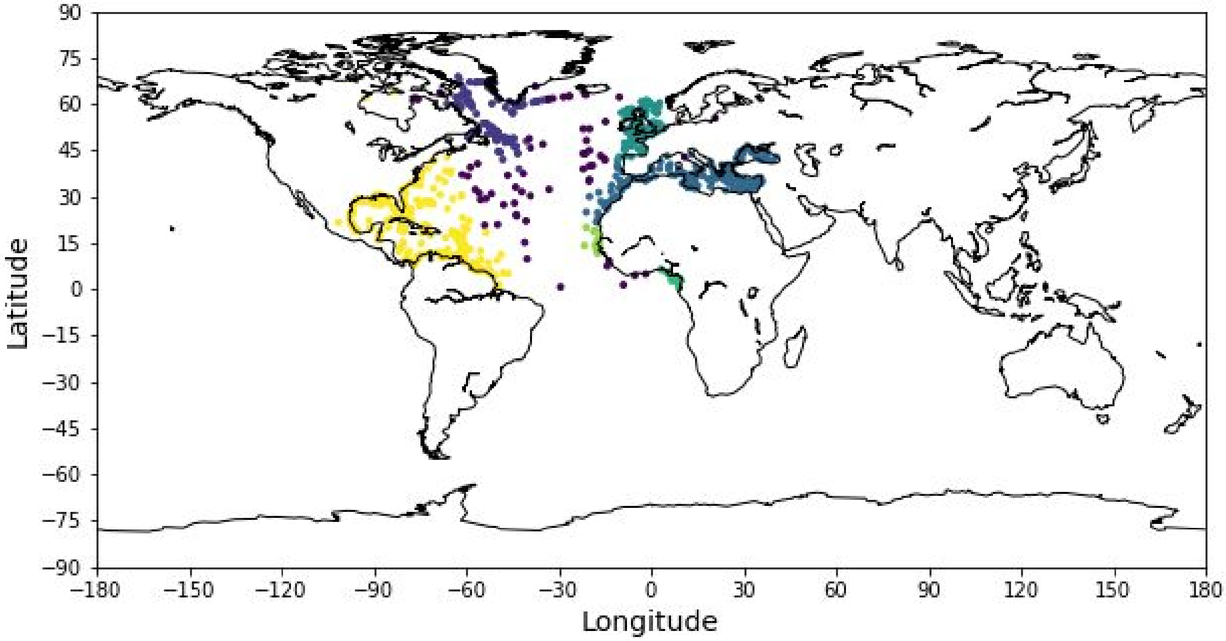
The hydroLant region encompasses the Southern portion of the Atlantic Ocean, between South America and Africa. The epsilon chosen for this area was 0.08 with 10 min_samples resulting in a Silhouette Coefficient of 0.69.

Figure-18: ***Clusters in HydroPac (Area 2)***



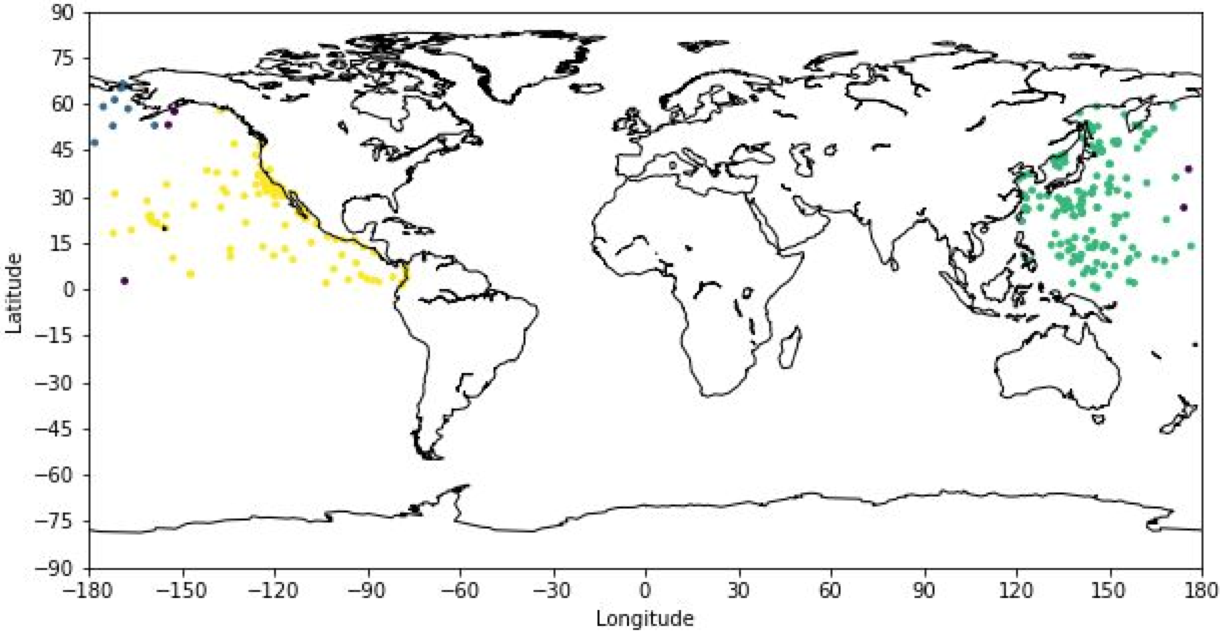
The hydroPac region (Figure-18) covers the Southern portion of the Pacific Ocean. As the Pacific Ocean spans around the globe, it is important to consider the clustering results for this area. The epsilon chosen for this area was 0.22 resulting in a Silhouette Coefficient of 0.62.

Figure-19: ***Clusters in NavareaIV (Area 3)***



The navareaIV region (Figure-19) covers the Northern Atlantic Ocean between North America and Europe. This is a very busy shipping and transport channel so we would expect to see more warnings in this area. The optimal value chosen for the epsilon for this area was 0.07 and 10 min_samples, resulting in a Silhouette Coefficient of 0.49.

Figure-20: ***Clusters in NavareaXXI (Area 4)***



The navareaXXI region (Figure-20) covers the geographical area of the Northern portion of the Pacific Ocean from Canada to Asia. The epsilon chosen for this area was 0.23 with 10 min_samples which resulted in a Silhouette Coefficient of 0.698.

Table-5: ***Silhouette Coefficient and number of Clusters by Area***

Area	**Silhouette Coefficient**	**Number of Clusters**
hydroArc	0.726	7
hydroLant	0.69	3
hydroPac	0.623	3
navareaIV	0.495	3
navareaXXI	0.698	3

The DBSCAN reveals that there are patterns within the navigational warnings as clusters exist within the dataset. The hydroArc region had the best Silhouette Coefficient score with 0.726. While this value seems like a good validation of the results, the geography of that area presents a challenge as the algorithm does not account for the round shape of the globe but interprets the data points as being spread on a flat surface. This means that where the clustering results are projected onto a coordinate system this is likely result in some distortion of the clusters as there is a discrepancy in distances. The navareaXXI and hydroLant regions also had good Silhouette Coefficient scores, indicating that the points across the clusters have been allocated correctly (Table-5). As the densities of warning varying greatly in some regions, this is a consideration when analysing the results of the DBSCAN. The epsilon chosen for each area may not be optimal to recognize clusters of varying densities within each region. Some points may be classed as noise that would otherwise be included in clusters if a similar epsilon value was chosen.

