Case studies

This chapter describes different case studies on the use of AIS. They range from faster economic indicators, maritime indicators, maritime statistics, inland waterway statistics, fishery and ships in distress. The studies were performed by different institutions.

- Faster economic indicators: Time in port and port traffic (UK ONS)
- Maritime indicators (UNCTAD/MarineTraffic)
- Official Maritime statistics: Port visits (Eurostat)
- Completing statistics on Inland waterways (Statistics Netherlands as part of the ESSnet Big Data II: Tracking ships 2018-2020)
- Mapping fishery activities (EU JRC)
- Ships in distress (UN Global Pulse)
- NOx, SOx, and CO2 Calculation (Université du Havre)
- Nowcasting Trade Flows in Real-Time (IMF)
- Experimental Statistics of Daily number of vessel (Statistics Denmark)

Faster economic indicators: Time in port and port traffic (UK ONS)

Background

Policymakers ask for faster economic indicators to be able to act and adjust policy more quickly in response to economic changes. The Faster indicators of UK economic activity project, led by the Data Science Campus at the Office for National Statistics (UK ONS) develops such indicators. The goal is to identify close-to-real-time data sources representing useful economic concepts to create a set of indicators for early identification of large economic changes and to provide insight into economic activity. One of these indicators is to develop an early picture of international trade in goods. For the UK, maritime shipping is an important modality in the international trade in goods. AIS provides an obvious source to explore shipping as it is available in almost real-time. UK ONS has investigated shipping activity for 10 major UK ports, focusing on two new monthly shipping indicators: 'Time-in-port' and 'Total traffic'. These indicators, derived from AIS data, are likely to offer a fast indication of the level of shipping activity, which in turn should be related to economic development. AIS data was compared to official statistics for gross value added and international trade in goods statistics.

Data

AIS data from July 2016 to August 2018 was used for 10 UK ports. AIS data was obtained from the Maritime and Coastguard Agency. Messages of types 1, 2 and 3 were used, containing the automatic position report from the ship transponder. Some quality issues were considered. Exploration of the data for UK ports specifically suggests that events in which AIS equipment was turned off (rendering a port visit partially detected or not detected at all) do not occur often and for the purpose of this research their impact on the outputs can be ignored. Additionally, some of the route information in AIS must be entered manually by crew members. This includes useful information such as destination and previous/next port of call. However, the manual entries are not always completed on time, and may not always be accurate. For example, the final route destination is reported for only 41% of journeys. To avoid any potential bias in the outputs which might arise from quality issues associated with the manual entry of data, this work has focused on developing methods relying only on the data which is generated automatically. To deal with the inherent noise present in the signal, Kalman filtering was used to improve tracking of the ship's locations and movements. The higher accuracy is a result of the estimation of a joint probability distribution over the positional parameters for each timeframe. Overall, the less noisy estimated positions enable increased fidelity in identifying events such as entering and leaving port or estimating the docking positions. To deal with such a large amount of information, approximately 28 million messages per day, Apache Spark and a Hadoop cluster were used to manage the volume and velocity of the incoming AIS data.

Methodology
The ten biggest UK port areas from 2017 were manually defined from a map for each port using the typical berth positions, resulting in rectangular bounding boxes. In the case of the port of Grimsby & Immingham, due to the distance between the sites, two bounding boxes have been defined and the presence of a ship in either of them is considered in-port state. If the ship’s location did not fall in any of the defined port bounding boxes then the in-port state was default value 0, defining a group of ships that are out of port. A ship is considered in port if its reported position was inside the port bounding box. The in-port states are marked from 1 to 10, corresponding to the port numbers. The data were grouped by the in-port state of the ship to produce the outputs for each of the ten ports individually and aggregated for all ports. The ‘time-in-port’ indicator was computed by summing all the periods of the time spent of ships having in-port state corresponding to the port over each port and each month. If a ship’s AIS transponder was switched off inside a port, the time in port was counted only if the following message received from the ship was also within the same port. This rule eliminated the outliers in data resulting from moored ships switching off their AIS equipment and later leaving port without reactivating it or from ships that for some reason change their Maritime Mobile Service Identity (MMSI) while in port.

Similarly, the ‘total traffic’ indicator was computed by grouping the data by port and month for the number of unique ships, based on MMSI number. As the total traffic indicator measured the number of unique ships entering port each month, it is not sensitive to ships that spend very long periods in port, e.g. pilot boats, or to have frequent port calls, e.g. ferries. On the other hand, the ‘time-in-port’ captures all time that ships spend in port and it may increase relative to the ‘total traffic’ indicator if either there are delays in port, or it takes longer to upload ships due to more cargo on board.

The data on port activity indicators were compared to monthly economic statistics:

- Gross value-added, chained volume measure, seasonally adjusted (source, ONS)
- Trade in goods: ONS, imports and exports, current prices defined as change in international ownership in accordance with European System of Accounts 2010 (CP, SA, source: ONS)
- UK overseas trade statistics: HMRC, imports and exports defined as the movement of goods across international borders (CP, NSA, source: HM Revenue and Customs)

**Results**

Both the month-on-month growth rate for time-in-ports and for port traffic had a weak positive relationship with GVA. Trade in goods is only one component of GVA, where it is defined by change in international ownership in this National Accounts estimate, rather than transfer across international borders. Therefore this weak positive correlation was to be expected. Seasonal adjustment of the shipping indicators, were it possible, might improve this relationship. However, as only two years of AIS was available, this is too short to perform seasonal adjustments.

Both ONS and HMRC trade estimates showed a reasonable correlation with both time-in-ports and port traffic indicators. This was a bit of a surprise, as it was expected that shipping would correlate better with the movement of goods, instead of a change in international ownership. Again, as there is no longer time series, no seasonal adjustment on the relationship between shipping could be performed.

