Analysing port and shipping operations using big data

Christopher Bonham, Alex Noyvirt, Ioannis Tsalamanis and Sonia Williams

Data Science Campus, ONS, Christopher.bonham@ons.gov.uk
18 June 2018

Data Science Campus

Office for National Statistics

Executive summary

The maritime freight industry is of critical importance to the economic output of the UK, with almost half a billion tonnes of freight being handled by UK ports in 2016. The Freight Transportation Association estimate that delays on both side of the Channel cost the UK logistics industry £750,000 a day[1]. As the demands upon shipping freight are likely to increase in the future, a more in-depth understanding of the UK maritime shipping industry becomes increasingly more important.

This report outlines the work undertaken by the Data Science Campus to explore the operation, use and relationships between ports in the UK at a macro level and the behaviour and operational characteristics of ships at a micro level, specifically:

- national and international relationships
- traffic at ports and related factors
- inbound delays
- capacity utilisation

Two sources of data are utilised:

- Automatic Identification System (AIS). AIS data records the position, speed, heading, bearing and rate of turn for each ship, at frequent time intervals throughout its voyage
- Consolidated European Reporting System (CERS). CERS data is collected at a higher level and records details such as destination port and expected time of arrival for the voyage of each ship

A means of storing, decoding and processing AIS data is proposed. A means by which AIS and CERS data can be merged is presented, allowing a more comprehensive analysis to be undertaken when compared with exploring each dataset in isolation. Exploratory analysis of both datasets uncovers several insights for ships using the largest UK ports and Felixstowe in particular. These insights include:

- port traffic and utilisation
- shipping movements
- port network analysis
- movement of hazardous materials
- delays at port

A novel unsupervised machine learning approach using K-means clustering is applied to AIS data aggregated over a time-based window. This is used to classify the behaviour of a ship into one of six unique groups at every point throughout its voyage. These classifications give a more meaningful and interpretable representation of ship behaviour and intention over time when compared with raw positional AIS data.

This classification along with a series of additional non-AIS related features are used to explore the feasibility of using supervised machine learning techniques to predict the likelihood that a ship will be delayed arriving at port. Random Forests, AdaBoost, Gradient Boosting and XGBoost algorithms are applied to shipping data taken from in and around the port of Felixstowe. Results are promising with the XGBoost algorithm being able to correctly identify a ship delay in nearly 70% of test cases.

These initial results suggest that additional focus should be placed on further development of both the classification and delays models. A means by which these predictions can be used to explore, simulate and optimise the operational efficiency ports throughout the UK is also discussed. The report concludes by discussing how the tools and technique used in the project may be applied to a broader set of applications lying outside of the maritime field.

[1] https://fta.co.uk/press-releases/20150724-port-delays-cost-freight-industry-750000-a-day-says-fta

Contents

1. Introduction 4
2. Background Research 5
3. Data Exploration 7
   3.1. Consolidated European Reporting System (CERS) 7
   3.2. CERS Insights 8
   3.3. Automatic Identification System (AIS) 17
   3.4. Processing AIS data 19
   3.5. Decoding AIS data 20
   3.6. Visualising AIS data 22
4. Classifying AIS data 26
   4.1. Voyage Classification 34
5. Predicting delays 36
6. Future work 42
7. Potential applications outside the maritime industry 45
8. Summary 46

All relevant code relating to this project is available in the Github repository.

1. Introduction

The maritime freight industry is of critical importance to the economic output of the UK. In 2016, 484 million tonnes of freight were handled by UK ports with 303 million tonnes being imported and 181 tonnes exported (UK Port Freight Statistics 2016, Department for Transport). Of the 120 commercial ports in the UK, 51 are classified as being "major", handling over one million tonnes annually and almost 98% of the total imported and exported freight. The European Union receives 66% of all UK outbound traffic with 16% being transported to the Asian continent, of which China accounted for 6%. Liquid bulk such as Liquified Natural Gas (LNG), crude oil and other oil-based products constituted 40% of all handled freight with dry bulk such as coal, ores and agricultural products making up another 20%. In 2016, 10.2 million TEUs[2] of container-based traffic passed through UK major ports.

It is therefore unsurprising that in recent years the amount of work that has focused on this area has increased dramatically. This report adds to this knowledge base by detailing the work undertaken by the Data Science Campus (Campus) at the Office for National Statistics (ONS). The DSC explored the operation, utilisation and relationships between ports in the UK at a macro level and the behaviour and operational characteristics of ships at a micro level, specifically:

- national and international relationships
- traffic at ports and related factors
- inbound delays
- capacity utilisation

This was done by understanding and analysing major UK port operation and utilisation using available ship geolocated big data and port itinerary reports provided by the Maritime and Coastguard Agency (MCA). This report begins by reviewing the latest research in the application of AIS data within these research areas. Section three explores the data sources used through the project, specifically CERS and AIS. Data issues encountered during the project are discussed and addressed with high-level actionable insights being drawn from both data sources. The report continues in section four by proposing an unsupervised learning approach to classify the behaviour of a ship into one of six segments, allowing the user to explore the behaviour of a ship at a more meaningful and insightful level. Section five presents a supervised machine learning technique that predicts the likelihood that a ship will be delayed arriving at its destination, these predictions can be used to predict port loading at a point in time and can support subsequent operational port planning. Model performance is explored, the most significant model features identified and their effect upon the likelihood of delays outlined. Weaknesses within the model are then discussed and areas of development are suggested. The report concludes by discussing potential areas of future work and exploring areas outside of the maritime industry where the work and findings discussed in this report may be applied.

[2] The twenty-foot equivalent unit (TEU) is a unit of cargo capacity used to describe the capacity of container ships. It is based on the volume of a standard sized individual 20-foot-long container that can be easily transferred between different modes of transportation, such as ships, trains and trucks.

2. Background research

This work is primarily using the AIS and CERS datasets provided by the Maritime and Coastguard Agency (MCA) for this study. The latest version of the CERS reporting system and dataset are used to fulfil MCA's reporting obligations under European legislation but, however this rich source of data has not been widely used to support additional research. Previously, AIS has been primarily used as an automatic tracking system, widely adopted to identify and locate vessels by electronically exchanging data with other nearby ships. In recent years, with the increase in the affordability of on-board data acquisition, storage and processing infrastructure and the development of modern distributed systems, AIS data has been used as a valid source of important information about vessel movement around the world. As Cabrera and others (2015) describes, AIS has often been used in the industry for numerous different types of applications like real-time statistics on ship traffic and congestion, operational management at ports, sustainable solutions on goods transport, route optimisation and many more. More specifically, there is a lot of work around the use of statistical methodologies on large numbers of trip trajectories to obtain motion patterns and route definitions around the globe. Real-time and historical AIS data can be used to forecast trajectories based on historical routes and allow for anomaly detection, collision prediction and route planning.

Anomaly detection

Real-time anomaly detection can identify potential security and navigation hazards and therefore is a useful feature not only for an on-board intelligent navigation system like AIS but also for the port authorities. It is based on creating motion patterns from historical data and using them to identify cases that deviate significantly. The normal ship motion is usually predictable as it follows a pattern, but the irregular movement characteristics of a ship are less predictable and a bigger challenge to identify. These vessels increase the risk of accidents or collisions in busy areas like ports and traffic lanes.

The anomalies in ship behaviour can be grouped in three main categories: position, speed and time. Different algorithms can detect different types of anomalies and Tu and others (2016) categorises the anomaly detection algorithms in two categories based on the learning characteristics of the models: geographical model-based and parametrical model-based methods. Geographical model-based methods are area specific models that are trained on local traffic data and are superimposed on a geographical map of the locale to detect anomalies. The following are examples of geographical model-based methods:

- Normalcy box described by Rhodes and others (2005)
- fuzzy ARTMAP described by Bomberger and others (2006)
- Holst model explained by Holst and Ekman (2003) and Laxhammar (2008)
- potential field method mentioned in Osekowska and others (2013)

Parametrical model-based methods are based on the development of parametric models of normalcy that are independent of training region maps. Some examples are:

- Trajectory Cluster Modelling (TCM) applied by Kraiman and others (2002)
- Gaussian Processes (GP) explained by Rasmussen (2006)
- Bayesian Networks (BN) used by Johansson and Falkman (2007)
- Support Vector Machines (SVM) applied in Handayani and others (2013)

Route estimation

Route estimation involves the development of models that can capture the motion characteristics of a moving vessel and accurately estimate the position and path of the vessel from that model. This information can be then used as an indicator of possible delays in the ship's arrival times or predictor variable in prediction models to forecast actual arrival delays. In general, the methods used to define the trajectory of a ship can be categorised in three main classes: physical model-based methods, learning model-based methods and hybrid model-based methods. In the physical model-based methods the motion characteristics of the vessels are calculated by using physical laws and mathematical equations that represent all possible factors that can influence the movement of the vessel. The

curvilinear model described by Best and Norton (1997) is a common general motion model that covers linear, circular and parabolic motion. The ship model proposed by Pershitz (1973) and Li and Jilkov (2003) is a dynamic model that considers the physical characteristics of the vessel and can describe and predict its motion.

