Mining Vessel Trajectory Data for Patterns of Search and Rescue

*Konstantinos Chatzikokolakis**, Dimitrios Zissis*†, Giannis Spiliopoulos* and Konstantinos Tserpes***
** MarineTraffic, London, United Kingdom*
Email:{konstantinos.chatzikokolakis, giannis.spiliopoulos} @marinetraffic.com
† Department of Product and Systems Design Engineering, University of the Aegean, Syros, Greece
Email: dzissis@aegean.gr
*** Department of Informatics and Telematics, Harokopio University of Athens, Greece*
Email: ktserpes@hua.com

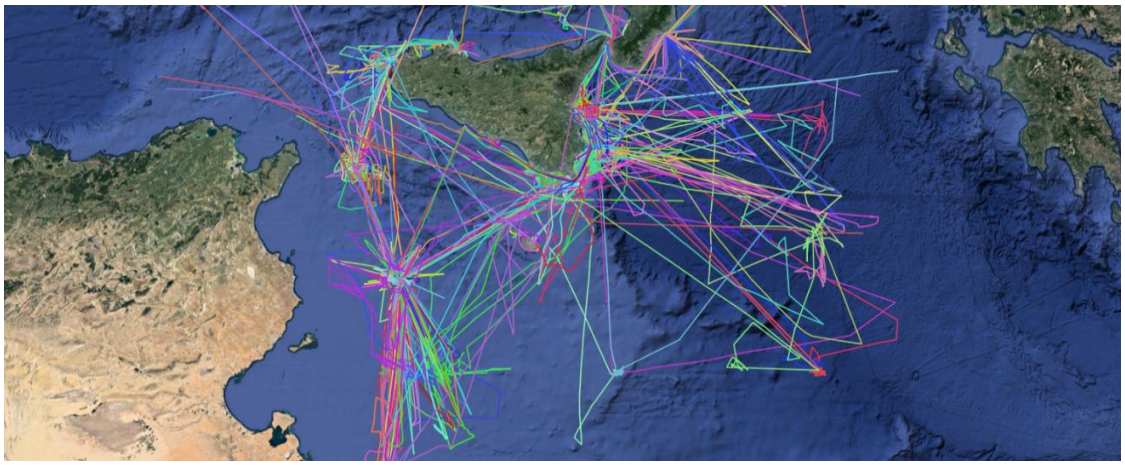

Figure 1. Visualization of SAR activity in the Mediterranean Sea during July-September 2015

ABSTRACT

The overall aim of this work is to explore the possibility of automatically detecting Search And Rescue (SAR) activity, even when a distress call has on yet been received. For this, we exploit a large volume of historical Automatic Identification System (AIS) data so as to detect SAR activity from vessel trajectories, in a scalable, data-driven supervised way, with no reliance on external sources of information (e.g. coast guard reports). Specifically, we present our approach which is based on a parallelised, nonparametric statistical method (Random Forests), which has proved capable of achieving prediction accuracy rates higher than 77%.

1 INTRODUCTION

For many years, North Africa has served as the jumping off point for refugees and migrants hoping to cross the Mediterranean Sea to Europe. Since the Syrian war in 2011, there has been a rapid increase in the number of people crossing; a trend which is not expected to stop any time soon. According to the UN Refugee Agency, this year alone, at least 2,030 people have died or gone missing on the voyage, with the greatest number of fatalities occurring along the so-called Central Mediterranean Route, through Libya [23]. Although under maritime law, any vessel in the area of a vessel in distress is obliged to offer assistance, numerous national and international missions have been launched on the EU borders and in the international waters of the Mediterranean, so as to assist in Search and Rescue (SAR) operations, such as Operation Mare Nostrum led by Italy, Operation Triton led by Frontex, NATO Operation Sea Guardian and the EU operation Sophia. Many of these operations were not designed with SAR as a primary mission goal. Due to this numerous Non-Governmental Organisations (NGO) have stepped in and have been performing SAR operations in the area; these include

Migrant Offshore Aid Station (MOAS), Doctors Without Borders, Sea-Watch and others. According to the UNHCR an overall 41% of those rescued have been by the NGOs.