Table-6: ***Homogeneity and Completeness Score by Area (10 Classes)***

Area	**Homogeneity Score**	**Completeness Score**
hydroArc	0.301	0.284
hydroLant	0.535	0.769
hydroPac	0.056	0.140
navareaIV	0.240	0.279
navareaXXI	0.102	0.188

Table-7: ***Homogeneity and Completeness Score by Area (5 Classes)***

Area	**Homogeneity Score**	**Completeness Score**
hydroArc	0.100	0.086
hydroLant	0.008	0.013
hydroPac	0.004	0.011
navareaIV	0.028	0.027
navareaXXI	0.020	0.035

It was hypothesized that there could be a relationship between the topic modelling classes and the DBSCAN algorithm results, in that clusters would be likely to form around singular classes due to the geography and characteristics of certain areas likely resulting in specific navigational warnings.

As a score of 1 would indicate true homogeneous and complete clusters, the homogeneity and completeness scores of each area (Table-5) show that there is a weak relationship between the clusters and their topic classes (scikit-learn, 2018d). The clusters do not demonstrate singular classifications from the previous analysis, however they reflect that the topics within the clusters are not completely random.

There was an assumption that the greater number of classes available, the more diverse that the warnings within a cluster would be. In order to test this theory, the classes were reduced and the Homogeneity and Completeness scores were run again. Reducing the size of the classifications (from 10 to 5) as displayed in Table-7 has an overall negative impact on the homogeneity and completeness score which is unexpected.

Table-8: ***Adjusted Rand Index by Area***

Area	**Adjusted Rand Score**
hydroArc	0.406
hydroLant	0.783
hydroPac	0.090
navareaIV	0.139
navareaXXI	0.079

The Adjusted Rand Scores shown in Table-8 that classes are split across the clusters in each region and not allocated randomly.

Limitations of the Methodology

While the clustering is able to provide valuable insight into the clusters in the data, the limitations within the data must be discussed. Due to the nature of nautical warnings there is a limitation to the areas that the data points, and therefore the clustering, can occur. In this analysis, the data points are limited to bodies of water and cannot occur on land so this a point to consider in assessing the accuracy of the DBSCAN. This can also be seen as a potential benefit of the DBSCAN in this scenario. When choosing the epsilon distance, there is a limited amount of points that can exist as the data points are limited to the ocean area.

We must also consider that the points used for analysis are the centroids of the GPS coordinates instead of the exact longitude and latitude coordinates as some of the navigational warnings came as polygons with multiple coordinates for the area. This can cause an issue with outliers or in the case of the DBSCAN, ‘noise’, points from being excluded from the dataset (DataCamp, 2019). For this analysis, this is not a concern as the coordinates are not meant to represent exact locations but will generally cover a wider area in the water.

Overall, the DBSCAN was successful in producing clusters within the data that could be validated with statistical checks.

Section 3: Forecasting Risk

Author: John R Hoopes IV

With approximately four months of global georeferenced data, we sought to quantify the risk profile entailed in a ship's journey. We developed two risk quantification methods, one using a quadrat method and the other a two dimensional spatial kernel density estimation. Our conclusion that warnings tend to cluster in space based on the K-function point pattern analysis justified our decision to predict the risk of encountering warnings in future voyages on historical locations of observations.

Data Preparation

NAVWARNING Data

Navwarning data was issued daily by the US Government in KML format. Preparing this data for analysis entailed converting the KML to Geojson, then parsing the text to extract dimensions including Message, Chart, Area, Authority, and Start and End date. Furthermore, the classification methods described in Section 1: Topic Modelling and Class Labelling provided an additional categorical dimension to each observation. This additional information enabled us to calculate a disaggregated risk profile, providing more nuanced insight into the potential risks a vessel might encounter in their journey.

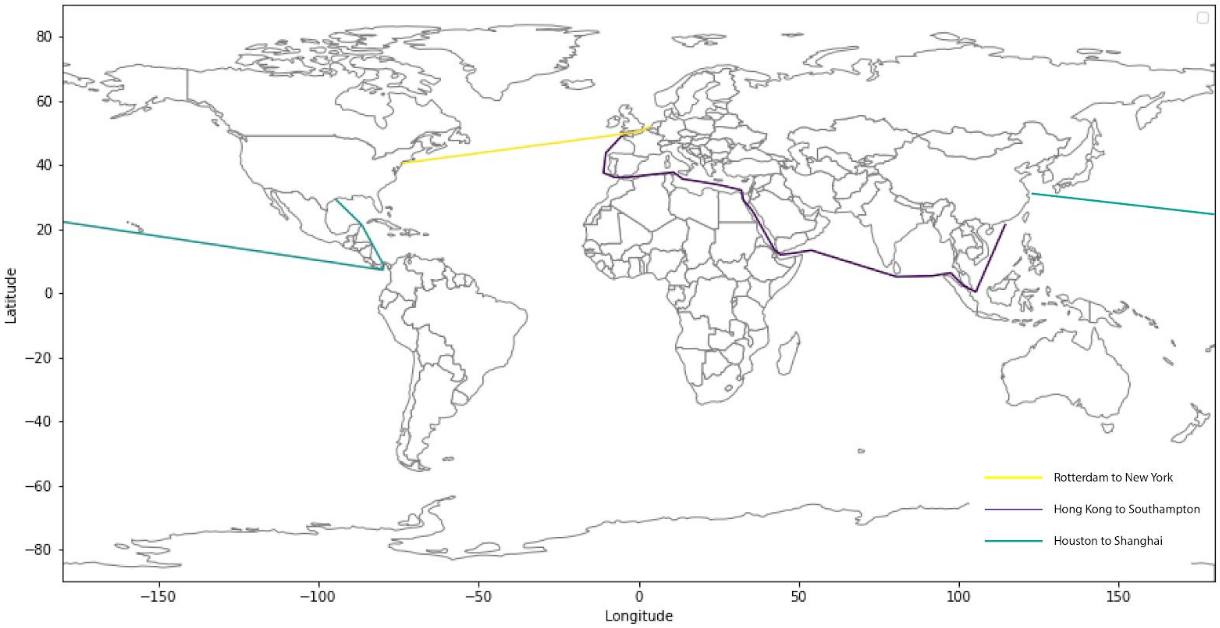
Example routes to forecast

To demonstrate our method we generated three example linestrings representing common shipping routes transiting heavily-trafficked maritime corridors (Stewart 2018). Route attributes are presented in Table-9.