Although the overall correlation is reasonably good, individual points can deviate strongly. Therefore, it is not recommended to use these indicators as predictors of GDP or other headline economic statistics on their own. Differences between the official statistics and the shipping indicators will also arise because:

- all ship types were included, not just cargo ships
- not all imports to the UK come via the sea
- some shipping will be between UK ports, rather than international voyages
- AIS does not contain information on the value or type of goods being transported.

**Way Forward**

As the individual shipping ports have different import and export product profiles, the intention is to study the links between the port indicators and port product profiles at the individual port level. Products can be linked to specific industries, allowing a deeper analysis of the relationship between shipping and the UK economy. Furthermore, the known cargo specialization of certain berths, due to the installation of specific loading equipment, creates a unique opportunity to obtain sub-port versions of the indices and explore their links with sectors of the economy. Another direction assumption is that movements of bigger ships are more important to the economy than these small leisure boats. Therefore, if suitable ship register data becomes available, the motion patterns of different groups of ships will be investigated. Different groups of ships might have different links to economic indicators and combining developments in sub-groups might reflect development in underlying economic trends. Ports can also have changing patterns of activity. For example, it could be interesting to monitor changes in both the number and type of ships at each port, for both international and UK-to-UK journeys.

The timeliness and coverage of the shipping indicators have the potential to offer real-time insights into how the UK and global economy is evolving. This might also be the case for other countries where shipping is an important part of the economy. Expanding the study on a global scale might promote an understanding of how the global economy is evolving in close to real-time. It will also allow for investigation of network effects and studying of ship movement patterns between trade partners and how these may evolve over time.
Maritime indicators (UNCTAD/MarineTraffic)

Background

Port performance is a determining factor in the level of a country's international trade level. In order to determine container port performance for different countries, UNCTAD developed the Liner Shipping Connectivity Index (LSCI). This index comprises six components: The number of scheduled ship calls, the deployed annual capacity in TEU (Twenty-Foot-equivalent Units), the number of regular liner shipping services, the average size in TEU of the ships deployed by the scheduled service with the largest average vessel size and the number of other countries that are connected to the country through direct liner shipping services. In 2019 this indicator became more detailed by providing information on a port level.

While the LSCI to date is generated purely from the scheduled of the ships, not their actual movement, in future it can be complemented by AIS data on the actual movement. Also, to date, the LSCI is limited to container shipping, given their regular scheduled services. In future, it could be envisaged to cover other vessel and service types, building on AIS data.

In addition, on a country level more detailed information on port call and performance statistics became available. In this new comprehensive table that features port calls by country, the typical turnaround time as well as the average size and age of ships. The statistics are derived from automatic identification system (AIS) data in collaboration with MarineTraffic.

Data

Calculations on port visits are based on data provided by MarineTraffic. Aggregated figures are derived from the fusion of AIS data with port mapping intelligence by MarineTraffic, covering ships of 1000 GT and above and not including passenger ships. In total, based on AIS data for the world commercial fleet of ships of 1000 GT and above, there were 1,884,818 port calls recorded in 2018.

Methodology

To calculate the number of port calls, only arrivals are selected. Cases with less than 10 arrivals or 5 distinct vessels on a country level per commercial market as segmented are not included. Passenger ships and RO/RO ships are excluded from the time at port calculations.

Time is calculated as the median time. Due to statistical outliers, the average time vessels spend in port is longer for practically all countries and markets. Ships can spend a long time in a port, for example for repairs skewing the data.

The data comprises 8 markets (based on ship type): passenger ships, wet bulk, container ships, dry breakbulk, dry bulk, RO/RO (roll-on/roll-off), LPG, LNG. For these markets the following variables are computed:
- Number of arrivals
- Median time in port (days)
- The average age of vessels (each ship is counted as often as it is called in the country's ports)
- Average size (GT) of vessels: (each ship is counted as often as it is called in the country's ports)
- Average cargo carrying capacity (DWT) per vessel (each ship is counted as often as it is called in the country's ports)
- Average container carrying capacity (TEU) per container ship: (each ship is counted as often as it is called in the country’s ports)
- Maximum size (GT) of vessels
- Maximum cargo carrying capacity (DWT) of vessels
- Maximum container carrying capacity (TEU) of container ships.

Preliminary Results

The main finding was that shorter times spend in port is a positive indicator of a port’s efficiency and trade competitiveness. In 2018, the median time of ship spent in port during one port call port was 23.5 hours. In general, dry bulk carriers spent 2.05 days during a port call, almost three times the median time of a container ship. Countries with more port calls usually have shorter turnaround times. The first year of coverage is 2018, with updates scheduled every six months.

If ships are larger, other things being equal, turnaround time should be longer, as there will be more cargo to be loaded and unloaded. At the same time, ports that can accommodate larger ships will usually also be more modern and efficient. UNCTAD analysis shows that there is a negative correlation between the size of the largest ship that calls at a country’s port and the median time ships spend in port, while there is a slight positive correlation for most market segments between the average size of vessels and the time spent in port. In other words, being able to accommodate very large container ships is an indicator that a port is fast and efficient, while ports that receive large ships will on average also take slightly longer to load and unload the higher cargo volumes.

Way Forward
In future, it may be that the schedules are combined with AIS data, i.e. the LSCI could be improved by verifying the schedules with the actual movements. By combining AIS data with the schedules, the schedule reliability can be measured and reported, and the LSCI itself can be fine tuned.