Recently though concerns have been raised about the possible interactions between NGOs and smugglers. A report published by the EU agency Frontex stated that there were "clear indications before departure on the precise direction to be followed in order to reach the NGOs' boats"[4]. According to this same report, during 2015, and the first months of 2016, smuggling groups instructed migrants to make satellite phone calls to the Maritime Rescue Coordination Centre (MRCC) in Rome so as to initiate targeted rescues on the high seas. During this period, SAR operations were mainly undertaken by Italian law enforcement, EUNAVFOR Med or Frontex vessels with NGO vessels involved in less than 5% of the incidents. From June to October 2016, however, the pattern was reversed. "Satellite phone calls to MRCC Rome decreased sharply (down to 10%) and NGO rescue operations rose significantly to more than 40% of all incidents. Since June 2016, a significant number of boats were intercepted or rescued by NGO vessels without any prior distress call and without official information as to the rescue location" according to Frontex [4].

Maritime Domain Awareness (MDA) is the effective understanding of activities, events and threats in the maritime environment that could impact global safety, security, economic activity or the environment [5]. Whilst in the past, MDA had suffered from a lack of data, current tracking technology has transformed the problem into one of an overabundance of data and information. Currently, huge amounts of structured and unstructured data, tracking vessels during their voyages across the seas, are becoming available, mostly due to the Automatic Identification System (AIS) that vessels of specific categories are required to carry. The AIS is a collaborative, self-reporting system that allows maritime vessels to broadcast their information to nearby vessels and coastal based stations [26]. AIS transceivers allow real time information exchange between vessels and shore based stations through digital radio signals transmitted over dedicated channels in VHF band. The major challenge faced today, is exploiting these vast amounts of data and transform it into actionable information. Discovering patterns emerging within these huge datasets is of great importance so as to provide critical insights into the patterns vessels follow during their voyages at sea.

The main objective of our work is to explore the possibility of leveraging these huge mobility datasets so as to automatically detect vessels performing SAR operations. Towards this direction we adopt a practical data mining and machine learning approach which is capable of overcoming the shortcomings and difficulties presented by AIS data (highly skewed, non-uniform, reception errors etc.) [6]. In sum, this work presents novelties on two fronts:

- **Domain Specific**: The overall aim of this work is to explore if it is possible to automatically detect SAR activity from open data (such as AIS), even when a distress call has not been received. This work has an important social impact, as it can help improve coordination of SAR efforts and understanding of implicated activities (e.g. response time).
- **Algorithmic**: We extract patterns of "rescue-like behavior" from billions of records of spatio-temporal (AIS) data and apply Random Forests, which is a parallelised nonparametric statistical method, evaluated as capable of achieving prediction accuracy rates of more than 77%, even when applied to large volumes of highly skewed geospatial data. To the best of the authors knowledge, no previous work has considered deriving SAR activity from AIS data.

The rest of the paper is organized as follows: Section 2 shortly presents previous work in this domain, while Section 3 describes our approach and Section 4 presents the preliminary results while section 5 concludes this paper by briefly outlining the main contributions of this work and suggesting future improvements.

2 RELATED WORK

The rise in the availability of larger quantity and better quality mobility data, has increased the interest of researchers in data driven knowledge discovery. Some of the typical mining tasks in the spatio-temporal context include, frequent pattern discovery, trajectory pattern clustering, trajectory classification, forecasting, and outlier detection. Recent works on pattern discovery are based on online event recognition systems that recognize suspicious and illegal vessel activities of compressed routes (i.e., only critical points of routes are preserved)[17]. Although this solution identifies complex events, it does not classify those to specific vessel operations (e.g. tugging, fishing, search and rescue, etc.). The merits of this work have been extended in where vessels' moving pattern analysis is performed through an ontology-based system[14]. Trajectory classification, includes constructing a model capable of predicting the class labels of moving objects based on their trajectories and other features [9]. Trajectory classification has been applied in many mobility applications and numerous methods have been proposed throughout the given literature, however less attention has been paid to the maritime domain and classifying a vessel's type with regards to its trajectory. For example, in [9], authors propose a feature generation framework TraClass for trajectory data from satellite images and trace gas measurements, which generates a hierarchy of features by partitioning trajectories and explores two types of clustering: (1) region-based and (2) trajectory-based. In

this paper, hierarchical region-based and trajectory-based clustering after trajectory partitioning is performed, and a vessel classification rate as high as 84.4% is reported, but unfortunately information on how many vessel types are included in the dataset is not provided [9].