In the learning based-model methods, the ship's motion is modelled by a learning model that is trained using historical data, in this case AIS historical location points and movement characteristics. The ship's manoeuvring system is being treated as an entire system and the model is being trained to mimic the system's function using the historical data. Neural networks presented in Haykin (2004), can fit complex functions and perform regression, making them the most common such models. They have been studied extensively throughout the years, they offer a stable and good performance, but their training process can take significant periods of time. Gaussian processes, as mentioned before for anomaly detection, are also very powerful on predicting the trajectory of a vessel. Extended Kalman filtering, as described by Hamilton (1994) and Grewal (2011), is a recursive estimator that consists of the prediction and update phases. Finally, Minor Principal Component Analysis has proven to be an accurate route estimation algorithm, as described by Bartelmaos and others (2005) and Peng and Yi (2006). It is a similar method to the Principal Component Analysis (PCA), simple to implement but might have limited ability to model nonlinear behaviour.

The hybrid model-based methods for estimating the trajectory of a vessel are combinations of physical and learning model-based methods to achieve better performance. An example of this is the combination of a curvilinear model to describe the common ship movement patterns and used as the motion model in the extended Kalman filtering, as described in Tu and others (2016). Another example is the combination of two different learning algorithms to achieve even more accurate estimation of the route, one to learn the characteristics of the ship's movement and the other to optimise the overall model performance. Such examples can include the combination of least square support vector machine (LS-SVM) and particle swarm optimisation (PSO) described by Zhou and Shi (2010), the combination of Kalman filtering and neural networks described by Guo and others (2009) and Stateczny and others (2011) and the combination of neural networks and genetic optimisation described by Khan and others 2005.

Path planning

In cases where high risk of collision is detected or alternative routes need to be found by the ship navigators, AIS can be used to provide the necessary information. Path planning is the process of finding a new safer route with the minimum cost with respect to time, distance, changes to the route and delays. In the past experienced navigators did this, but nowadays intelligent path-planning algorithms can take into consideration many factors and provide optimal alternative routes, as described by Cummings and others (2010). There are several path planning methods in the literature, like the shortest graph path method, evolutionary algorithm method and evolutionary set method, as described by Hornauer and others (2015), Lazarowska (2014) and Szlapczynska (2013).

The work around vessel behavioural segmentation presented later in this report is directly related to anomaly detection and route estimation, and might prove useful in enhancing the performance of some of the techniques described in this section. By detecting unexpected changepoints in the ship's behaviour segmentation, one can identify anomalies in the position, speed or time characteristics of the ship's voyage and feed these to the anomaly detection algorithms. Also, by analysing the historical behavioural segmentations of a specific ship or ships that travel through popular shipping lanes, new features can be engineered and used to estimate the future path of a ship. Finally, the study around delays prediction based on the motion behaviour of the vessels and other external parameters like weather can be linked to path planning and provide a more accurate target field for the planning algorithms. By accurately defining arrival delays and managing to detect them in a timely manner and quantify them, more efficient route planning might be achieved and delays avoided.

3. Data exploration

The data used in this project was provided by the Maritime and Coastguard Agency (MCA) who authorised access to the CERS platform and provided an extract of AIS data covering UK waters for land-based AIS transmitters. These data sources are discussed in the following subsections.

3.1. Consolidated European Reporting System (CERS)

The Consolidated European Reporting System (CERS) was originally created in 2006 to ensure the UK met its reporting obligations under European legislation. It is used by masters, shipping agents and port authorities to provide mandatory reportable information when a vessel arrives at a port in the UK. It captures ship arrival and departure notifications, dangerous or polluting goods notifications and notifications of port waste and bulk carrier infringements, for all the ports within UK waters. The information stored within CERS is forwarded onto SafeSeaNet (SSN), the central European data collection system in accordance with the EU Vessel Traffic Monitoring and Information System Directive (2002/59/EC)

A CERS report must be made at least 24 hours in advance of arrival or departure by the following:

- all ships of 300 gross tonnage and above
- all recreational craft of 45 metres length and over
- all ships regardless of size, when carrying dangerous or polluting goods, either departing from or bound to a UK port

The CERS system is not a dataset, but rather a tool that can be used to create records of voyages using a Windows-based user interface. Once complete and verified, records are passed in XML format to SSN. The MCA has more information relating to the CERS system.

The user interface consists of the following three sections.

- Port level information (one row per port), fields include:
    - port identifier and name
    - port address
    - port authority
    - port size
    - number of voyages

- Vessel level information (one row per ship), fields include:
    - Maritime Mobile Service Identity (MMSI), the unique ship identifier
    - name
    - callsign
    - gross tonnage
    - certification details

- Voyage level information (one row per ship, per voyage), fields include:
    - Maritime Mobile Service Identity (MMSI), the unique ship identifier
    - current, previous and next port of call
    - actual, estimated time of arrival and departure at current port
    - inbound and outbound hazardous material flags

All historical records can be downloaded for offline processing. Additional information including detailed waste and hazmat manifests was extracted (with permission) by creating an automated process to directly retrieve data from the relevant dialog screens within CERS.

3.2. CERS exploratory analysis

An extract of CERS data covering the 2017 calendar year was taken. This contained information relating to 120,000 voyages into and out of 172 UK ports. A high-level analysis of this data gives several headline insights relating to the operation, utilisation and relationships between ports in the UK at a macro level.

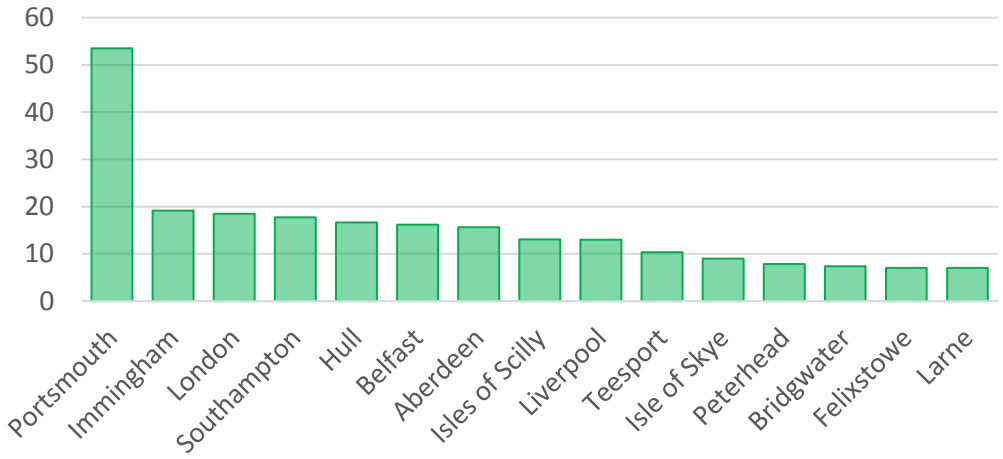


Figure 1: Average number of visits (call to port) per day
Orkney Islands (530), Gills Bay (209.4) and Penzance (207) omitted for clarity

Port loading

Port loading can be explored by plotting the average number of visits per day (Figure 1) and the average vessel size per visit (Figure 2). Here it can be seen that there are a handful of ports that receive very large volumes of visits per day. The Orkney and Gills Bay ports are major links in the oil and gas network and are also linked by frequently running ferries. The large number of daily visits to both Portsmouth and Southampton ports reflect their status amongst the largest passenger ports in the UK; this is further reflected in the relatively low average vessel size per visit. Surprisingly Felixstowe, the largest freight container port in the UK, has a comparatively low number of daily visits.

Turning to the average vessel size per visit (Figure 2). The first fours ports (Shetland, Hound Point, Tetney and Orkney) are all ports that predominantly serve the gas and petrochemical industries and are therefore most likely to be visited by the very large super tankers. Of the non-petrol chemical related ports, Felixstowe receives the largest ships by gross tonnage.