**Official Maritime statistics: Port visits (Eurostat)**

**Background**

In 2016 Eurostat (the statistical office of the European Union) started the ESSnet to integrate big data in the regular production of official statistics. One of the projects investigated whether AIS data can be used to improve the quality and internal comparability of existing statistics and for new statistical products. Initially, the focus was on maritime statistics. Most maritime statistics involve information on goods transported, however, AIS does not (directly) contain information on goods. Therefore, the first step was to investigate the use of AIS for port visit statistics. This statistic covers visits to European seaports by vessels (gross tonnage >=100) to load or unload goods or passengers if their voyage was undertaken wholly or partly at sea. There are a number of advantages of using AIS to produce the port statistic: for some ports, reporting the data still results in administrative burden and some ports do no report at all. Furthermore, basing this statistic on AIS could result in an improvement of quality and internal comparability in a statistical system.

**Data**

AIS dataset was obtained from Dirkzwager, a commercial party. The dataset contained 6 months of AIS data (8 October 2015- 12 April 2016), originating from land-based stations only, covering Europe and some non-European countries. The data contained both dynamic messages (e.g., location and speed) and static information (containing information on identity). Note that only MMSI-number (Maritime Mobile Service Identity) is present in both types of messages. AIS data was compared to ports statistics Poland and the Netherlands for 1 day. The identifying variable for ships used here was the IMO-number.

**Methodology**

The first step was to construct a reference frame of ships, for two reasons. Maritime statistics only apply to maritime ships carrying goods, however other ships entering ports also emit AIS signals. Filtering data from other ships reduce the huge amount of data. The other reason is the need for the frame of ships to categorize ships. Maritime ships can be identified on the basis of their so-called IMO number (International Maritime Organization). This number is not included in the dynamic messages (containing information needed to determine the location of a ship), they only contain the MMSI number. These MMSI-numbers also originate from non-maritime ships that should not be included for maritime statistics. Furthermore, not all statistical offices (can) collect MMSI or IMO numbers, some collect call sign as a shipping identifier. Thus, the reference frame of maritime ships should include all three ship identifiers: MMSI, IMO and call sign. Still, AIS data does not provide sufficient information for a complete reference frame of maritime ships as maritime statistics require more specific information on the identity of vessels. For example, the type of vessel defined in AIS data is less detailed than required by maritime statistics based on Directive 2009/42/EC. Therefore, in order to get more detailed ship information, AIS data has to be linked to existing ship dictionaries which was not present.

The reference frame was constructed by linking all MMSI-IMO couples in the static messages. First, couples with invalid IMO or MMSI numbers were filtered om the basis on length for MMSI and for IMO the check digit. To generate port visits on the basis of AIS data, the reference frame was used to investigate which ships had been in the port on one day for Poland (winoujcie) and the Netherlands (Amsterdam). For MMSI’s in the reference frame, it was checked whether their location (latitude and longitude) were present in a certain area. Statistics Poland did this for one day for the external boundaries for the port of winoujcie. Statistics Netherlands did this for the port of Amsterdam using a simple bounding box. Ships that were identified were compared to ships from the port statistics.

**Preliminary Results**

All the ships from the maritime statistics data were all present in AIS data. This indicated that AIS data had complete coverage of maritime ships in the port. However, using the first rough reference frame of maritime ships resulted in many more visiting ships in the AIS data. Closer inspection showed glitches in AIS data elements that could cause deformation of all of the elements in a message, sometimes resulting in valid fields. For instance, by coincidence, the resulting MMSI could be technically valid, but incorrect. These errors could arise for every variable, also resulting in erroneous latitude and longitude sometimes, where the faulty locations could be quite far away from the actual location of a ship.
First, this showed that the reference frame of maritime ships had to be improved. As one MMSI could occur coupled to different IMO-numbers. This was done by selecting only correct pairs by filtering out the most frequent MMSI-IMO pair over a 6-month period for each MMSI. This improved reference frame resulted in a smaller number of ship visits. However, still, maritime ships were found in port that was not actually present in the maritime statistics.

Second, the algorithm counting ships had to be improved, as the simple algorithm counted ships that were in the port, not their arrival per se. Thus ships that arrived days before but stayed for multiple days, were also counted those other days. Furthermore, the noisy nature of AIS data resulted in some ships appearing in the port for one message, although they were not actually present. This noise also resulted in messages of ships in ports, shortly appearing outside the port and then appearing again. To deal with the noise in the data, resulting in ship’s location for one glitch that was far away from their actual location, a filtering method was applied on latitude and longitude. Using a 10-minute averaging filter would not solve the problem, as noise-points far from actual locations, still resulted in faraway locations. Therefore a median filter over 10 minutes was used. Then, instead of simply checking if a ship was in the port, the algorithm counted ships entering the port. By splitting up the ships’ route into locations outside the port (sea) and inside the port (port). The point where the location changed from sea to port was counted as an entry. https://github.com/mputs/WP4/tree/master/Portvisit

Way Forward

The resulting number of ship visits was still a bit higher than the number of visits reported by maritime statistics. However, there were valid reasons for why they should be counted for the port visits, but were missing from original statistics (e.g., administrative reasons leading to miscounting by port authorities). Note however, that AIS does not give information on the action ships undertake. For example, ships might also enter to perform bunkering. Also, the type of ship might also not provide sufficient information, as towing ships can carry goods. The latter, however does not occur often. All in all, it was concluded that AIS could be of use in producing these statistics, if statistics cause an administrate burden or if the information is simply not available. At the moment, Greece (as part of ESSnet Big Data II) is working on improving the AIS for port visits. Other statistical areas were also identified where AIS could improve statistics. For example on missing information on the next destination of departing ships (to and from traffic matrixes), an improved average distance or information on fluvio-maritime transport (transport by maritime ships in Inland Waterways or Inland Waterway ships in maritime waters).