Table-9: ***Example route attributes***

Origin	Destination	Distance (km)	Time (hrs at 25 kts)	Time (days at 25 kts)
Hong Kong	Southampton	19 344.94	417.82	17.41
Rotterdam	New York	8 790.18	189.85	7.91
Houston	Shanghai	20 799.95	449.24	18.72

Figure-21: ***Example route linestrings***



Quadrat density estimation

Methodology

To create a quadrat risk map of the Earth's maritime space based on the dataset of navigational warnings under consideration, we divided the Earth's surface into 648 regular square grids 5 degrees to a side. The sum of warnings contained within each grid cell was calculated and assigned to that geometry, then normalized. In the resulting matrix, values represented an estimation of the likelihood of encountering a navigational warning in that area of the ocean.

Considerations

Ecology

This quadrat method is a simple density estimation performed by binning observation frequencies based on somewhat arbitrary delineations in the study space. This is effectively a two dimensional histogram; selection of grid cell size affects the resultant risk grid and the calculated voyage risk forecast, at times substantially changing the estimation (Lerner 2013, VanderPlas 2016). This is an example of the modifiable areal unit problem (Openshaw and Taylor 1979), “which refers to the fact that statistical results are often contingent upon how data are aggregated using different areal units or scales” (Sui 2009). Figure-22 and Figure-23 illustrate this effect.

Our decision to use 5° quadrats was based on a visual inspection of various sample risk grids created for different cell sizes. Further investigation into the effects of cell size selection, as well as boundary line placement, is warranted, but beyond the scope of our analysis. We suspect that a dataset spanning a longer duration would enable the creation of a finer-grained quadrat map without overfitting the model.

Figure-22: ***Quadrat Density Estimation - 1° quadrats***

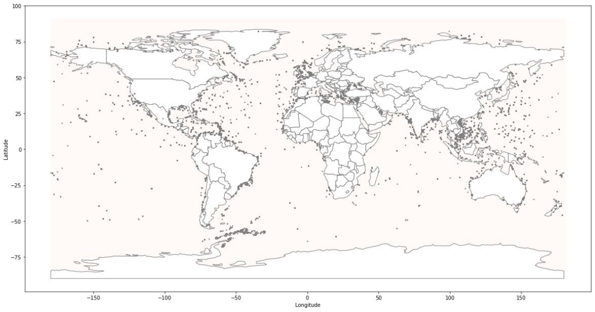
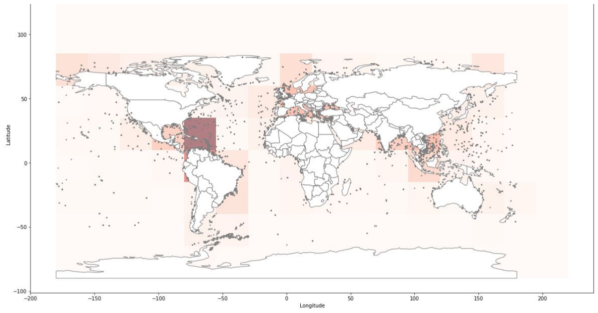


Figure-23: ***Quadrat Density Estimation - 25° quadrats***



Higher dimensional geometries

Navigational warnings were sometimes issued with line and polygon geometries, rather than simple point coordinates. For our effort we calculated geometry centroids and treated each warning as a point, thereby losing some information. A more sophisticated model could account for risks of encountering warnings with associated geometries, including lines and polygons.

Risk Profiling

To predict a voyage's risk profile based on the quadrat risk estimation method, only quadrats intersected by the ship track's linestring are considered. The relative length of the linestring intersecting each cell is used to scale that cell's risk score, which is then added to voyage's cumulative risk score. We chose this approach to address the boundary placement problem described above: if a voyage only passes through the corner a high risk cell for a short distance, exposure to the risk represented in that cell is proportionately limited in our quantification.

Our initial design assumes that ships travel at a constant speed, so relative length - distance - is proportionate to the amount of time spent in a geometry. Future forecasting models will accommodate variability in ship steaming speeds.

Results

Risk Profiles

Based on the 5° quadrat risk grid generated from the navigational warnings dataset, this voyage's forecasted risk scores for all warnings and each warning class are displayed in Table-10.

Table-10: ***Cumulative risk scores from quadrat approach by classification***

Route	Classification										
	All	0	1	2	3	4	5	6	7	8	9
Hong Kong to Southampton	6.9068	0.0000	1.5097	9.0664	3.6290	2.2630	0.4828	0.8752	1.4991	3.4413	5.7675
Rotterdam to New York	3.1967	0.0000	1.1493	2.7413	1.1345	2.2287	0.2495	0.0000	0.1881	1.2940	5.1514
Houston to Shanghai	2.4846	0.0205	1.7103	1.1997	0.7360	1.5498	0.1561	0.2187	0.4849	0.3082	1.9971

Values represent a quantification of the risk of encountering a warning of each classification along the route specified. Since risk values were normalized, cumulative risk scores are only comparable across voyages within columns, not across classes.

Spatial Kernel Density Estimation

Methodology

To address several of the constraints inherent in the simple quadrat density estimation method employed above, we developed a more sophisticated tool to quantitatively forecast the risk of a vessel encountering navigational hazards in their route using a spatial kernel density estimation. Using Python's scikit-learn library, we fit two-dimensional kernel density estimation models for the full dataset of navigational warnings, as well as for each class of warnings based on observation [longitude, latitude] centroid coordinate pairs.

A kernel density estimation is a probability density function that allows us to calculate the estimated likelihood of encountering an event in the study area (SciPy 2019). Our approach employed a Gaussian kernel, in which likelihood decays as distance from observed events increases based on the following formula (Wikipedia 2019):

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Since our analysis space represented a sphere, we employed a Haversine distance formula to calculate the distance between points on the surface of a sphere, accounting for the volume's curvature and topological continuity. Coordinate points were converted from decimal degrees to radians before fitting the model.

Considerations

Similar to quadrat cell size, bandwidth is an adjustable parameter, the selection of which will affect the resultant model and likelihood scores. Our model was calculated using a bandwidth of 0.04 radians, approximately 6 hours steaming at optimal speed, again selected based on visual inspection of various parameters. In future iterations of our risk model this hyperparameter will be empirically tuned using the k-fold cross-validation technique (VanderPlas 2016), which was too computationally intensive to complete on the hardware used in our analysis. The effects different bandwidths have on risk profile estimation is worth further investigation, as is exploring the use of KDE models fit with different bandwidths in different spatial extents.

Our kernel density estimation approach addresses many of the shortcomings related to cell size and boundary selection inherent to the quadrat approach employed above. Furthermore, it is possible to estimate the likelihood of encountering a warning at arbitrary points along a path through the study space, allowing us to forecast voyage risk with more precision than the quadrat method affords.