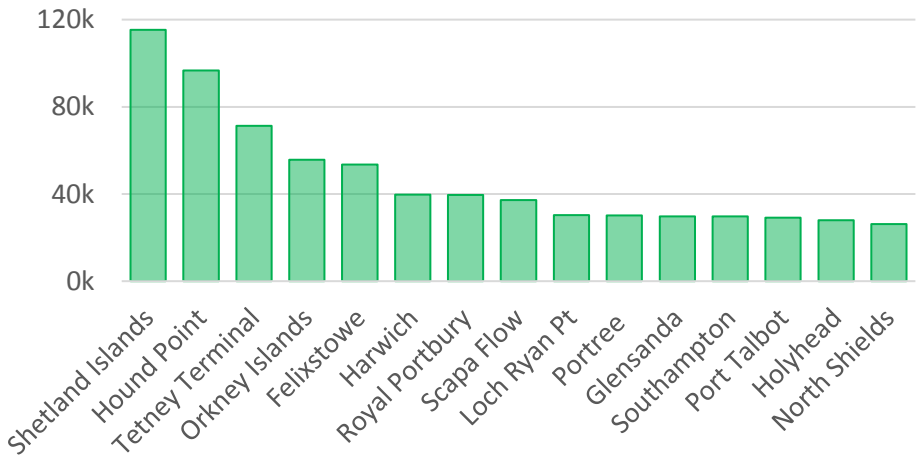


Figure 2: Average vessel size per visit (tonnes)

Focusing on Felixstowe[3], arrivals of ships into port broken down by month, day and time of the day are shown (Figure 3 to Figure 5). The figures are expressed as indices relative to the expected average for that time window. For instance, a daily value of 1.2 indicates arrivals that are 20% higher than the expected daily average. The indices for monthly arrivals are very small and caution must be exercised; however, they suggest there may be a small seasonal trend with fewer than expected ships docking at Felixstowe over the winter months (September to February) and larger volumes over the summer months (May to August). Turning to the daily breakdown, arrivals into port are lower over the weekend and higher during the working week. One exception to this is Monday where arrivals are almost 20% lower than the expected daily average. A stronger message can be seen when arrivals are broken down by time of the day; here Felixstowe has at least 20% fewer than average arrivals between early morning and lunchtime hours (0400 to 1300), with the quietest period generally being between 0800-0900 where arrivals are over 40% less than expected. Afternoons and evenings (1300-2200) are generally much busier with arrivals being up to 30% more than average.

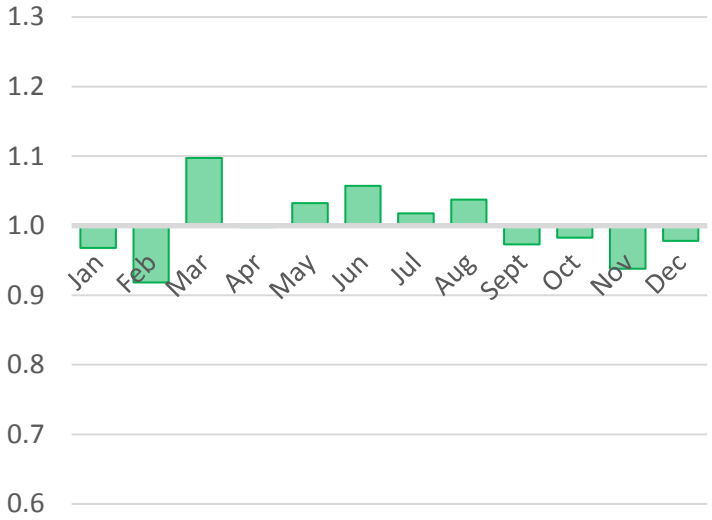


Figure 3: Monthly arrivals at Felixstowe
Figures are expressed as an index relative to the monthly average

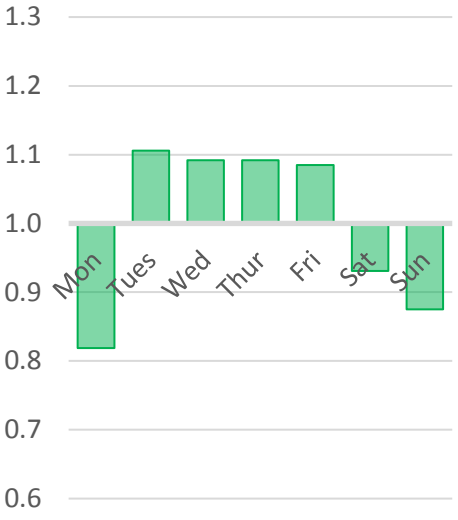


Figure 4: Daily arrivals at Felixstowe
Figures are expressed as an index relative to the daily average

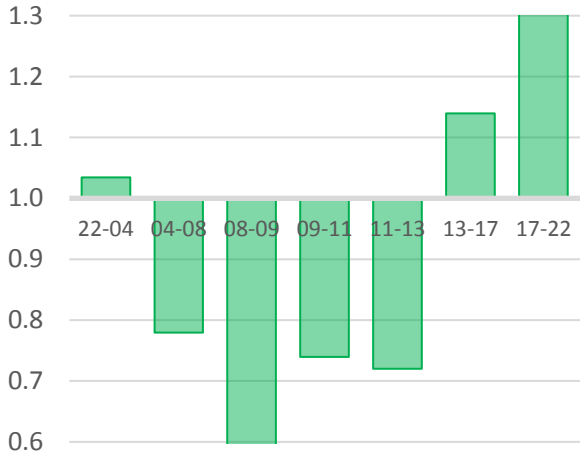


Figure 5: Arrivals at Felixstowe throughout the day
Figures are expressed as an index relative to the average for that time window

[3] The charts may be recreated for any port within CERS, however this section focuses on the port of Felixstowe reflecting its status as the predominant container port in the UK

Shipping movements and port links

The relationship between shipping movements for the ports of Belfast, Felixstowe and Milford Haven are shown below (Figure 6 to

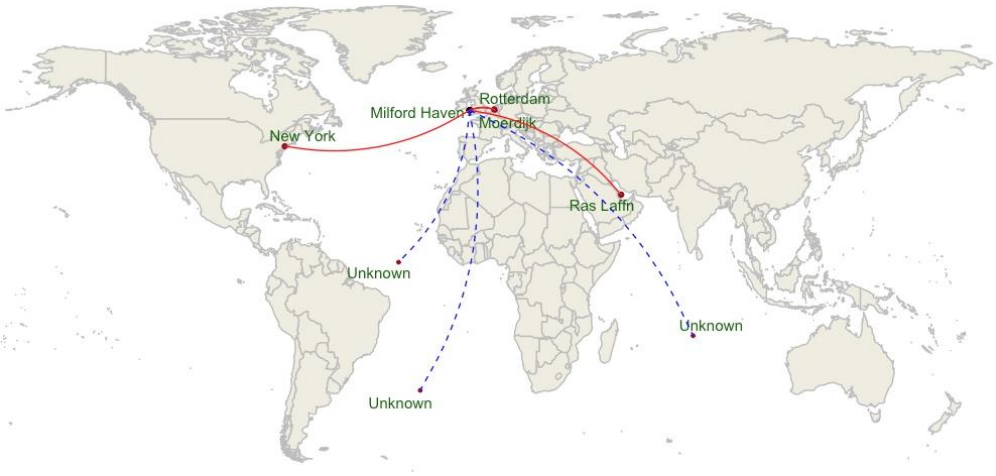
Figure 8). These show common port links for outbound journeys (as a percentage of total voyages). All three ports serve very different geographic regions. The ships leaving the port of Belfast predominantly sail to destinations within UK waters, with 60% of ships sailing to the ports of Loch Ryan, Birkenhead and Heysham. Ships leaving the port of Felixstowe generally travel to ports within the continental mainland with nearly 70% terminating at the ports of Rotterdam, Antwerp, Hamburg, Bremerhaven and Amsterdam. Milford Haven almost exclusively serves international destinations, with 88% of ships sailing to unspecified international ports, New York and Ras Laffan in Qatar.

Loch Ryan	35%
Birkenhead	15%
Heysham	10%
Unknow Int.	6%
Greenock	2%

Figure 6: Port of Belfast. National port links
(percentages give proportion of voyages from Belfast terminating at each port)

Rotterdam	51%
Antwerp	10%
Hamburg	8%
Bremerhaven	5%
Amsterdam	4%

Figure 7: Port of Felixstowe. Mainland continental port links
(percentages give proportion of voyages from Felixstowe terminating at each port)



Unknown Int[4].	86%
Rotterdam	4%
New York	1%
Ras Laffan	1%
Moerdijk	1%

Figure 8: Port of Milford Haven. International port links
(percentages give proportion of voyages from Milford Haven terminating at each port)

Further insight relating to the operational links between ports may be explored by applying network analysis.