Completing statistics on Inland waterways
(Statistics Netherlands as part of the ESSnet Big Data II: Tracking ships 2018-2020)

Background

For the Netherlands, statistics on the transport of goods via Inland Waterways (IWW) is based on information that has to be provided when inland waterway ships pass by a lock. At the locks information is gathered on the origin and destination of a ship, the goods the ship carries and some ship’s characteristics. However, not all shipping routes pass by a lock, resulting in missing information for these particular journeys. Furthermore, not all locks render complete information. Thus, information on certain routes of ships and the goods transported is incomplete. One of the projects from Eurostat’s “ESSnet: tracking ships” focuses on completing this missing information using AIS.

Data

Both the data used here, lock information and AIS, come from Rijkswaterstaat, the Dutch authority on the design, construction, management and maintenance of the main infrastructure facilities in the Netherlands. AIS data comes from all Dutch mainland receivers from Rijkswaterstaat. The data used in the preliminary study comes from July 2015 and comprised of all message types, with a size of 60 GB raw NMEA coded AIS messages. Note that for IWW transport a different type of AIS message has to be used: message types 5 and 8. The latter is specifically used by inland ships and contains static information on the ships.

Methodology

In the Dutch statistical process of IWW, the operationalization of a journey is a movement between two locations where loading or unloading takes place, or the location where a ship moves outside Dutch inland waterways. To complete information on missing journeys, information from AIS can be used. To link AIS data to the IWW statistical data, the same concept of a journey is applied for AIS data. The method comprises 3 steps: preprocessing, deriving journeys and linking AIS and traditional journey information.

Preprocessing
From the AIS data, dynamic and static messages are separated. For each ship, a file is constructed with all dynamic messages and one file with all static messages. The dynamic messages contain information such as the ship’s location and speed. The static messages contain information on the ship characteristics such as identity, size, and ship type. The dynamic information is filtered to contain messages where the ship’s speed is faster than 0.2 knots and latitudes and longitudes are in the Dutch range. Navigational status was not used, because these were not always filled in or not correctly filled in. In addition, messages are filtered that do not have valid MMSI (i.e. not having 9 digits). The file with static messages is deduplicated on the basis of ship type (as a ship might have different uses through time). Then, the single static message is added to the dynamic messages.

**Deriving journeys**

The operationalization of a journey is the movement of a ship between two locations where goods are (un)loaded or the point where a ship moves outside Dutch IWW. For this, each file containing information per ship is used. When the time interval between two successive data points is over an hour or the difference in location is larger than 200 meters, the first data point is considered to be the end of a journey and the succeeding data point the start of a new journey. Each location is linked to a register of terminal locations using the nearest node method. As the definition requires locations where ships (un)load information, the location register also has to contain locations where ships regularly stop for other activities such as waiting areas for locks or sleeping places. This enables differentiating locations for (un)loading conditions and other stopping reasons. This process requires a good register with terminals.

**Linking AIS and traditional journey information**

Journeys from AIS (lasting from tA1-tA2) were linked to “lock” journeys (lasting from tB1-tB2). The following criteria were used: the starting time of the “lock” journey had to be before the end time of the AIS journey and end time of “lock” journey had to be after beginning time of AIS journey: tA2 > tB1 and tB2 > tA1. Then, for the “lock” journeys the amount of overlap with the AIS journey is calculated using the following link index:

\[
\text{Link index} = \frac{\left(\frac{tA2-tB1}{tA2-tA1}\right)}{\left(\frac{tB2-tA1}{tB2-tB1}\right)}
\]

If the time matches, link index will be around a value of 1, otherwise the value will lower. Values with the highest link indices are matched. AIS journeys that are not linked to “lock” journeys are taken to be missing journeys from the lock data.

**Preliminary Results**

AIS data was useful in completing the information on missing journeys. Using AIS resulted in extra journeys that were missing from the lock data. However, the journey linking algorithm was not optimal. Both false positives and false negatives occurred. In the extra journeys from AIS data, the problem also lies in the definition of a journey. The start and end of a journey should be characterized by a loading or unloading action. However, ships can also stop to rest, wait for a lock or to get fuel. As mentioned, this requires an accurate database of stopping locations that are functionally defined. However, the problem is that some stopping locations can be both a waiting area for a lock and a loading dock. Besides, the location database is not always accurate and complete. The final problem is that AIS data does not contain information on the type and quantity of the goods loaded or unloaded. The type of ship and the type of terminal can provide some information on the nature of the type of goods.

**Way Forward**

At the moment, AIS data is implemented in the production process of IWW transport statistics. The linking algorithm, in which AIS and lock travel are linked is not optimal and will be improved. Also, the location database should be improved. In addition to the coverage in the data from locks being incomplete, it is not obligatory for certain ships to report goods transported. This means that in multiple cases information on goods transported is missing. One of the plans is to implement information from other ships to characterize the type of terminal. Employing clustering available shipping information, useful information might be obtained to estimate the type of goods ships has carried.

**Mapping fishery activities (EU JRC)**

**Background**
Information and understanding of fishing activities at sea are indispensable to fisheries science, public authorities, and policy-makers. The EU JRC performs research into fishing activities. Fisheries research in the EU relied heavily on effort, catch and fleet capacity data from the fleet register, the logbooks, the sales notes and the Vessel Monitoring System (VMS) established by the control regulation (Council Regulation (EC) No 1224/2009). VMS data provides detailed information on the vessel tracks at the high spatial-temporal resolution, while the logbooks include essential information on the gear used, species and volume of the catches. However, the problem is access to the VMS data and the relatively low temporal resolution (2 hours). Since AIS became compulsory for fishing vessels of more than 15 meters of length, this opened up a new source to monitor fishery activities. Not only is the source easier to attain, but it also has a higher spatial resolution. The drawback, however, is that spatial coverage is lower, only 100 nautical miles from the coastline (and for S-AIS the temporal resolution is lower again). To investigate the usefulness of AIS to study fishing activities, 2016 the JRC (Vespe, Gibin, Allesandrini, Natale, Mazzarella and Osio) published the first map of EU fishing activities based on AIS data in 2016.