Several studies have proved the value of using AIS data for data driven knowledge discovery in this domain [12, 15, 16]. An interesting trajectory classification case that has caught researchers attention, is that of fishing activity detection; especially for applications such as illegal fishing, where the task can be defined as given a ship trajectory T, predict a label y_i for each data point t_i where $y_i \in$ {Fishing, NonFishing}[21]. In [21], authors develop three different models to detect potential fishing behavior according to the type of fishing activity; for trawlers a Hidden Markov Model (HMM) is developed using vessel speed as observation variable; for longliners a pattern recognition approach named Lavielle's algorithm has been applied; and for purse seiners a multi-layered filtering strategy based on vessel speed and operation time was implemented. Validation against expert-labeled datasets showed average detection accuracies of 83% for trawler and longliner, and 97% for purse seiner. Although these methods were designed for wide applicability, high accuracy results are only achieved by preprocessing AIS data, where wrong detections, noise and faulty out-of-bounds data (e.g. observations on land) are previously removed [21]. The use of AIS data poses a series of data management and data processing challenges linked to the treatment of large volumes of data which may heavily reduce the applicability of the approach. Many traditional data mining approaches assume that the underlying data distribution is uniform and spatially continuous. This is not the case for global AIS data, as it is often to have large geographical coverage gaps, message collisions or erroneous messages especially when processing large areas [18, 25].

In [11] Mazzarella, Vespe, Damalas and Osio focus on discovering and characterising fishing areas by exploiting historical AIS data broadcast by fishing vessels. Specifically, they focused on detecting the behavior of fishing boats that are probably actively fishing. The methodology used for the identification of fishing activity was based on assuming a fishing behaviour highly dependent and characterised by speed. Detecting changes and frequency of speed could help identifying which part of the vessel track can be considered as fishing and which not [13].Their approach relies on DB-SMoT [20] and DBSCAN [3] but unfortunately it is difficult to evaluate the overall accuracy of their results due to the limited availability of ground truth data.

In [24], authors make use of trajectory kernels in combination with a Support Vector Machines (SVM) to detect fishing activity from AIS data, which was collected in a 50km radius around the Port of Rotterdam. For their classification experiments they use the four most common vessel types: cargo ship, tanker, tug and law-enforcement vessel with the best accuracy score being 76.25%. Jiang, Silver, Hu, De Souza, and Matwin in [8], also make use of AIS data and compare Autoencoders with SVMs and Random Forests. In their work they suggest that autoencoders can perform at least as well as and sometimes better than SVM and Random Forests on classification fishing activities, achieving up to 85% accuracy [8]. However, the nature of the autoencoders is to capture as much information as possible and not as much relevant information as possible and since this work utilised only a small dataset it would be difficult to have only a small part of the input that is relevant to the considered problem. Furthermore, SVMs do not work well with categorical features and often fail to handle larger datasets as they pose significant memory requirements and computational complexity in such cases. Other studies indicate the superiority of Random Forests when used for classification tasks, compared to SVMs and back propagation neural networks [10].

Random Forests, which are based on decision trees combined with aggregation and bootstrap ideas, were first introduced by Breiman in 2001 [2]. They are a powerful nonparametric statistical method allowing to consider in a single and versatile framework regression problems, as well as two-class and multi-class classification problems [19]. Random Forests can deal with large numbers of predictor variables even in the presence of complex interactions, and have been applied successfully in genetics, clinical medicine, and bioinformatics within the past few years. Random Forests have been shown to achieve a high prediction accuracy in such applications and to provide descriptive variable importance measures reflecting the impact of each variable in both main effects and interactions [22]. They are considered capable of good accuracy, relatively robust of outliers and noise, can be pararellised and are thus considered suitable data mining algorithm for big data [1, 2].