Network analysis

Applying network analysis to the outbound and inbound voyages at ports allows for visualisation of the most important routes for Great Britain. Figure 9 shows the voyages from Great Britain to countries within 3,500km, highlighting the Netherlands, Belgium and Germany as the most important neighbours. Figure 10 shows the voyages to Great Britain from countries within 3,500km, also highlighting the Netherlands, Belgium and Germany as the most important neighbours plus Spain. For all inbound and outbound voyages, 37% are between ports within Great Britain, with 19% between Great Britain and the Netherlands, 8% between Belgium and Great Britain and 7% between Germany and Great Britain.

[4] Unknown Int. is a catch all destination classification that relates to all unknown or unspecified international ports

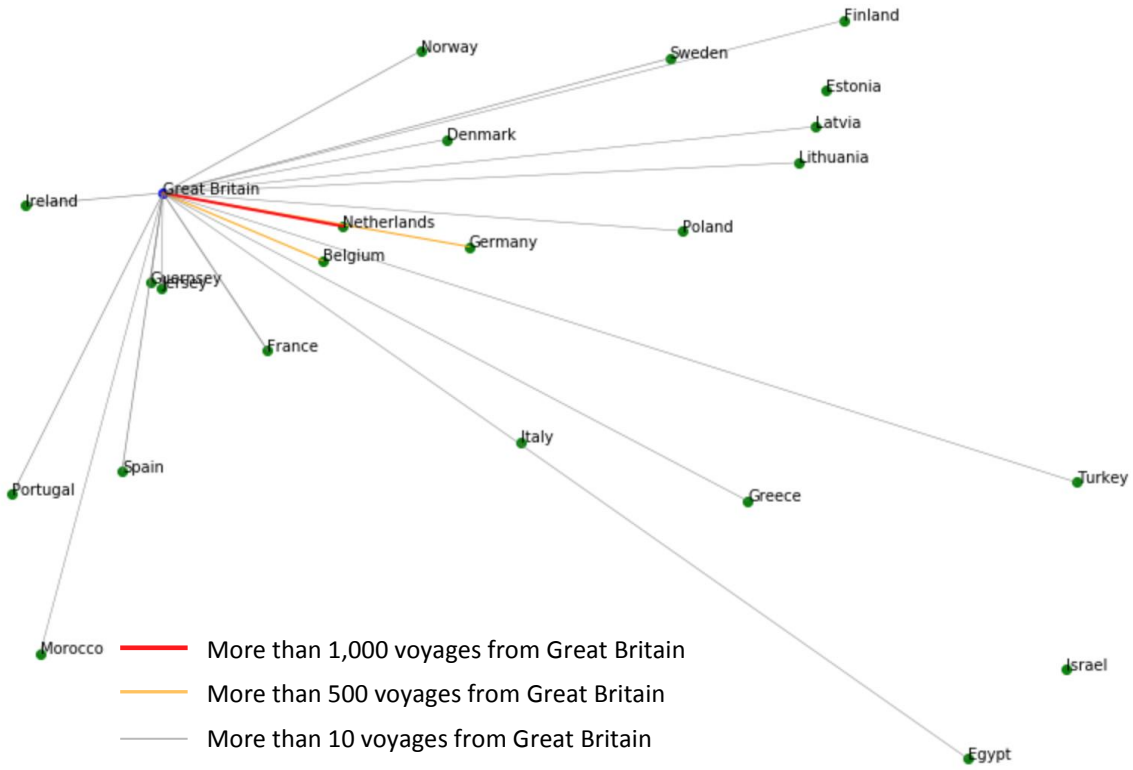


Figure 9: Voyages from Great Britain to countries within 3,500km

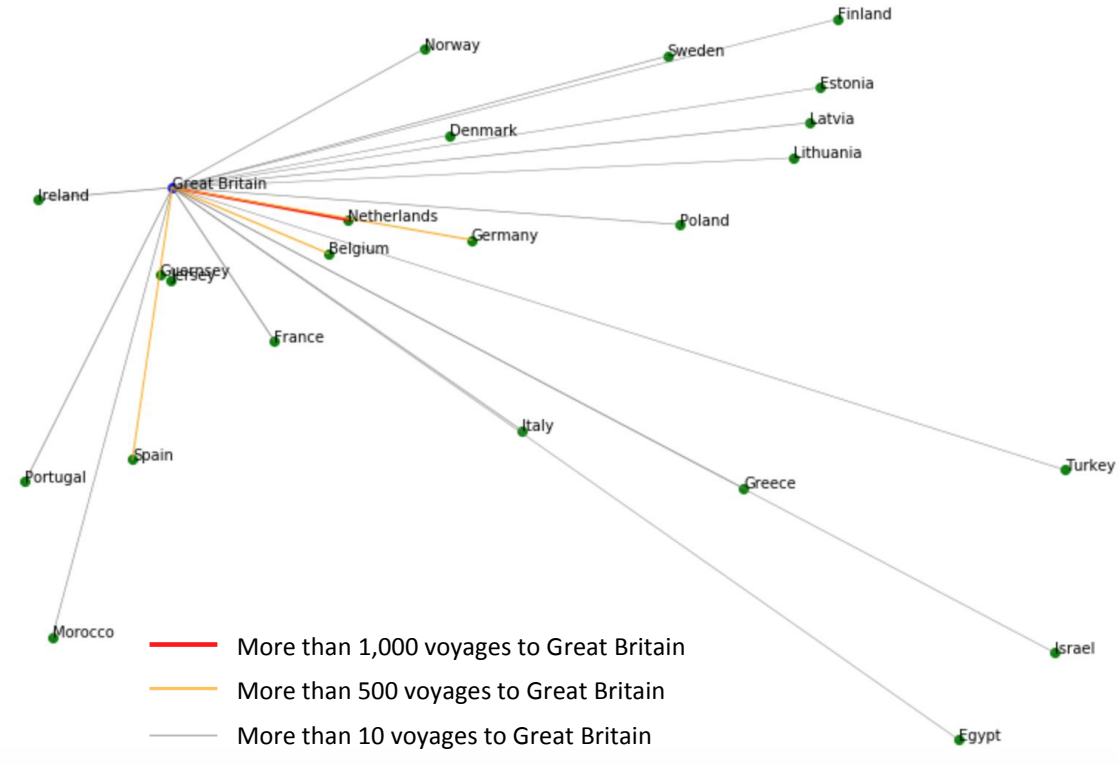


Figure 10: Voyages to Great Britain from countries within 3,500km

Focusing on Felixstowe emphasises the importance of Felixstowe as a port to Great Britain, with Felixstowe's most important neighbours aligning with those for Great Britain. Of all Felixstowe voyages, 25% are between Felixstowe and the Netherlands, 17% between Germany and Felixstowe and 10% between Belgium and Felixstowe.