Data

In Europe (EU Dir 2011/15/EU), fishing vessels down to 15 m in length are required to be fitted with AIS. To study fishing activities, terrestrial AIS historical data from September 2014–September 2015 was processed. After aggregating the data from multiple providers, spatial coverage is computed in order to give an estimate of the reliability of the results in different areas. The AIS data set contained fishing vessels only. Fishing vessels were identified by linking vessel identifiers (MMSI) from statics AIS messages to the EU fishing fleet register through call sign and name. In addition to this resulting list of EU fishing vessels, information on the primary and secondary gears from the fleet register is used to define specific fishing categories (e.g., trawlers, purse seiners). Here, only fishing activities by trawlers, the largest portion of the EU fishing vessels above 15 m of length, were analyzed. Once linked to the fleet register, the data set was anonymized.

Methodology

First, the data was cleaned. For this, the position and speed data cleaned and filtered to an interval of 5 min between consecutive observations (reducing 120 mln to 60 mln messages). Messages with zero-velocity messages relating to periods when the vessel is likely to be in a port were excluded. Resulting speed profiles were analyzed. This resulted in speed profiles with a bi-modal distribution, where corresponding to fishing behavior is extracted. This algorithm is based on the assumption that speed during fishing activities is lower than during steaming (Mazzarella et al., 2014). Note that, factors such as vessel size, area, and fishing gear result in specific mean and standard deviation values of the speed bi-modal distributions. For this reason, the identification of fishing behavior has to be implemented for each individual vessel (Natale et al., 2015). For the map, aggregate results were turned into density maps: the resulting points classified as fishing were aggregated into 1 km² cell.

Preliminary Results

This data was plotted, showing transit legs containing the high-speed points. The low-speed points, likely indicating fishing grounds, were usually in the middle of the sea at the far ranges of the track. By applying this approach for all fishing vessels in a specific area, it was possible to specify the fishing grounds and map high-intensity fishing areas. The first emerging pattern is that all the continental shelf area in the EU Mediterranean countries is almost all subject to high intensity of trawled gear fishing.

Way Forward

At the moment worldwide studies are being performed into the quality of coverage of AIS to study fishing behaviors. For example, for the Northeast Atlantic, it is clear that most vessels over 15 m long broadcast AIS using Class A devices mainly. This implies that a large network of terrestrial receivers enhances satellite coverage. However, several offshore regions beyond the coverage of terrestrial receivers have poorer reception (e.g. some areas in the North Sea, the Bay of Biscay, and areas east of Ireland and the United Kingdom). Vessels with AIS Class B devices generally limit their operations to coastal areas and are well covered by terrestrial receivers. Observation in detail can highlight substantial differences in patterns and intensities mainly due to uneven spatiotemporal coverage and inadequate gear information in the AIS dataset. This means that specific types of fishing vessels cannot be clearly distinguished. Note that in 2012 the European Data Protection Supervisor issued an opinion on the use of AIS and VMS data. The opinion states that as long as the data can be linked to identified or identifiable individuals it entails the treatment of personal information. Under such circumstances, the treatment of the data should follow a general principle of “limitation of purpose” and be confined to general law enforcement and objectives connected with the Common Fishery Policy.

Ships in distress (UN Global Pulse)

Background
In 2016, over 363,000 migrants and refugees are known to have arrived in Europe by sea (International Organization for Migration (IOM), 2017). To provide a better understanding of the context of search and rescue operations, UN Global Pulse performs research using new big data sources. Migrants and refugees are typically counted in two ways: (a) when they reach Europe and their arrival is recorded and administratively processed by the authorities, or (b) when they are reported dead or missing at sea. However, there is much less quantitative data available on the conditions of their journeys across the Mediterranean. AIS data and broadcast warnings have the potential to advance the understanding of migrants’ and refugees’ journeys in the Mediterranean, and to better explain the events that lead to migrants and refugees dying or going missing. This study investigated how different data sources can be combined into a quantitative rescue framework. Machine learning is then used to perform automated rescue detection based on vessel trajectory information.

Data

AIS data was obtained from different providers. Other big data sources are broadcast warnings (short, text-based radio alerts used to update ships on nearby activities, dangers, and emergencies) and Twitter. Other information on rescue operations is also used from both official (e.g., Italian Coast Guard and the UN Refugee Agency) and non-official (e.g., Rescue NGOs) sources.

Methodology

Information from the different sources on rescue sequences is combined into a so-called quantified rescue. This approach is used to unify narrative threads from a variety of sources, tying the observable physical traces of a rescue operation to qualitative sources of information on what happened and how events unfolded. This reduces “subjectivity” in the data. To produce a quantified rescue, incident descriptions were used to identify the ships involved in a rescue. Then, timestamps were used to link these descriptive statements or tweets to the AIS coordinates that are closest in time. This is not always possible as descriptive data can be vague or incomplete. Another problem was that for some AIS data the frequency was too low. As rescues involve sudden movements, the interval between two data points cannot be too long, otherwise, deviations in trajectories cannot be picked up. Also, the amount of available information varied substantially by the incident.