Separate KDE models were trained for the full dataset and for each class, to provide an overall risk score and a disaggregated view of what class of warnings a ship might encounter on its route. Risk profiles and cumulative risk scores are therefore only comparable across routes. Quantitative values across classes for the same voyage are not comparable since the calculated risk scores represent the likelihood of encountering a specific type of warning at a sample point.

Risk Profiling

To calculate a risk profile based on fitted kernel density estimation functions, likelihoods were calculated for an array of points spaced at regular intervals along the linestring under consideration. Again, this sampling method assumes that ships travel at constant rates during transit. Likelihoods were calculated along the linestring profiled at an interval of 0.01 radians, or ~63.78 km. Cumulative risk scores between different linestrings are only comparable if they employ identical sampling frequencies.

Results

Visualizing kernel density estimations

Figures 24 and 25 display heat maps depicting sampled scores from our set of kernel density estimations.

Figure-24: ***KDE - All navwarnings***

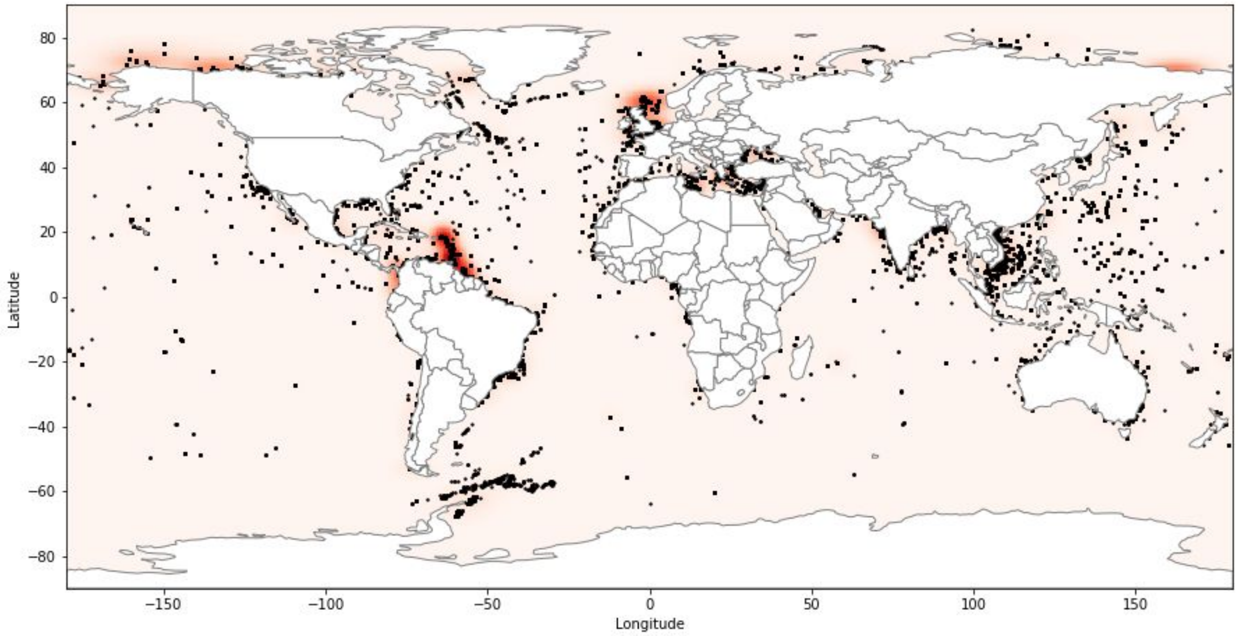
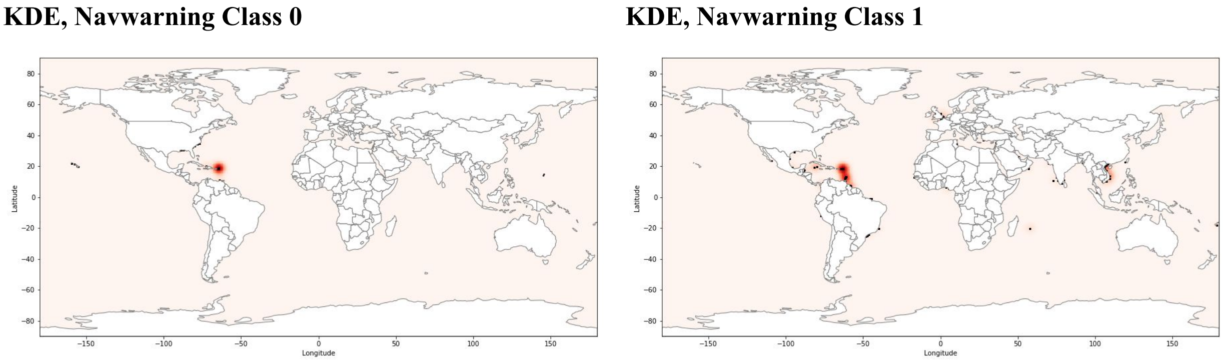
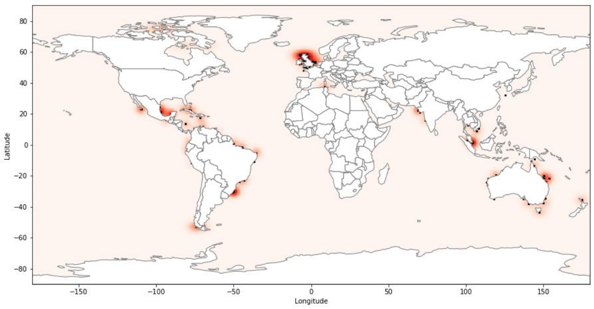