3 PROPOSED APPROACH

Our aim is to explore the possibility of automatically detecting SAR activity from open data (such as AIS), even when a distress call has not been received. The task can be formulated as given a set of vessel trajectories T, predict a label y_i for each trajectory t_i where $y_i \in$ {SAR, Non-SAR}. A trajectory T is a set of AIS messages monitoring a vessel's movement from a departure port to a destination port.

3.1 Dataset description and processing requirements

According to International Organisation for Migration, more than 360.000 migrants have arrived to EU by sea in 2016, mainly at Italy, Greece and Spain [7]. With respect to the spatial coverage, our analysis has been focused on a bounding box covering the Central Mediterranean Route,

where most of the refugee fatalities have been observed. Figure 2 below illustrates the bounding box taken into account in conjunction with the refugee fatalities in 2016. It should be noted that our approach relies only on AIS data and the migration fatalities dataset visualised in Figure 2 is used only as a reference to define the bounding box area.

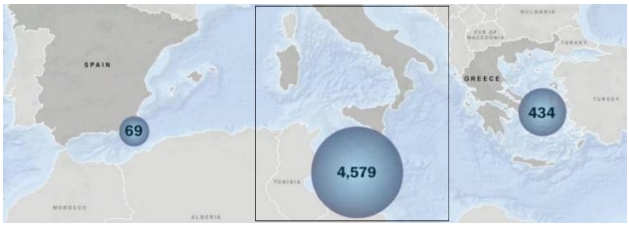

Figure 2. Spatial coverage in conjunction to migration fatalities for 2016

The considered dataset includes the 6 most relevant to navigation AIS messages out of the 27 AIS message types defined in ITU 1371-4 report [26], which are used in approximately 90% of AIS-based scenarios. More specifically, the dataset includes messages of types 1, 2, 3, 5, 18, and 19 out of which 1, 2, 3, 18 and 19 are position reports, including latitude, longitude, speed-over-ground (SOG), course-over-ground (COG), and other fields related to ship movement, while type 5 messages correspond to static-and voyage information, including the IMO identifier, radio call sign, name, ship dimensions, ship and cargo types.

Each vessel's type can be deduced using the information contained in these messages that the vessel is transmitting. This piece of information, typically referred to as AIS SHIPTYPE, usually consists of two digits, the first one ranging from 1-9 indicates the general category of the subject vessel (e.g., Special Category, Passenger, Cargo, etc.), while the second one provides additional information regarding the vessel's type of cargo in certain vessel categories (e.g., Cargo Ships, Tankers, etc.). The vessel's crew or the accountable officer are responsible for correctly entering information into the AIS transponder and although there are explicit types for SAR vessels, it is frequently the case that vessels participating in SAR operations are not declared as such. Furthermore, only the fact that a vessel's type is SAR does not necessarily infer that each voyage of the vessel is linked to SAR operations (e.g., such vessel could travel between ports for maintenance purposes). Data volume included in our analysis demands large computational power and a parallel processing approach, due to the fact that traditional analytics fail to handle such volumes of data in a considerable time frame. Consequently, we have deployed our approach in Microsoft Azure which is a distributed computing framework capable to process large amount of data fast. Particularly our system included two Head D12v2 nodes, and six D13v2 Worker nodes summing up to a Spark cluster with 56 cores and 392GB memory (in total). The worker nodes have 8 processing cores and 56GB of memory each and the head nodes have 4 processing cores and 28GB of memory each.