To produce a large-scale training dataset for machine learning, 77,372 points were manually geotagged, spanning four search and rescue ships over a period of 100 days. Points were labeled according to two estimated activity types—“Rescue” and “Non-rescue”—based on the shape of the trajectory and the corresponding rescue ship’s speed. Models of rescue behavior were trained to predict this binary outcome based on the characteristics of trajectory points, including speed, course over ground, latitude, longitude, day of week, hour of day, and month of year. Two approaches were taken:

1. First, classification algorithms were trained on the raw point dataset.
2. Second, classification algorithms were trained on clusters of points. Specifically, points were clustered prior to classification using the Cluster-Based Stops and Moves of Trajectories (CB-SMoT) algorithm (Palma et al. 2008). Density is defined using two parameters: eps, a distance parameter, and mintime, a time parameter. CB-SMoT uses these parameters to classify points into three categories: (1) Core points, from which the object travels less than the distance threshold eps in either direction within a period of length mintime; (2) Border points, falling within eps of a core point; and (3) Other points, which are neither border nor core. Together, core and border points form clusters that represent periods of slow motion. Although the algorithm is designed to find stop points, it was used here to break trajectories into segments with different motion profiles by allowing sequences of “other” points to form their own clusters.

Automatic characterization was applied using standard binary classification algorithms—AdaBoost, Support Vector Machines (SVM), and logistic regression—with 10-fold cross-validation.

Preliminary Results

A systematic analysis of quantified rescues can yield insights into exactly how many rescues ships are conducting, where and how they happen, and how long they take. In addition to revealing how multiple vessels may conduct a single rescue, quantified rescues can demonstrate how multiple migrant boats may be discovered in sequence by a single rescue vessel, forcing it to work continuously to bring people on board until it reaches full capacity. Thus, quantified rescues can help to systematically profile rescue activity patterns, and create data models of rescue operations.

With respect to automatic detection of rescue activity, features such as speed and course over ground appear to be relatively good predictors of whether a vessel is conducting a rescue operation or not, with most rescues associated with speeds of less than two nautical miles per hour. More generally, using the full set of input features, the top-performing models correctly classified over 96% of points, on a dataset for which approximately 75% of the points are labeled as non-rescues. There was a key limitation however: classifier performance was location-dependent. A ship’s location near the Libyan coast was too dominant in identifying rescue behavior, and performance declined when location features were removed from the data; the next steps include improving model performance when location features are not included. Finally, although predictive performance was lower on the clustered dataset, CB-SMoT appears to be a promising approach to detecting rescue operations because of two key features: its natural incorporation of both time and distance traveled, and its ability to work with sequences.
The information contained in quantified rescues could be used to measure the conditions under which rescues are most effective; to predict which rescues are likely to be associated with fatalities, or to produce input for operational predictive modeling. Once the model performs well in separating rescue from non-rescue points, AI can then be used to pre-process data, with predicted rescues acting as a building block to making larger-scale abstract inferences about the situation at sea.

Despite the promise of these new data sets, their biases pose an important challenge for analysis. While this research has shown that AIS datasets can be used to understand rescue operations, they are generally less useful for identifying the movements of migrant vessels. In many cases, the vessels carrying migrants and refugees may have their transmitters off, may transmit wrong information, or may be simply too precarious to be equipped with a transmitter in the first place. An additional challenge related to the automated detection of rescues is that features which may have been key predictors of rescue activity in previous years – such as a specific latitude and longitude – might no longer be relevant today. Similarly, while the identification of rescues could potentially be framed as an anomaly detection problem, rescue activities in the Central Mediterranean have become so concentrated, and so frequent, that they might now characterize normal behavior for the area.

Way Forward

There are a number of potential policy and operational applications for quantified rescues and the automatic classification of rescue patterns at scale. First, they can be used to study the geographic evolution of rescue operations. Second, they can be used to retrace potentially contentious rescue maneuvers. Third, they can inform the theory and practice of conducting rescue operations, helping to optimize maneuvers according to e.g., conditions at sea. Fourth, they can be used to facilitate coordination between rescue ships, and to ensure proper spatiotemporal coverage of rescue operations. Finally, they can be used to identify rescues that might otherwise go unreported, such as those conducted by commercial ships.

Ongoing work departs in two directions. First, trajectory data will be converted into images to test the performance of out-of-the-box deep learning image classification algorithms on these rescue “signatures”. This will eliminate the need for a two-stage clustering-and-classification approach since deep learning algorithms essentially implement and tune a multilevel learning strategy automatically. Architectures to be tested include Convolutional Neural Networks (CNNs), which are useful in recognizing patterns with varying location and scale; and Recurrent Neural Networks (RNNs).

Second, the process of trajectory labeling will be improved. When tagging trajectories by visual inspection, it is difficult to determine when a rescue starts and ends. A ship may spot a migrant vessel and move toward its position, however, ships can also be called on to assist migrant vessels that are hours away. Similarly, ships seem to move along the coast while patrolling for migrant vessels, creating the appearance that the ship is engaging in an abnormal rescue-like maneuver. Rescues and transfers can also be easily confused, given that they both occur at slow speeds. As a result, it can be hard to determine a ship’s intent from trajectory characteristics alone. Future work will leverage ground-truth rescue data (reported by MSF), which corresponds to the much narrower period of time when people pass from a rescued ship to a rescue ship.

By combining data from very different sources, new inferences can be drawn. This suggests the need for a real-time common database for keeping track of the date, time, vessel, location and number of passengers for all rescues in the region to assist with coordination and analysis of the situation.

NOx, SOx, and CO2 Calculation (Université du Havre)

Background

The maritime vessels which normally operate in international waters are subject to the regulations of the International Maritime Organization (IMO). Annex VII of the Marpol Convention (1997, 2010 and 2015) establishes emission control zones (ECAs) for NOx and SOx any deliberate emission of substances that deplete the layer of ozone. Since 2011, there are four ECAs that exist worldwide in the Baltic and North Sea for sulfur emissions (SECA); in North America and the Caribbean maritime area of the United States for emissions of sulfur and nitrogen oxides and particulate matter.