Figure-25: ***KDE results for all classes***

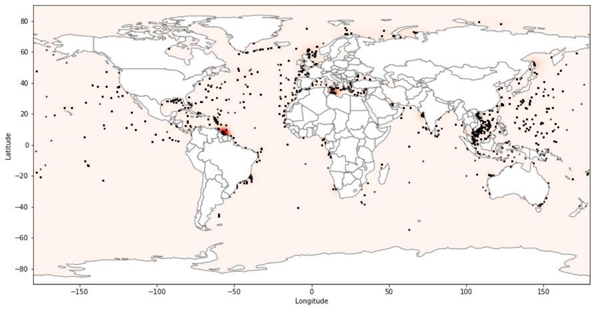

KDE, Navwarning Class 0

KDE, Navwarning Class 1

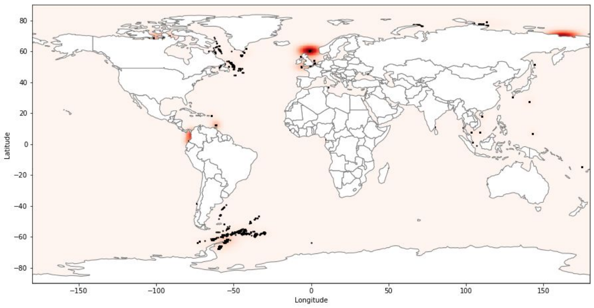
KDE, Navwarning Class 2



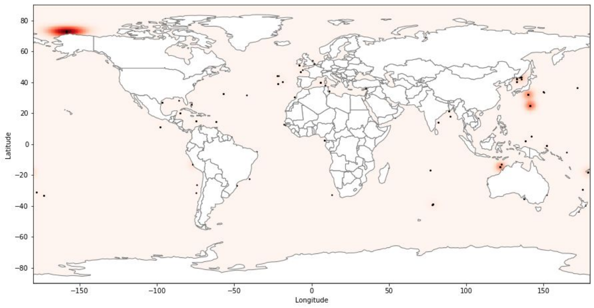
KDE, Navwarning Class 3



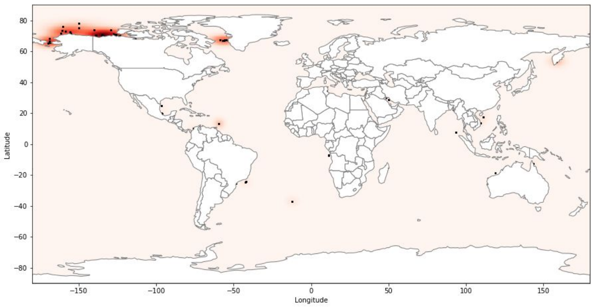
KDE, Navwarning Class 4



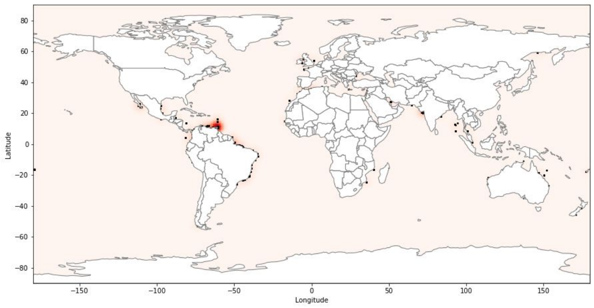
KDE, Navwarning Class 5



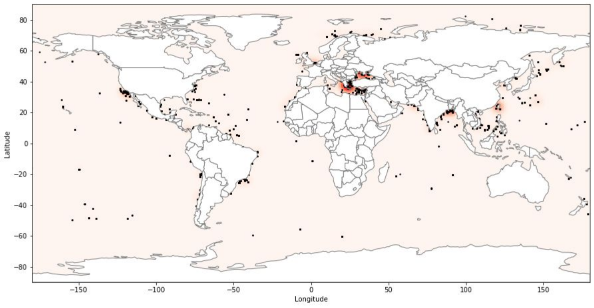
KDE, Navwarning Class 6



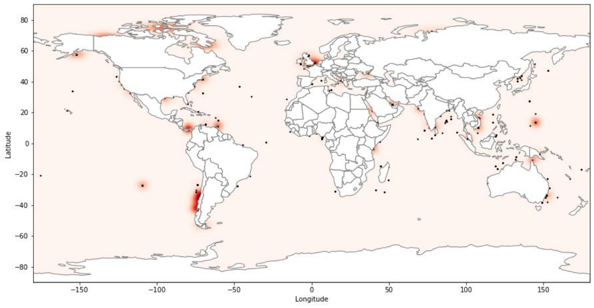
KDE, Navwarning Class 7



KDE, Navwarning Class 8



KDE, Navwarning Class 9



Risk Profiles

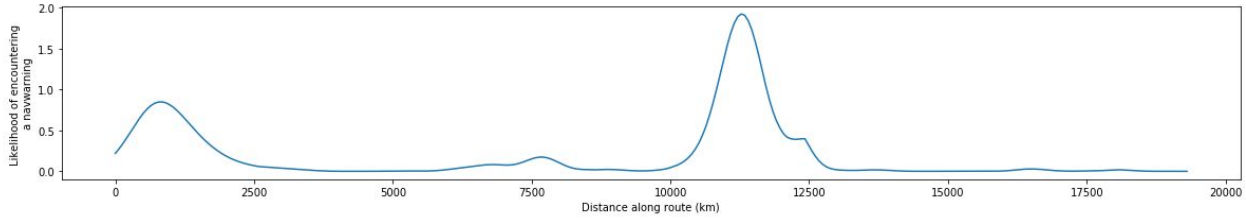
Table-11: ***Cumulative KDE risk scores by classification***

Route	Classification										
	All	0	1	2	3	4	5	6	7	8	9
Hong Kong to Southampton	57.3179	0.0000	7.2630	1.2042	23.4651	7.8528	107.3658	207.9340	12.3373	206.5523	31.0412
Rotterdam to New York	0.2649	0.0000	0.0037	0.0000	0.7019	0.0000	0.0000	0.0000	0.0275	0.0150	1.7976
Houston to Shanghai	1.1815	9.3609	0.0000	0.0000	1.4406	0.0000	0.0047	0.0000	0.0000	0.6065	10.9253

Values represent the sum of likelihoods sampled at 0.01 radian intervals along the route linestring.

Figure-26: ***KDE risk profiles, Hong Kong to Southampton***

All navwarnings



Classified navwarnings

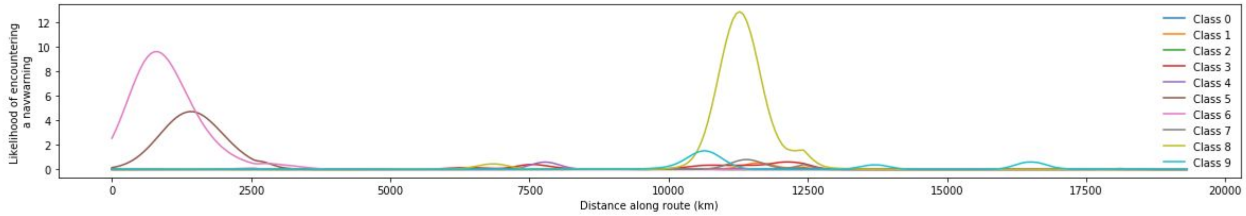
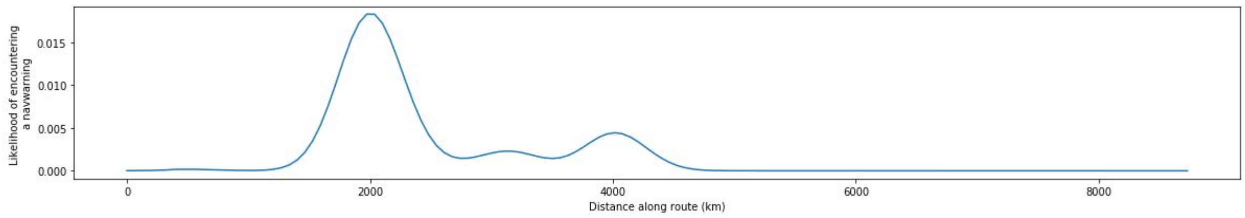