3.2 Data processing and analysis

The dataset used for this study consists of all the voyages of 2016 that intersect with the bounding box shown in Figure 2. More specifically this includes 275.657 (SAR and non-SAR, according to the reported AIS SHIPTYPE) voyages made by 12.291 vessels. These correspond to 54.766.629 AIS observations. After processing the initial data we used an algorithmic approach we have introduced in [6], which determines departure and destination port for each AIS message, thus transforming them into specific voyages. Each voyage includes the vessel's trajectory as well as its static and voyage information described in the previous sub-section. Then, a data curation process was performed, to discard voyages with insignificant amount of positions (e.g. statistically too few to be representative). More specifically, all the voyages that included less than 50 positions were removed (as the geographical area selected covers a distance of over 1500 kilometers, trajectories with only 50 reported positions translate to a sample rate of less than one sample per hour). Such voyages suffer from gaps of communication, which will affect the accuracy and the effectiveness of the proposed. After the curation process the dataset included 114.762 voyages, performed by 10.816 vessels, containing 52.505.718 AIS records. However, the SAR data available in this geographic area for 2016 are more than 100-times less compared to the data of non-SAR voyages. More specifically, the dataset includes 114.377 non-SAR voyages of 10.788 vessels which include a total of 52.429.521 AIS messages while the SAR voyages are 385 made by 28 vessels with 75.797 AIS records. For evaluating the approach, the dataset was split into training and test data; the training set included 70% of the SAR voyages and in order to avoid having imbalanced training data or having imbalanced evaluation metric of the classifier (e.g. true positive rate at some false positive threshold), we subsampled the non-SAR voyages (i.e., randomly selecting a subset) included in the training data. Particularly the training data included 1.544 non-SAR voyages and 261 SAR voyages made by 949 and 26 distinct vessels respectively. The rest of the data (i.e. 30% of the SAR voyages and all the non-SAR voyages not included in the training set) constituted our test data.

For all the records in the dataset we filter the following attributes which will be used in our analysis for distinguishing SAR patterns:

a. Ship id: This is a unique identifier for each vessel

b. Ship type: This is a two-digit code that corresponds to the general category of the vessel and the vessel's type of cargo in certain vessel categories
c. Latitude, Longitude: These represent the geographic location of the vessel
d. SOG: This is the speed over ground of the vessel measured in knots
e. COG: This is the course over ground of the vessel measured in degrees with 0 corresponding to north
f. Heading: This attribute represents the ship's heading in degrees with 0 corresponding to north
g. Timestamp: This is the full UTC timestamp that the AIS message was received by MarineTraffic

It should be noted that COG and Heading may be different, due to weather conditions such as wind speed and direction, wave height and currents (e.g. when vessels are drifting). COG on the one hand is the actual moving direction of the vessel, while heading simply indicates where the ship is pointing compared to north. Based on all these attributes and in conjunction with other datasets that assist on determining the boundaries of a port the following additional attributes were calculated:

a. Departure port id: This is a unique identifier of the port from which the vessel departed
b. Departure timestamp: Full timestamp of the first AIS message outside of departure port geometry
c. Departure port name: This is the name of the departure port
d. Departure port type: This attribute determines the type of the port (e.g., port, anchorage, etc.)
e. Departure country code: This attribute indicates the country of the departure port

Similar attributes related the arrival of each vessel to a port have been also calculated.

3.3 SAR Motion analysis

All these attributes have been used to transform raw positional data into vessel voyages. However, in order to distinguish SAR trajectories from other voyages it has been required to delve into more details on the motion patterns during SAR operations and focus on maneuverability of such vessels. The methodology used for the identification of SAR activity is based on assuming that SAR behaviour is highly dependent and characterised by frequency of speed changes, frequency of turns, departing and arriving at the same port or anchorage and voyage duration. Detecting changes and frequency of speed as well as departing and arriving at the same port will help distinguishing SAR trajectories from typical voyages (i.e., travelling from one port to another). However, there are also other types of ships that may follow similar patterns. For instance, inland vessels tend to have frequent changes in course over ground and heading due to the voyage area topography. Another example are tugboats that maneuver other vessels by pushing or towing them. Such vessels typically operate in crowded port or narrow canals and perform various maneuvers leading to increased frequency of turns. Furthermore, tugs typically have the same departure and arrival port as they are called to leave a port (i.e., depart), reach the vessel to be towed (or pushed) and return to the same port. One of the distinguishing factors between such vessels and SAR is the voyage duration. In many SAR operations, once vessels recover migrants from sea, they return to the same port from which they departed so as to disembark rescued people and return back to the SAR operation area. Furthermore, SAR vessels patrolling tend to have a steady course, while when they are engaged in rescuing operation they perform complex maneuvers to collect migrants. In some cases, it has been observed that vessels patrolling an area, may be at open sea (i.e., outside of port boundaries) for several days (or even weeks) traveling in a rather small bounding box (compared to the overall time of their voyage).