The authors aim to measure the impact of potential emissions of NOx, SOx and CO2 on the environment by analyzing AIS data. Initially, the first methodology uses three maritime indicators: speed, gross tonnage and year of construction. In addition, weather conditions (i.e., wind speed) at a specific time and type of engine are also integrated to refine calculation methodology.
Data

AIS data used in the study were sourced from the AIS-HUB network containing 1048576 geolocation data points with total of 1430 vessels. The data was taken for a period of 7 June 2017 to 19 June 2017.

Methodology

The authors identified three structural variables and one geographical variable which can be used to calculate emissions from maritime traffic. They are as follows:

- Gross tonnage (DWT): implying the need of energy for moving the vessels
- Age of vessel (Ag): recently constructed vessels have a more efficient propulsion system or equipped by “Dual Fuel”. It is categorized into three periods: before 1990, between 1990 and 2005 and after 2005
- Speed of vessel (SOG): which has an impact on fuel economy (i.e., "Slow Steaming" could reduce fuel consumption up to 50%)
- Geographical concentration of traffic

The variables are weighted based on their impact on the production of emissions (Ag=3, DWT=2, SOG=1). The formula of emission potential is as follows:

\[ ((Ag \times 3) \times (DWT \times 2) \times (SOG)) / Ut \]

With Ut representing a unit of time in observation

Preliminary Results

Containers ships are the top CO2 emitter in international shipping due to the use of fuel. Since 2009, the ships' operator has been reducing the speed of their fleet to reduce fuel consumption. However, the fall of fuel prices may make the operators resume the average speed and as a consequence would amplify pollution by vessels.

Nowcasting Trade Flows in Real-Time (IMF)

Background

According to the United Nations Conference on Trade and Development (UNCTAD), maritime shipping was the mode of transport for about 80 percent of trade volume and for about 70 percent of trade value in 2018. Modern technologies like the automatic identification system (AIS) for vessels enable to track these trade flows in real time. Consequently, data coming from the AIS have the potential to serve as a fast and granular indicator for trade and maritime activity which could help to detect turning points on the economic cycle. The study described in the IMF working paper uses official statistics from Malta as a benchmark to evaluate their indicators based on AIS data. It shows that trade in goods (not in services), trade volume (not value), gross trade (not re-exports), and trade by broad groups (rather than specific goods) can be measured with the AIS data.

Data
The study is based on port call data for two ports in Malta between January 2015 and December 2018, i.e., it focuses on vessels near these ports in this time period. MarineTraffic generates this more structured data set of port calls from AIS data. Using port calls reduces the size and complexity of the data while still keeping track on incoming and outgoing vessels. Furthermore, port call data tends to be more accurate than full voyage information due to better AIS-receiver coverage close to ports. Typically, the port call will include information about the port name, vessel identifier, gross tonnage, deadweight tonnage, and draught of the vessel, vessel type, a timestamp with arrival and departure times, and actual departure time from the last port and estimate arrival time to the next port. The port area is defined with bounding boxes.

Methodology

The methodology used on the study follows two steps:

1. Cargo ships are identified by a filter and static and voyage-related information for the identified ships is aggregated:
   
   The filter identifying the cargo ships follows three rules: 1) Bunkering tankers providing fuels to vessels located at seaports, 2) ships arriving but not departing, and 3) ships that stay in the port boundaries only for a short time or for too long are omitted.

2. High-frequency indicators are derived: On a weekly basis, the cargo number indicator that counts the number of incoming ships and the cargo load indicator based on information on the ship's deadweight tonnage and the reported draught are calculated.

To verify the indicators, they were compared with the official Maltese statistics.

Preliminary results

The study finds that port calls derived from AIS data can be used to nowcast trade flows. The two assessed indicators, cargo number and cargo load, are evaluated with official trade statistics for Malta. While the cargo number indicator exceeds the number of ships relative to the data from official port statistics, the cargo load indicator seems to reflect the trade volume for Malta. Since international trade presents a crucial share of the GDP in several countries, policy makers could benefit from the real-time information about international trade derived from these indicators. The approach will be most valuable for small island countries with open economies. Furthermore, the indicator predict trade in goods, not services, and do only provide on total traded goods and not on detailed types, except oil and gas.

Way forward

Even though the study is conducted for Malta, the approach can easily be transferred to other countries. Especially in countries where shipping is the most common mode of transport, the indicators may serve as valuable real-time information about the trade volume. In future studies, this hypothesis should be assessed by testing the robustness of the indicators with a longer time period and other countries. Furthermore, the cargo load indicators may be improved by having a better understanding of the relationship between actual cargo size and AIS-reported draught. Finally, the filter algorithm to identify ships that actually load and unload cargo can be improved by using the features speed and course over ground in the AIS data.

Experimental Statistics of Daily number of vessel (Statistics Denmark)

Overview of Methodology

Statistics Denmark has published index of daily number of vessels visiting number Danish ports using Danish AIS-data. It is accessible from https://www.statistikbanken.dk/aisdag. The data processing steps consists of the following

1. Data reduction
2. Selection of arrival and departure observations
3. Delimitation of ports
4. Linking arrival and departure observations with ports and creating statistics.

Data reduction
The first step is data reduction. The sole purpose is to reduce the amount of data that has to be processed, and it consists of three parts: geographical reduction, reduction of observation frequency and reduction of analysed vessel types.

AIS data contains all the information from AIS that the Danish receivers have registered. That includes multiple observations that do not relate directly to Denmark but rather Germany, Norway or Sweden. The method used for geographical delimitation is simple as data is delimited to be within a square that covers all of Denmark and, consequently, the southern part of Sweden also. The Swedish ports will be excluded later on in the process. Here it is possible to reduce the observations to ones that are within the Danish waters.