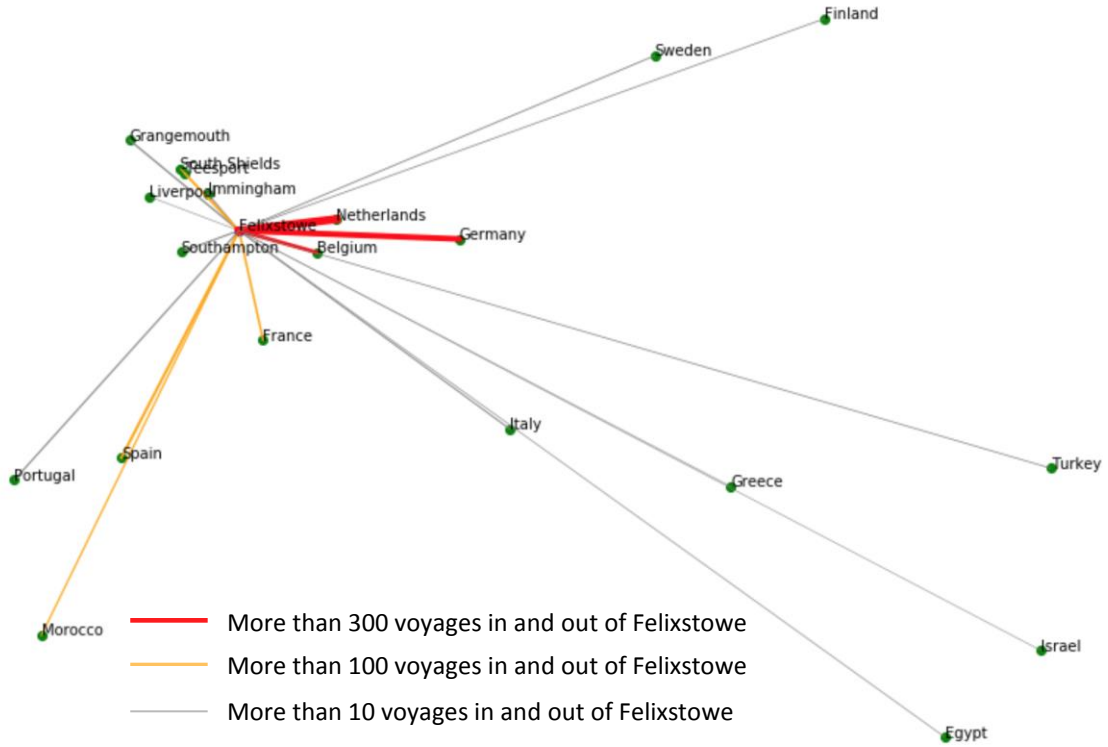


Figure 11: Voyages from countries within 3,500km in and out of Felixstowe

Hazardous materials

The movement of hazardous materials (hazmat) within UK ports is of particular interest. There are two fields within the CERS database that specify whether a ship enters and leaves a port carrying hazardous materials. These data were used to identify ports where a large proportion of ships either load or unload all or some of their hazmat cargo. In most cases, there are no significant differences between the proportions of ships entering and leaving a port carrying hazmat. However, in a few cases a difference is noted. Figure 12 shows that in Belfast 28% of ships enter port carrying hazmat and 48% leave carrying it, for Holyhead these figures are 25% and 80% respectively, which suggests that both ports send hazmat to other ports. Conversely, 37% of ships entering the port of Hull carry hazmat whilst only 31% leave Hull carrying it, suggesting that Hull accepts hazmat from other ports.

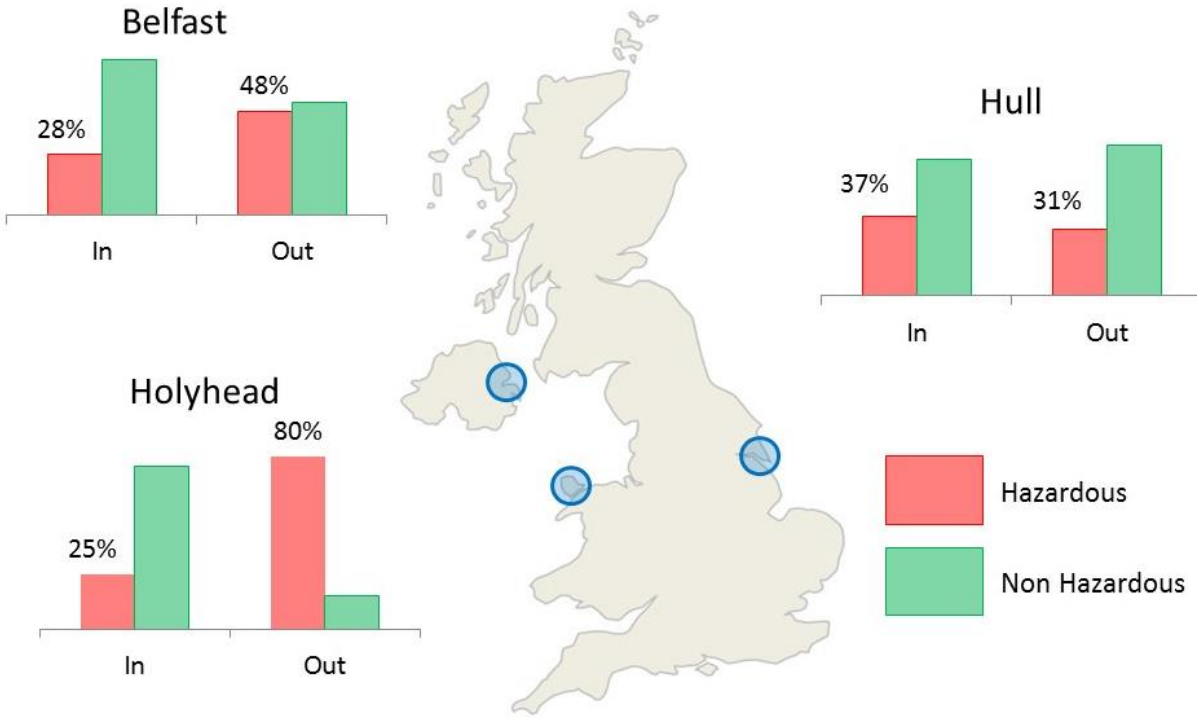


Figure 12: Moment of hazardous material in and out of port

Exploring these differences further, CERS data can be used to understand where the hazmat leaving Belfast and Holyhead goes to and where the hazmat unloaded at Hull comes from. In the case of Belfast (see Figure 13) the majority travels to Birkenhead, Heysham and Loch Ryan ports. In the case of Holyhead, almost all travels to the ports of Dublin (see Figure 14). Turning to imported hazardous material and the port of Hull: large proportions are imported from ports outside the UK, specifically Rotterdam, Antwerp, Oxelösund in Sweden and Luanda in Angola (see Figure 15).

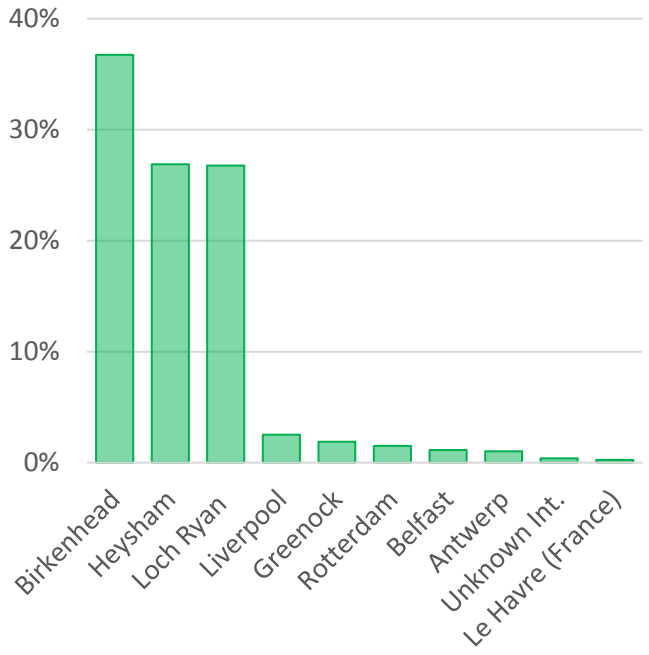


Figure 13: Destination of hazardous material loaded at Belfast as a percentage of all voyages

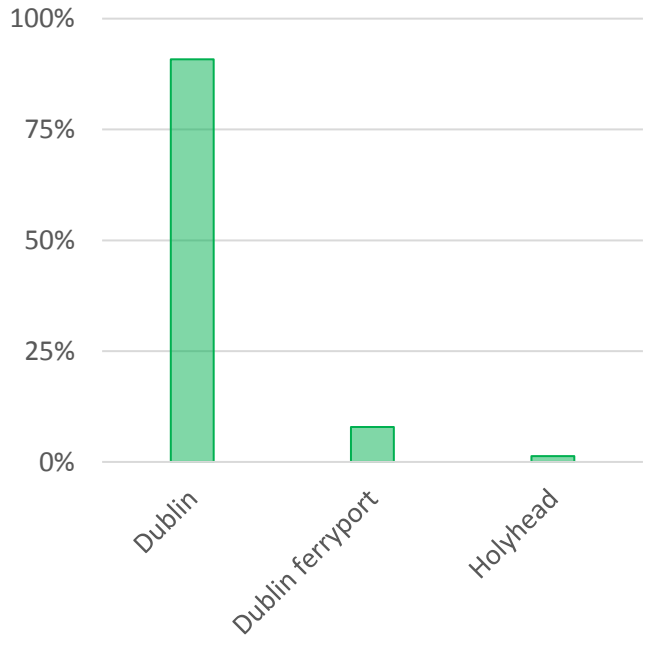


Figure 14: Destination of hazardous material loaded at Holyhead as a percentage of all voyages