Based on those characteristics, we produced some additional attributes that have been considered as possible features for the classification process. For each voyage we have ordered the AIS messages received chronologically and we calculated COG, SOG and Heading deltas for each pair of (chronologically) consecutive messages. Negative values in the COG delta feature indicate moving to the left, while positive values indicate moving to the right. Similarly, negative values in the SOG delta feature indicate speed decrease, while positive values indicate speed increase. Finally, negative values for the Heading delta imply a turn of ship's heading to the left, while positive values indicate a turn to the right. In our analysis we use the absolute values of COG, SOG and Heading deltas, which capture the magnitude of change of the corresponding attributes. In addition, two extra features have been added to the dataset. The first one is a Boolean value indicating whether the vessel has the same departure and arrival port has been added to the dataset, while the latter one is the voyage duration.

After constructing these last features, we were able to measure the quantiles for the COG, SOG and Heading deltas and it has been observed that SAR operation voyages have different behavior compared to other voyages. More specifically, non-SAR voyages seem to have low values even for large quantiles (i.e., 75%, 80%, 85% etc.) compared to the SAR voyages, meaning that in most observations the COG, SOG and Heading deltas are typically small, while for SAR voyages those quantiles had large values. Thus, we added to our dataset the 50%, 75%, 85% and 95% quantiles for each of those voyages.

4. RESULTS AND DISCUSSION

The focus of this work is on exploiting large volumes of historical AIS data so as to identify SAR operations from trajectories in a scalable data-driven and supervised way. Our approach is based on a parallelised, non-parametric statistical method, the Random Forests. To evaluate the approaches' performance, we conducted a series of experiments that showcase its effectiveness to unseen real-world data. Firstly, we applied a multiple fold cross validation procedure and measured the F1 score. This score given by the Equation (1) below is the weighted average of Precision and Recall taking both false positives and false negatives into account. Then, using the best model derived through the cross-validation procedure the algorithm classified the test data.

$$\text{F1 Score} = 2 * \frac{\text{Recall} * \text{Precision}}{Recall + Precision} \qquad (1)$$

4.1 Random Forest training and validation

The training dataset described in subsection 3.2 has been used to train and validate the Random Forest model using the features analysed in subsections 3.2 and 3.3. The dataset has been repeatedly partitioned, following the well-known k-fold cross-validation procedure, into training and validation pairs. The partitioning process has been repeated 5 times (i.e. 5-fold cross validation) each time leading to different training and validation pairs. In each partition we have split the dataset into five parts. Four of them used as training set and one of them as validation set with the former set utilised to create the model of the Random Forest and the latter one used for predicting the class of the observations and comparing it against its actual value. Each Random Forest model derived has 10.000 trees and the F1 metric has been measured, leading to an average score of 0.946 for all the 5 folds. The best model derived from the cross-validation process has been retained and used for predicting the values of the test set. Finally, it should be noted that, although classification has not been applied afore for SAR missions, the Random Forest algorithm shows similar performance compared to other classification schemes used for identifying other types of vessels' motion patterns such as fishing [8][21][24].

4.2 Random Forest prediction model evaluation

The best model obtained through the 5-fold cross validation process has been used for predicting the labels of the test dataset. To evaluate the performance of the model against first seen data, we measured the F1 score, the Accuracy, the weighted Recall and the weighted Precision presented in Table 1 below. Accuracy is the most intuitive performance measure giving the ratio of correctly predicted observation to the total observations. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High Precision relates to the low false positive rate. Finally, Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

Table 1: Prediction model metrics scores

Metric	Value
F1 score	0.986
Accuracy	0.975
Weighted Recall	0.975
Weighted Precision	0.998

The results, show high scores in all the metrics. This occurs due to the highly imbalanced test dataset. More specifically it shows that the model can distinguish non-SAR voyages and classify them as such. The ROC curve and the Area Under ROC curve shown in Figure 3 below indicate also the capabilities of the derived model to classify SAR and non-SAR voyages, as the area under ROC is equal to 0.86.