The ships transmit signals at an interval, which is determined by speed and type of activity as well as the type of transmitter/transponder. The most frequent signals are received at intervals of a few seconds. However, since the objective is to examine port calls that level of data frequency is not necessary. Consequently, data is reduced to the first observation per minute.

The activity in the ports can cover many types of activities. In order to support existing port statistics that are centred on cargo handling in the ports, data is reduced to vessel types that are only used for cargo transportation: freight and container vessels.

**Determination of arrival and departure observations**

The next step in the process is to identify the ports calls with arrivals and departures. Each ship leaves behind a number of position data, and the objective is to identify the two observations, one that represents the arrival at a port and another that represents the departure from the port.

Fundamentally, the process is simple: The data set is sorted by identification of ship and time. Subsequently, all observations where the ship shifts from movement to a standstill are marked (where the navigational status shifts from “under way” to “moored” and the ship goes from moving (more than 1 knot) to (almost) a standstill). These are potential arrivals. The same approach is used for potential departures where the ship is moving and the navigational status shifts from “moored” to “under way”. All of the potential arrivals and departures are matched. For the majority it correlates well, but there are arrivals without departures and vice versa. Possible explanations are:

- **Lengthy stay in a port:** The average stay in a port is approximately 12 hours, but if the stay lasts for a number of days it can result in one of the matching observations missing, in the beginning or the end of the observed period. This is partly made up for by using data that falls outside the period, e.g. by not preparing the statistics for last month until 5-6 days after the cut-off-date. Long-term stays in a port is rarely connected with transportation of cargo, but rather a need for repair work or something else, and they are probably not important in relation to the objective of the statistics.
- **Turned off transponder:** It is not (normally) illegal to turn the transponder off, and a possible scenario is that the transponder is not turned on until the ship reaches the port or that is turned off after arrival at the port, and the crew forgets to turn it on again after departure.
- **Data disruption:** Data disruption can occur when data from a single coastal AIS receiver is not registered, data is not streamed from the Danish Maritime Authority or in cases where data is not collected and stored by Statistics Denmark. The latter was mainly an issue in the initial phase and disruption has been reduced significantly over time.

**Delimination of ports**
The third step is to exclude the observations that are not actually port calls. At this point, there is still a large number of arrivals/departures not close to a port. Furthermore, data should only include Danish ports. Observations from other ports, especially Swedish, are still a part of the data basis, including the largest cargo port in the Nordic countries, Göteborg. If data is reduced to Danish waters to begin with, this will no longer be an issue.

Once this third step has been carried out with, no less than, a couple years of data (statistics for a single month can be prepared without reducing the port data again. This reduces the monthly production time significantly and new data can be included in the reduced data after the production of the statistics. The process of this step consists of three parts:

• Gather observations (arrivals are used) in groups (clusters) based on the individual distance between the observations and the number of observations in close proximity.
• Make polygons that cover all the observations in the same cluster.
• Connect the individual polygons to actual ports.

The result of these three steps is a spatial look-up table that consists of polygons and the matching port. A port may be connected to a number of polygons whereas a polygon can only be connected to one port. If an arrival is located within a given polygon, we can then tell which port the port call belongs to. For larger ports, each polygon will typically represent a particular quai.

The first step is to calculate all the individual distances between the arrival points, and subsequently, the arrivals are gathered in groups. If the distance between two observations (irrespective of time) is less than e.g. 50 metres, they are connected. If a third observation is less than 50 metres from one of the first observations, it also becomes part of the cluster. All groups with more than e.g. 5 observations then become final groups or clusters. The number of observations and distance between them are parameters that can be adjusted. The less observations used, the larger the distance should be and the smaller the number of observations in a group should be. With three years of data, 70 metres and 5 observations are used. After the first step, all of the observations have been connected to a group (cluster). All the observations that do not meet the criteria are given cluster number -1 and are considered as invalid port calls.

In the second step, polygons that encircle the individual clusters are made, so that all observations in the same cluster are within or on the edge of the polygon. By looking at the location of these clusters, you will, without further processing, get a good idea of where ports are located. In countries where unofficial ports are established, these ports can also be identified.

In the third step, the individual polygons are connected to a port. This is done in an iterative process in which the basis is a centroid for the port (or something similar – most often, the port coordinate that is on the UnLocode list is used, but it can also be found by a simple search on e.g. Google Maps). The iterative process ensures that no clusters are connected to more than one port and that the polygon is connected to the port that is closest to the polygon. Separation of ports that are close to each other can result in a little extra manual processing, i.e. adjusting the reference points so the individual clusters are connected to the right port. Iteration is done by making a circle that gradually increases around the port’s reference point. If a circle overlaps a cluster polygon, this cluster is connected to the port in question, and it cannot, subsequently, be connected to other ports. The maximal distance from the port’s reference point to the clusters has to be defined by actual data. In the Danish data, the limit is based on a defined number of clusters for ships that are lying in a roadstead off the coast of Skagen. They are not considered to be lying in port and the limit is set so that these clusters are not included. Since the basis is Danish ports, other ports, primarily Swedish ones, are excluded in this process.

Finally, we end up with a look-up table, in which further information about the individual ports can be found, e.g. municipality, Unlocode, name, business identification number and coastal zone.

Linking arrival and departures with ports

The last step in the entire process consists in matching the port calls (referred to as arrivals and matching departures in step 2) with the polygons that define (part of) the ports. The result is a data set that contains information about the port call (primarily time, identification of the ship and the ship’s name) and the port in question (primarily the port’s name, code and region). Data can be supplemented with more detailed information about the ships from ship registers, e.g. size, flag, more precise vessel type and owner/operator.

From this point on, the process for production of statistics is the same as usual. Possible subclassification, indexing and seasonal adjustment and tabulation.