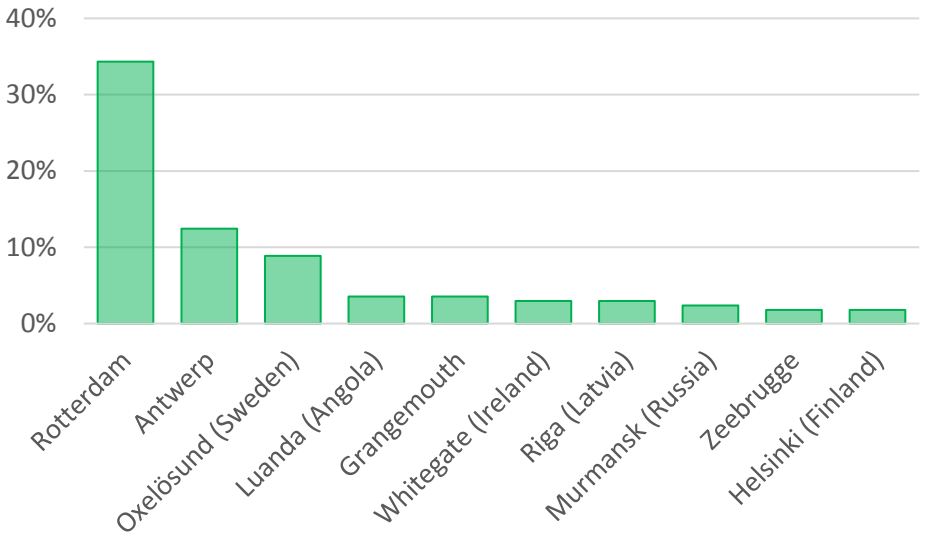


Figure 15: Source of hazardous material unloaded at Hull as a percentage of all voyages

Delays

Delays resulting from the late arrival of ships at port can have a significant operational and economic impact. Figure 16 gives the distribution of all arrival delays[5] within UK ports. As expected, delays are broadly normally distributed with the median value being located around zero. Just over 43% of ships are subjected to a delay, with 24% of all ships being delayed by an hour or more. This proportion drops to 17%, 12%, 8% and 6% for delays of two, three, four and five hours respectively.

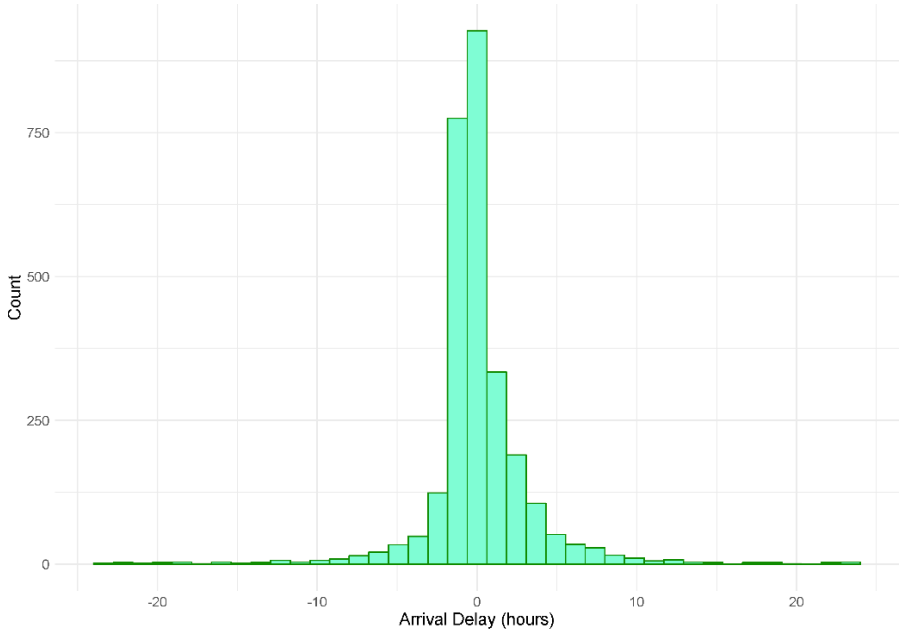


Figure 16: Distribution of arrival delay (measured in hours) for all ports
A positive value indicates a delay, whilst a negative value indicates early arrival

[5] For a full definition of delays see the 'Predicting delays' section of this report.

When the distribution of arrival delays is plotted for each day (Figure 17) the distribution changes little, suggesting that arrival delay is independent of the day of arrival. However, the charts suggest that an effect is evident when broken down by time of the day (Figure 18) and more notably seasonality (Figure 19). In the former example fewer ships are delayed at night time and in the early hours of the day whilst more ships are delayed during the start of the working day. In the latter case, fewer ships are delayed in the spring and summer seasons.

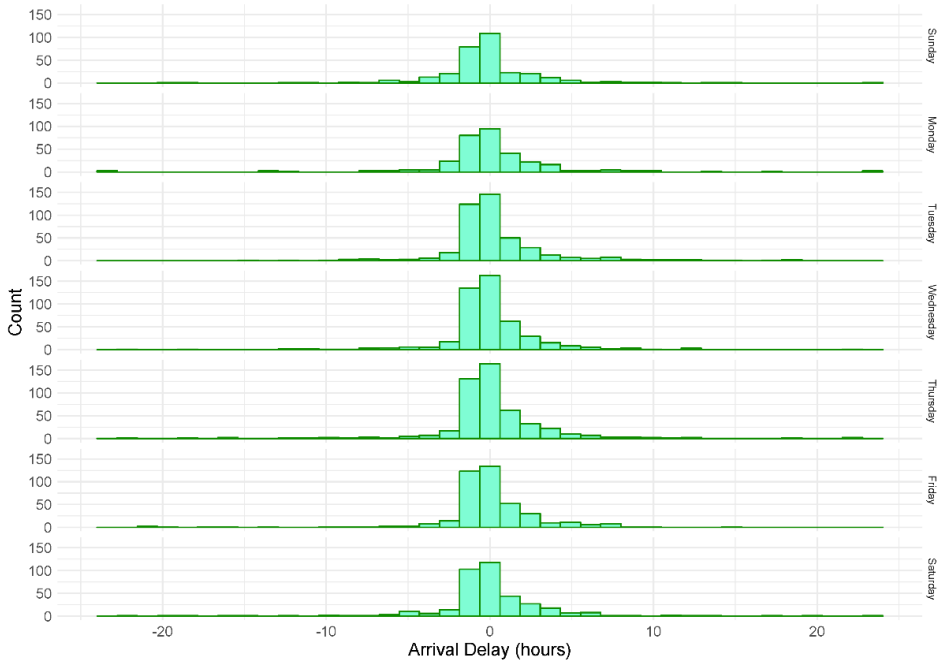


Figure 17: Distribution of arrival delay (measured in hours) for all ports, broken down by day of arrival
A positive value indicates a delay, whilst a negative value indicates early arrival

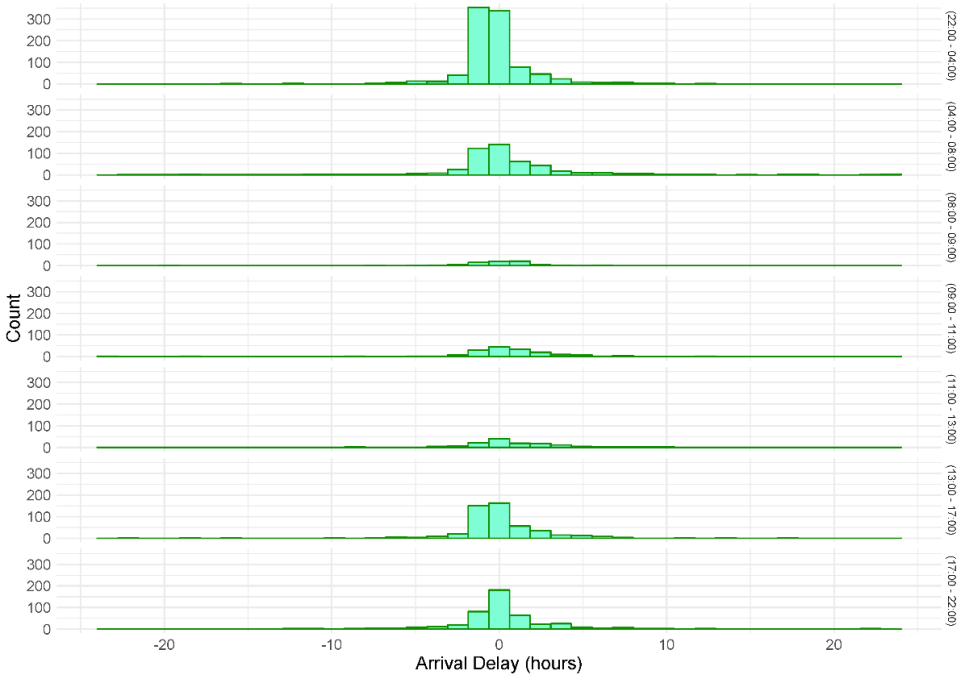


Figure 18: Distribution of arrival delay (measured in hours) for all ports, broken down by time of arrival
A positive value indicates a delay, whilst a negative value indicates early arrival