Figure-27: ***KDE risk profiles, Rotterdam to New York***

All navwarnings



Classified navwarnings

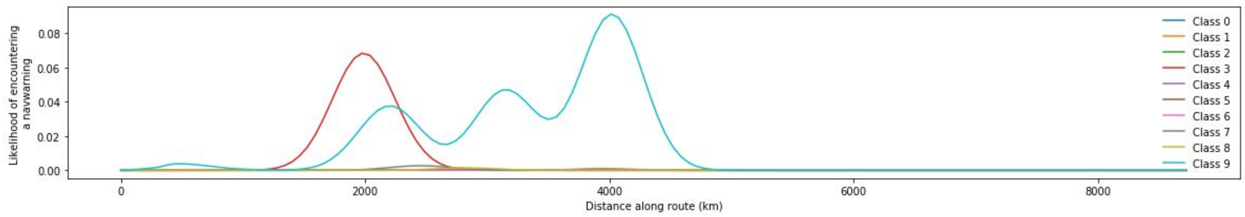
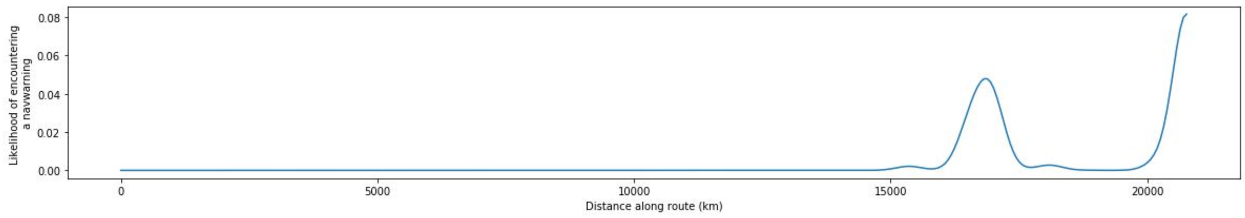
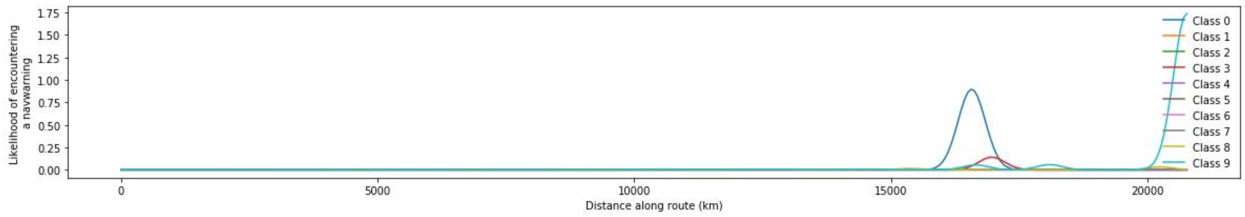


Figure-28: ***KDE risk profiles, Houston to Shanghai***

All navwarnings



Classified navwarnings



Further Research

Numerous opportunities have been unearthed for continued research in the quantitative assessment of risk for mariners at sea. Using our risk forecasting models, along with equations describing fuel usage, staff costs, insurance costs, and so on in a linear programming problem would allow the optimal route and steaming speeds for any origin-destination pair to be calculated, an application with potential to improve efficiency and safety in the global logistics industry. Further investigation into local point patterns of navwarnings might provide a more sophisticated understanding of risk, especially in highly-trafficked areas such as western Europe, the Suez and Panama Canals, and the Straits of Hormuz, Malacca and Singapore and Bab al-Mandeb.

Discussion

This project has been concerned with maritime navigational warnings, more specifically their content, extent and the potential implications of the two.

The first project extracted topics based on the messages in the warnings, which were used to label all observations. Compared with results based on K-NN, Latent Dirichlet Analysis yielded decent similarity as measured by the words found in top 4 and top 10.
These labels were used in the second project to assess the quality of the clusters detected using DBSCAN. The cluster analysis has been performed on each of the five regional areas, described in the introduction, instead of the broad area that the total warnings covers because vast distances are not captured well by the DBSCAN algorithm. Clusters are detected, albeit with varying quality, validated by the homogeneity (hs) and completeness score (cs). It is natural to consider if the varying quality of the clusters are due to the quality of the *ground truth labels* (actual class labels), from section 1, or due to the nature of the warnings. Looking at Table-5 and Table-6, showing silhouette scores (ss) and hs/cs, shows that the ss is fairly high and hs/cs is low, with average values of 0.65 and 0.25/0.33. The level of the scores indicates that the observations within a cluster fits the cluster it is assigned to (ss close to 1), but that the clusters are ambiguous when containing a singular class (low hs/cs). Looking at the scores for the individual areas indicates that regional navigational areas perhaps are not granular enough, and that the cluster

analysis is likely to benefit from finer division of the world than what the Navigational Warnings limits[1] prescribes.

Another suggestion is that the class labels derived from the topic modelling are not suitable. However, Table-3 indicates that fewer classes could be present. Equivalent results obtained using fewer classes resulted in poorer scores (Table-7), indicating that at this scale, fewer classes does not produce more homogenetic or complete clusters. Nothing indicates that there should be more classes. The labels reflect the topics detected, and the number of topics empirically chosen using LDA. It is likely that this procedure is too simple, and an alternative way to detect the number of topics, such as Hierarchical Dirichlet Process (HDP), had yielded different labels, which could be an opportunity for further investigation.

There is no doubt that the scale on which the analysis is conducted has the greatest implication for the results, i.e. it would be beneficial with a finer subdivision of the areas, albeit no such exists in simple format. Footnote 1 suggests 24 areas, but no boundaries are publicly available, limiting this extension, making it a topic for future research.

Focusing on the risk profiling of the three major shipping routes, shows that the shipping routes have varying risk, independently of the method used to determine the risk profile. Comparing Table-10 and Table-11 shows an interesting insight, namely that the simple quadrat risk method catches the takeaway from the first two projects, i.e. no area nor cluster contains unambiguous classes.

The KDE method is fitted based on classes, which covers enormous areas, which understates the risk in certain areas. The risk with KDE is that the method itself does not tell if the characterisation is correct, i.e. the fitting on the data. Forcing one distribution over something which is better characterised by multiple distributions, implies a result which is either smoothed or skewed, which in this case understates the risk. The understatement of risk in seen from Figure-25, where no profound areas of risk is located for class-3, yet this class covers heavily trafficked areas as Eastern Asia, and Class-3 is the class containing most observations. The

[1] https://msi.nga.mil/MSISiteContent/StaticFiles/Images/navwarnings.jpg

takeaway is that the KDE method, as well, would benefit from a finer division of the areas, to better characterise the risk.