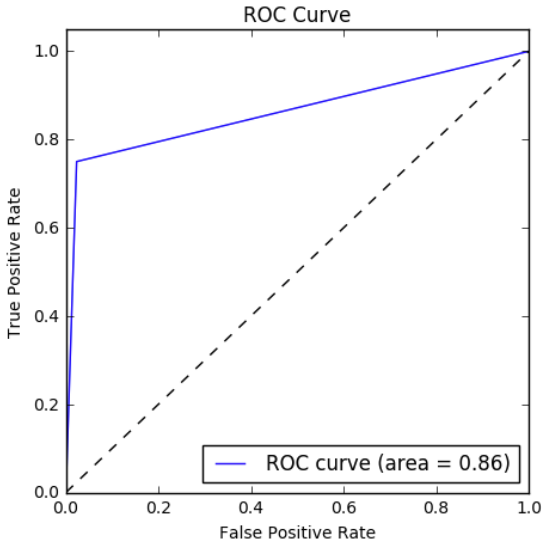

Figure 3: ROC curve and Area Under ROC curve of the Random Forest prediction model

However, since the test dataset is imbalanced, and in order to further investigate how well the algorithm identified SAR voyages we have measured the misclassification rate for each vessel type. Particularly the prediction accuracy of each vessel type class has been derived and Table 2 below includes the top-5 (i.e., with most misclassification) vessel types (i.e. false positives) and the misclassification of SAR voyages (i.e. false negatives). The

results show that the classification model labelled accurately 77,5% of the SAR voyages.

Table 2: Top-5 misclassified vessel classes

AIS Vessel type	AIS Vessel type name	# voyages	Misclassification rate (%)
51	SAR	124	22.5 (false negatives)
34	Dive Vessels	40	62.5
53	Port Tender	10	60
49	High-Speed Craft I	548	57.6
40	High-Speed Craft II	435	57.01
30	Fishing	1021	26.75

Though, the misclassification rate of the non-SAR voyages presented above is high, these classes represent a small portion of the overall test dataset, with only a few tens or hundred voyages. On the other hand, the classification algorithm achieved remarkable accuracy rate reaching up to 99.7% in classes with more voyages in the test set. Table 3 below includes the five vessel types with the most voyages in the test set and the misclassification rate for those vessel types.

Table 3: Top 5 vessels with most voyages

AIS Vessel type	AIS Vessel type name	# voyages	Misclassification rate (%)
70	Cargo	32.611	0.3
60	Passenger	17.253	1.64
71	Cargo – Hazard A	10.308	0.32
80	Tanker	9.599	1.43
69	Passenger	9.057	0.695

5. CONCLUSION AND FUTURE WORK

This work focused on the task of automatically detecting SAR vessels from maritime trajectory data. Specifically, we leveraged a large volume of historical AIS data and described our approach which is based on Random Forests, a parallelized nonparametric statistical method, with no reliance on external sources of information (e.g. coast guard reports), so as to detect vessels performing SAR operations in the Mediterranean Sea. The task was formulated as given a set of ship trajectories T, predict a label y_i for each trajectory t_i where $y_i \in \{\text{SAR, Non-SAR}\}$. Our proposed approach proved capable of classifying SAR trajectories at an accuracy higher than 77%. To the best of the authors knowledge, no previous work has considered deriving SAR activity from AIS data in a data driven approach. In the future, we will attempt to reformulate the problem towards a point based approach classification, such that given a ship trajectory T, predict a label y_i for each data point t_i where $y_i \in \{\text{SAR, NotSAR}\}$. Based on these labeled points, SAR time per area can possibly be calculated on any given scale.

ACKNOWLEDGEMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732310 and by Microsoft Research through a Microsoft Azure for Research Award.

REFERENCES

[1] A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment - IEEE Journals & Magazine: *http://ieeexplore.ieee.org/document/7557062/*. Accessed: 2017-11-30.

[2] Breiman, L. 2001. Random Forests. *Machine Learning*. 45, 1 (Oct. 2001), 5–32. DOI:https://doi.org/10.1023/A:1010933404324.

[3] Ester, M. et al. 1996. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (Portland, Oregon, 1996), 226–231.

[4] FRONTEX *Risk Analysis for 2017*.

[5] Galdorisi, G. and Goshorn, R. 2006. *Maritime Domain Awareness: The Key to Maritime Security Operational Challenges and Technical Solutions*.

[6] Giannis Spiliopoulos et al. 2017. A big data driven approach to extracting global trade patterns. (Sep. 2017).