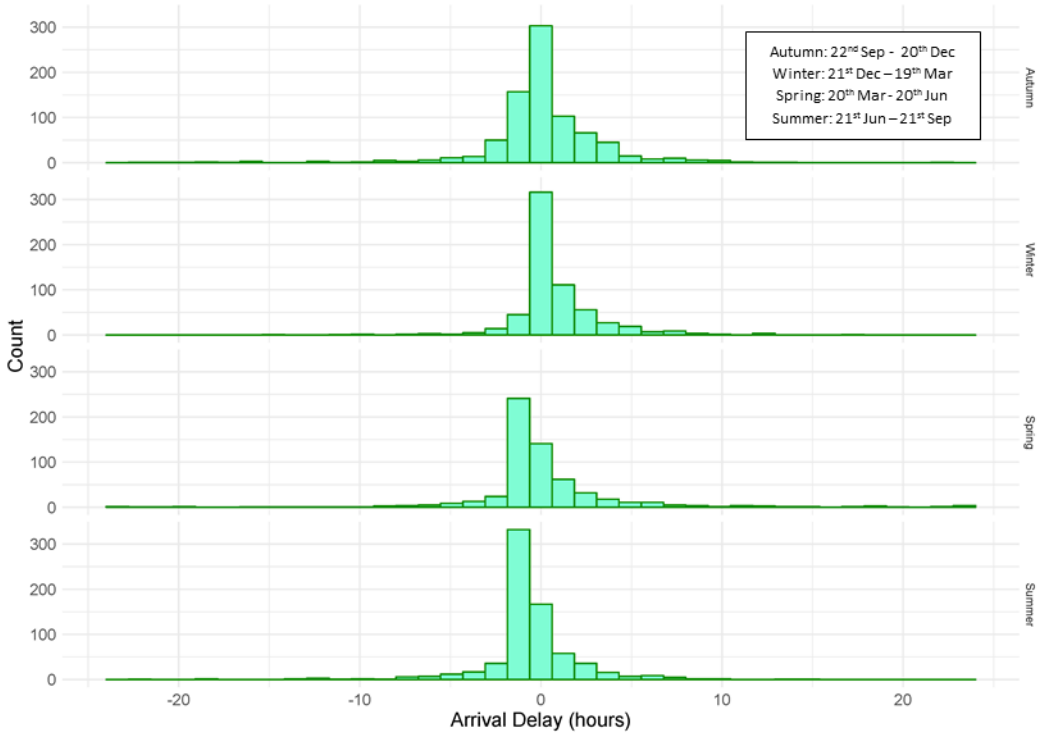


Figure 19: Distribution of arrival delay (measured in hours) for all ports, broken down by season of arrival
A positive value indicates a delay, whilst a negative value indicates early arrival

3.3. Automatic Identification System (AIS)

AIS was first developed in the 1990s for use as a short-range identification and tracking system used on ships and other marine traffic. On board AIS equipment (see Figure 20) allows ships to view traffic in their local area (10 to 20 nautical miles) and to simultaneously be seen by that traffic.

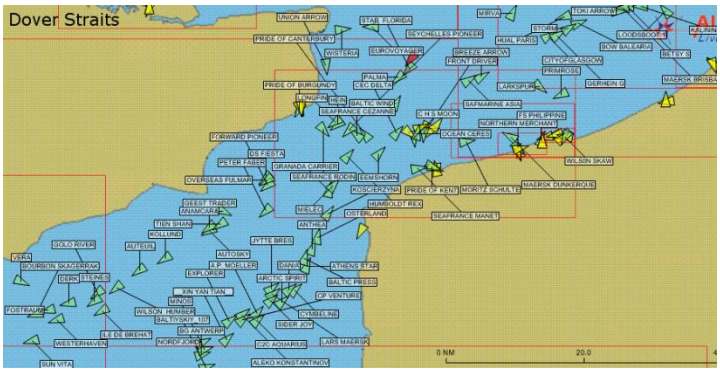


Figure 20: On-board AIS equipment and typical graphical display

More recently the AIS system has been used to support:

- collision avoidance, notably amongst vessels outside the range of shore-based systems
- fishing fleet monitoring and control
- vessel traffic services, used to augment existing systems such as local vessel traffic service (VTS)

- maritime security, to identify and monitor suspicious activity patterns
- aids to navigation, which may support or replace information generated by radar beacons currently used for electronic navigation aids
- search and rescue., coordinating on-scene resources of a marine search and rescue (SAR) operation
- accident investigation, AIS information is more accurate and comprehensive than transitional systems such as radar
- ocean current estimates
- fleet and cargo tracking

There are 27 types of AIS message. Of these, two are relevant to this report as they relate to ship position and dynamics. Class A messages are sent by large ships typically over 300 tonnes and those carrying passengers. Class B messages are used by lighter commercial and leisure craft. In both cases, an AIS transmitter will send the following information every 2 to 10 seconds when underway and every three minutes when stationary or at anchor:

- Maritime Mobile Service Identity (MMSI) the unique ship identifier
- navigation status, at anchor, under way using engine(s), not under command an so on
- rate of turn, right or left, from 0 to 720 degrees per minute
- speed over ground, 0.1 knot resolution from 0 to 102 knots
- longitude, accurate to 0.0001 minutes
- latitude, accurate to 0.0001 minutes
- course over ground, relative to true north to 0.1 degrees
- true heading, 0 to 359 degrees
- true bearing at own position, 0 to 359 degrees
- UTC seconds, the seconds field of the UTC time when the data were generated

Higher-level information is transmitted less frequently (every six minutes):

- Maritime Mobile Service Identity (MMSI), the unique ship identifier
- radio call sign, up to seven characters
- name of vessel
- type of ship and cargo
- ship dimension
- location of positioning system (such as GPS) antenna on board the vessel
- type of positioning system – such as GPS, DGPS or LORAN-C
- draught of ship – 0.1 to 25.5 metres
- intended destination
- ETA, estimated time of arrival at destination

A recent development of the AIS system is the ability to make it viewable on the internet without the need for a dedicated receiver. This removes the range limitation of the marine-based receivers and allows AIS data to be visualised over a wider area (see Figure 21). With permission, data may also be downloaded to be used for additional offline processing.