Conclusion

Topic modelling of the messages contained in the warnings has been explored, and with no single way of determining the topics in a corpus, an alternative way has been explored, by letting the data reveal the number of topics itself. The topics found using LDA were the basis for labelling the observations. However, this is no guarantee of the quality of the labels, as no actual labels exists, even though the topics found by LDA had decent similarity with results from K-NN. The labels made it possible to evaluate clusters found by DBSCAN, which showed that no cluster contained singular classes. This finding underlines the initial suspicion that warnings of the same type aren't limited to certain areas, as the variation in warnings reflects the dynamic beast that the seven seas are. Furthermore, the labels served as a means to creating the risk profiling tool, which in turn allowed quantification of the risk along any given route at sea.

This project has explored and investigated some interesting and important aspects of life at sea, even though there is room for improvement in all methodologies explored. This analysis serves as a minor step toward increasing safety and decreasing risk for those who explore, transport and operate at sea.

References

Cai, Zhiqiang et al., 2016, *Can Word Probabilities from LDA be Simply Added up to Represent Documents?,* Proceedings of the 9th International Conference on Educational Data Mining.

Dasarathy, B. V., 1991, Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, Las Alamitos, California.

Ester et al. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. [Online]. University of Munich. Available from: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf [Accessed 10 April 2019]

He, Ji et al., 2004, *Unsupervised Learning for Document Classification: Feasibility, Limitation and the Bottom Line.*

Human Environment and Transport Inspectorate. 2019. *Annex 1 IMO/IHO World-Wide Navigational Warning Service - Guidance document.* [ONLINE]. Available at: https://puc.overheid.nl/nsi/doc/PUC_1444_14/2/. [Accessed 26 April 2019.]

International Hydrographic Organization. 2014. *Joint IMO/IHO/WMO Manual on Maritime Safety Information*. [Online]. Available at: https://www.iho.int/iho_pubs/draft_pubs/Joint%20Manual%20on%20MSI%20_Clean%20Version_%20_2_.pdf. [Accessed 15 April 2019.]

Lerner, Michael. 2013. *Histograms and kernel density estimation KDE 2.* [Online]. Available from: https://mglerner.github.io/posts/histograms-and-kernel-density-estimation-kde-2.html?p=28. [Accessed 15 April 2019.]

Mukhurjee, Paromita. 2017. *Important Points For Dealing With Navigational Warnings On Ships.* [ONLINE]. Available at: https://www.marineinsight.com/marine-navigation/important-points-dealing-navigational-warnings-ships/ [Accessed 10 April 2019.]

National Geospatial Intelligence Agency. 2019. *Broadcast Warnings.* [ONLINE]. Available at: https://msi.nga.mil/NGAPortal/MSI.portal?_nfpb=true&_st=&_pageLabel=msi_portal_page_63. [Accessed March 1 2018.]

Paul, S. 2018. *DBSCAN: A Macroscopic Investigation in Python*. [Online]. Available from: https://www.datacamp.com/community/tutorials/dbscan-macroscopic-investigation-python [Accessed 19 April 2019.]

Rodrigue, Jean-Paul. 2017. *The Geography of Transport Systems*. Routledge. [Online]. Available from: https://transportgeography.org/?page_id=5955. [Accessed 15 April 2019.]

Rosetta Code. 2019. *Haversine formula.* [Online]. Available from: https://rosettacode.org/wiki/Haversine_formula#Python. [Accessed 15 April 2019.]

Rosenberg, A. and Hirschberg, J. 2007. V-Measure: A conditional entropy-based external cluster evaluation measure. [Online] Department of Computer Science, New York. Available from: https://www.aclweb.org/anthology/D07-1043 [Accessed 10 April 2019]

Scikit-learn. 2018a. *sklearn.neighbors.DistanceMetric*. [Online] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html [Accessed 15 April 2019]

Scikit-learn. 2018b. *Sklearn.metrics.silhouette_score*. [Online] Available from: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html [Accessed 15 April 2019]

Scitkit-learn. 2018c. *Sklearn.metrics.adjusted_rand_score*. [Online]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html [Accessed 15 April 2019]

Scikit-learn. 2018d. *2.3.9.3. Homogeneity, completeness and v-measure*. [Online]. Available from: https://scikit-learn.org/stable/modules/clustering.html [Accessed 15 April 2019]

Scipy. 2019. *Scipy.stats.gaussian_kde*. [Online]. Available from: https://docs.scipy.org/doc/scipy-0.18.1/reference/generated/scipy.stats.gaussian_kde.html. [Accessed 15 April 2019.]

Sheng, P. and Yin, J. 2018. Extracting Shipping Route Patterns by Trajectory Clustering Model Based on Automatic Identification System Data. *Sustainability.* 10(7), pp 1-13.

Stewart, Paige. 2018. *Major Shipping Routes for Global Trade*. [Online]. Available from: https://arcb.com/blog/major-shipping-routes-for-global-trade. [Accessed 21 April 2019.]

Sui, D. 2009. 'Ecological Fallacy.' *Elsevier Ltd*. pp 291-293.

VanderPlas, Jake. 2016. *Python Data Science Handbook.* O'Reilly Media. https://jakevdp.github.io/PythonDataScienceHandbook/05.13-kernel-density-estimation.html

Wikipedia. 2019. *Gaussian function*. [Online]. Available from: https://en.wikipedia.org/wiki/Gaussian_function. [Accessed 20 April 2019.]

Appendix

Categories:

1. **Casualties to lights, fog signals, buoys and other aids to navigation affecting main shipping lanes**
2. **The presence of dangerous wrecks in or near main shipping lanes and, if relevant, their marking.**
3. **Establishment of major new aids to navigation or significant changes to existing ones, when such establishment or change might be misleading to shipping.**
4. **The presence of large unwieldy tows in congested waters.**
5. **Drifting hazards (including derelict ships, ice, mines, containers, other large items etc.).**
6. **Areas where search and rescue (SAR) and anti-pollution operations are being carried out (for avoidance of such areas).**
7. **The presence of newly discovered rocks, shoals, reefs and wrecks likely to constitute a danger to shipping, and, if relevant, their marking.**
8. **Unexpected alteration or suspension of established routes.**
9. **Cable or pipe-laying activities, seismic survey, the towing of large submerged objects for research or exploration purposes, the employment of manned or unmanned submersibles, or other underwater operations constituting potential dangers in or near shipping lanes.**
10. **The establishment of research or scientific instruments in or near shipping lanes.**
11. **The establishment of offshore structures in or near shipping lanes.**
12. **Significant malfunctioning of radio-navigation services and shore-based maritime safety information radio or satellite services.**
13. **Information concerning events which might affect the safety of shipping, sometimes over wide areas, e.g. naval exercises, missile firings, space missions, nuclear tests, ordnance dumping zones, etc.**
14. **Operating anomalies identified within ECDIS including ENC issues.**

15. Acts of piracy and armed robbery against ships.

16. Tsunamis and other natural phenomena, such as abnormal changes to sea level.

17. World Health Organization (WHO) health advisory information

18. Security-related requirements.

Proposed Categories:

Proposal 1:

- Instruments: 1, 3, 8,
- Operations: 4, 6, 9, 10, 11
- Safety: 2, 5, 7, 13, 15, 16, 17, 18
- Malfunction: 8, 12, 14
- Emergency

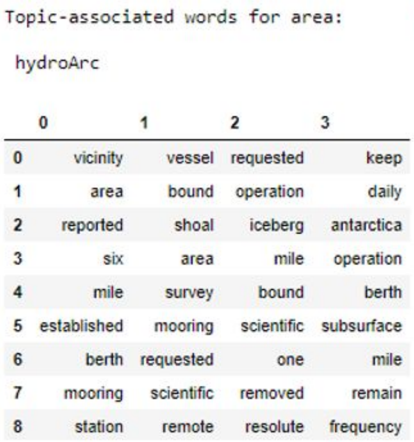
Proposal 2:

- Navigational change: 1,3,8
- Fixed hazards: 2,7
- Dynamic hazards: 4,5,16
- Scientific and commercial operations: 9,10,11
- Security and Safety: 6, 13, 15,17,18
- Information and communication: 12, 14

Proposal 3:

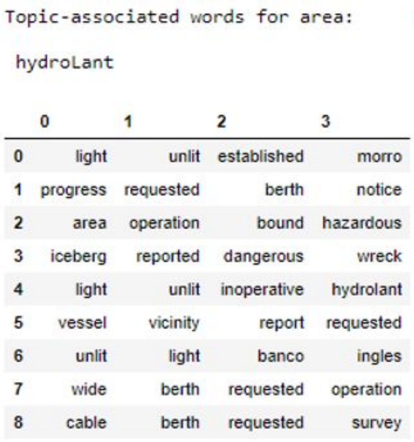
- Weather/Environmental: 5, 6, 7, 10, 16, 17
- Military/Security: 13,15, 18
- Operations: 3, 4, 9, 11
- Unexpected hazards: 1, 2, 8, 12, 14

Table-12 – ***Comparison of topic-associated words for the subareas***



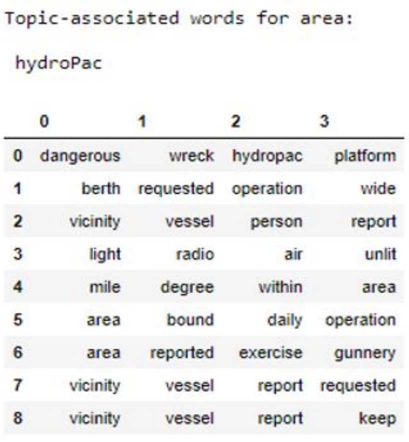
```
Topic-associated words for area:

 hydroArc
```

	0	1	2	3
0	vicinity	vessel	requested	keep
1	area	bound	operation	daily
2	reported	shoal	iceberg	antarctica
3	six	area	mile	operation
4	mile	survey	bound	berth
5	established	mooring	scientific	subsurface
6	berth	requested	one	mile
7	mooring	scientific	removed	remain
8	station	remote	resolute	frequency



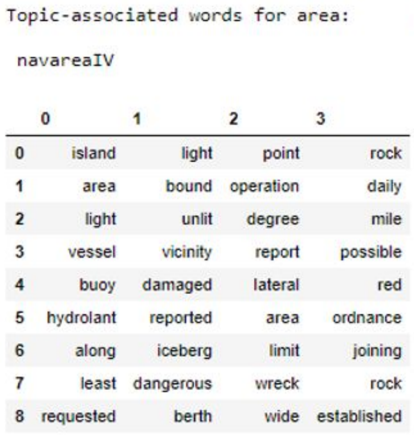
```
Topic-associated words for area:

 hydroLant
```

	0	1	2	3
0	light	unlit	established	morro
1	progress	requested	berth	notice
2	area	operation	bound	hazardous
3	iceberg	reported	dangerous	wreck
4	light	unlit	inoperative	hydrolant
5	vessel	vicinity	report	requested
6	unlit	light	banco	ingles
7	wide	berth	requested	operation
8	cable	berth	requested	survey



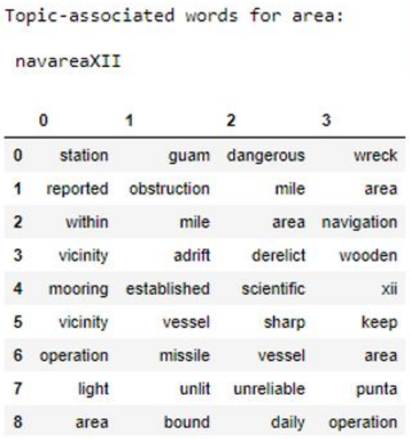
```
Topic-associated words for area:

 hydroPac
```

	0	1	2	3
0	dangerous	wreck	hydropac	platform
1	berth	requested	operation	wide
2	vicinity	vessel	person	report
3	light	radio	air	unlit
4	mile	degree	within	area
5	area	bound	daily	operation
6	area	reported	exercise	gunnery
7	vicinity	vessel	report	requested
8	vicinity	vessel	report	keep



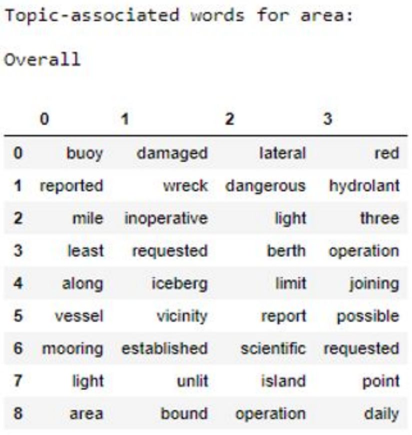
```
Topic-associated words for area:

 navareaIV
```

	0	1	2	3
0	island	light	point	rock
1	area	bound	operation	daily
2	light	unlit	degree	mile
3	vessel	vicinity	report	possible
4	buoy	damaged	lateral	red
5	hydrolant	reported	area	ordnance
6	along	iceberg	limit	joining
7	least	dangerous	wreck	rock
8	requested	berth	wide	established



```
Topic-associated words for area:

 navareaXII
```

	0	1	2	3
0	station	guam	dangerous	wreck
1	reported	obstruction	mile	area
2	within	mile	area	navigation
3	vicinity	adrift	derelict	wooden
4	mooring	established	scientific	xii
5	vicinity	vessel	sharp	keep
6	operation	missile	vessel	area
7	light	unlit	unreliable	punta
8	area	bound	daily	operation



```
Topic-associated words for area:

Overall
```

	0	1	2	3
0	buoy	damaged	lateral	red
1	reported	wreck	dangerous	hydrolant
2	mile	inoperative	light	three
3	least	requested	berth	operation
4	along	iceberg	limit	joining
5	vessel	vicinity	report	possible
6	mooring	established	scientific	requested
7	light	unlit	island	point
8	area	bound	operation	daily