[7] International Organization for Migration- UN *Mixed Migration Flows in the Mediterranean and Beyond*.

[8] Jiang, X. et al. 2016. Fishing Activity Detection from AIS Data Using Autoencoders. 33–39.

[9] Lee, J.-G. et al. 2008. TraClass: Trajectory Classification Using Hierarchical Region-based and Trajectory-based Clustering. *Proc. VLDB Endow.* 1, 1 (Aug. 2008), 1081–1094. DOI:https://doi.org/10.14778/1453856.1453972.

[10] Liu, M. et al. 2013. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. *Sensors and Actuators B: Chemical*. 177, Supplement C (Feb. 2013), 970–980. DOI:https://doi.org/10.1016/j.snb.2012.11.071.

[11] Mazzarella, F. et al. 2014. Discovering vessel activities at sea using AIS data: Mapping of fishing footprints. *17th International Conference on Information Fusion (FUSION)* (Jul. 2014), 1–7.

[12] Millefiori, L. et al. 2016. A distributed approach to estimating sea port operational regions from lots of AIS data. (Washington D.C., USA, 2016).

[13] Natale, F. et al. 2015. Mapping Fishing Effort through AIS Data. *PLOS ONE*. 10, 6 (2015), e0130746. DOI:https://doi.org/10.1371/journal.pone.0130746.

[14] OBDAIR: Ontology-Based Distributed Framework for Accessing, Integrating and Reasoning with Data in Disparate Data Sources (PDF Download Available):

https://www.researchgate.net/publication/319280828_OBDAIR_Ontology-Based_Distributed_Framework_for_Accessing_Integrating_and_Reasoning_with_Data_in_Disparate_Data_Sources. Accessed: 2018-02-02.

[15] Pallotta, G. et al. 2013. Traffic knowledge discovery from AIS data. *Proceedings of the 16th International Conference on Information Fusion* (Jul. 2013), 1996–2003.

[16] Pallotta, G. et al. 2013. Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction. *Entropy*. 15, 6 (Jun. 2013), 2218–2245. DOI:https://doi.org/10.3390/e15062218.

[17] Patroumpas, K. et al. 2017. Online event recognition from moving vessel trajectories. *GeoInformatica*. 21, 2 (Apr. 2017), 389–427. DOI:https://doi.org/10.1007/s10707-016-0266-x.

[18] Poļevskis, J. et al. 2012. Methods for Processing and Interpretation of AIS Signals Corrupted by Noise and Packet Collisions. *Latvian Journal of Physics and Technical Sciences*. 49, (Jan. 2012), 25–31. DOI:https://doi.org/10.2478/v10047-012-0015-3.

[19] Random Forests for Big Data - ScienceDirect: *http://www.sciencedirect.com/science/article/pii/S2214579616301939*. Accessed: 2017-11-30.

[20] Rocha, J.A.M.R. et al. 2010. DB-SMoT: A direction-based spatio-temporal clustering method. *2010 5th IEEE International Conference Intelligent Systems* (Jul. 2010), 114–119.

[21] Souza, E.N. de et al. 2016. Improving Fishing Pattern Detection from Satellite AIS Using Data Mining and Machine Learning. *PLOS ONE*. 11, 7 (2016), e0158248. DOI:https://doi.org/10.1371/journal.pone.0158248.

[22] Strobl, C. et al. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*. 14, 4 (2009).

[23] UN Refugee Agency *Refugee and migrant flows through Libya on the rise – report*.

[24] de Vries, G.K.D. and van Someren, M. 2012. Machine learning for vessel trajectories using compression, alignments and domain knowledge. *Expert Systems with Applications*. 39, 18 (Dec. 2012), 13426–13439. DOI:https://doi.org/10.1016/j.eswa.2012.05.060.

[25] Yang, M. et al. 2012. Collision and Detection Performance with Three Overlap Signal Collisions in Space-Based AIS Reception. *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications* (Jun. 2012), 1641–1648.

[26] 2001. *ITU Recommendation 1371-4, "Technical characteristics for an Automatic Identification System using time-division multiple access in the VHF maritime mobile band."* Tech. Rep. Recommendation.