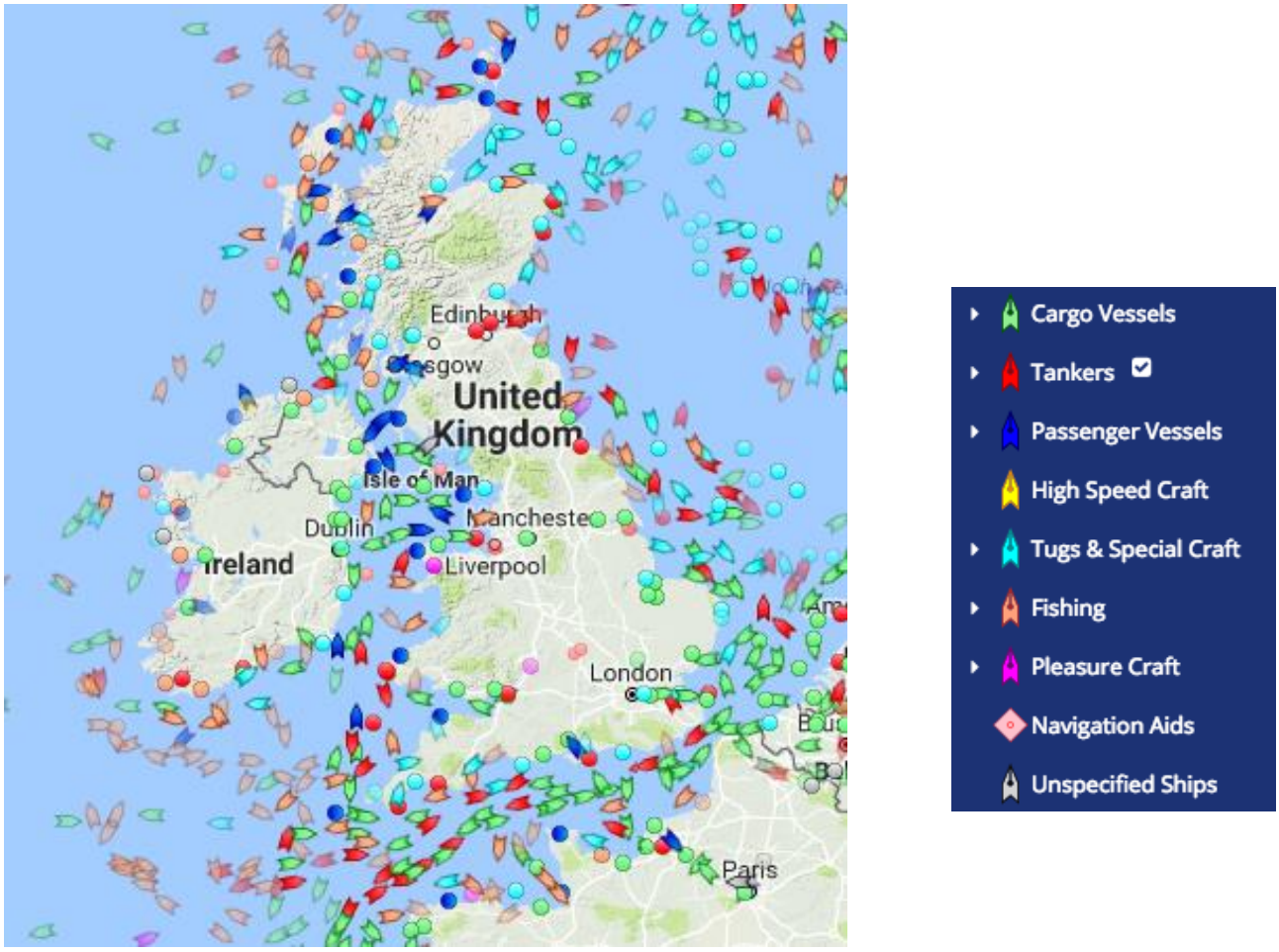


Figure 21: Online AIS data showing the position of ships around the UK at noon on the 25 April 2018

3.4. Processing AIS data

The AIS data used for this study spans for a period of 12 months, between 1 August 2016 and 31 July 2017. The raw data encoded using the National Marine Electronics Association (NMEA) format, was approximately one terabyte in size and consisted of almost three billion rows. A Hadoop Distributed File System (HDFS) environment was used to decode the encoded messages, process the new data, filter out the required types of messages and extract smaller slices of data based on certain criteria, as described in the next subsections.

Once the data was saved in the HDFS environment, it was accessed using Pig, Hive and Spark as part of the Hadoop stack. Decoders for individual types of message were written in Scala (see next section), the raw data was decoded in a table format and saved as Parquet files. Scala functions were also developed to filter out the data based on the type of message, unique ID of ship, timestamp, geographic location and other criteria. After filtering, the data was extracted from the Hadoop environment, stored in local machines and further processed using Python and R (see Figure 22).

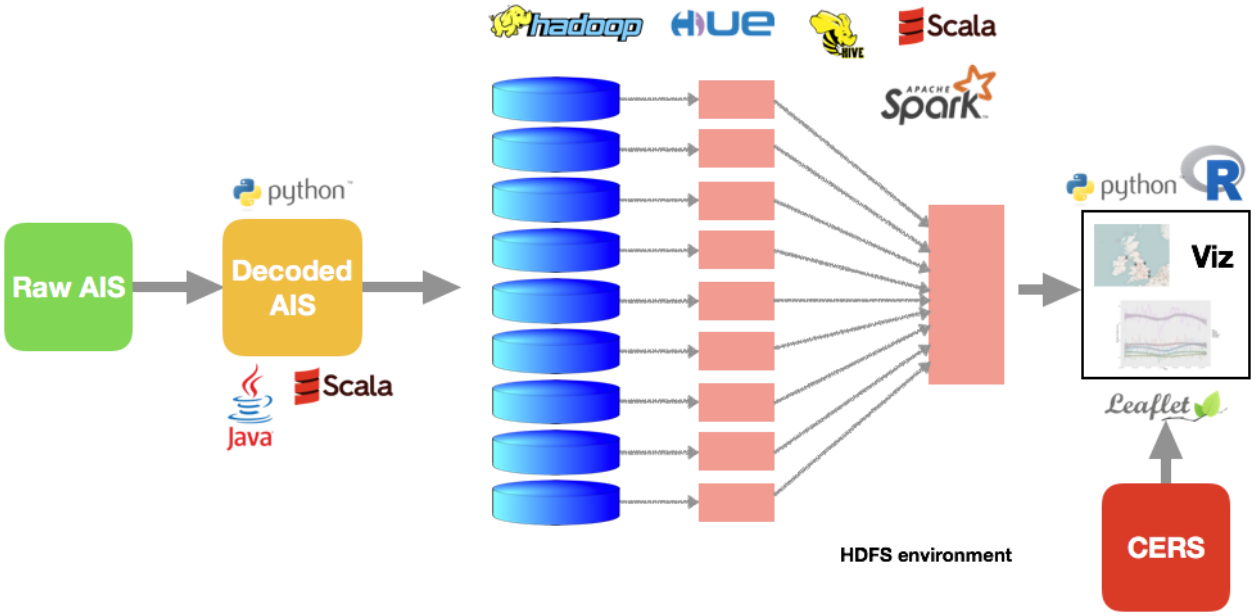


Figure 22: Cloudera distributed system environment

3.5. Decoding AIS data

AIS data was extracted in standard National Marine Electronics Association AIS message format; this is a text encoded binary format that was decoded before being split into a series of files based upon the AIS message type. As some messages are split over multiple lines (see Figure 23), they were concatenated before being decoded into real data that can be directly interpreted and processed.

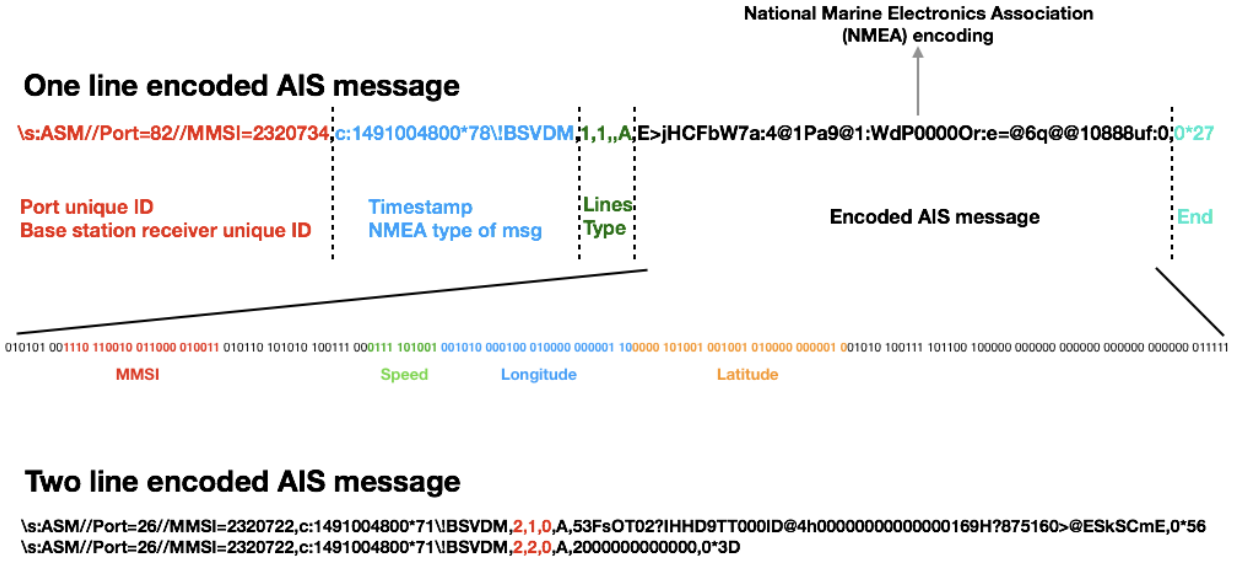


Figure 23: One-line and two-line examples of raw AIS messages

The component of the positioning Type 1 AIS message that was of specific interest is:

\!BSVDM,1,1,,A,E>jHCFbW7a:4@1Pa9@1:WdP0000Or:e=@6q@@10888uf:0,0*27

The message is comma-separated with each element of the message decoding to a unique piece of information (Table 1).

Component	Description
\!BSVDM	The NMEA message type
1	Number of message lines
1	Sentence number (1 unless it´s a multi-sentence message)
	Sequential message ID (for multi-sentence messages)
A	AIS Channel (A or B)
E>jHCF...	Encoded AIS data
0*	End of data terminator
27	NMEA checksum

Table 1: AIS decoded message components
(NMEA provide a full review)

Message integrity was checked using the NMEA checksum and corrupted messages discarded. The multi-sentence messages were then assembled together into single messages, through concatenation of the encoded AIS data parts. In NMEA AIS encoding, each ASCII character corresponds to six binary bits (unlike normal ASCII which uses eight). To account for this, the decoding algorithm steps through each character of the encoded AIS data and subtracts 48 from the decoded value. If the resulting number is a decimal number with a value greater than 40, the algorithm again subtracts eight. The resulting number is then converted to a binary string that is split into substrings using the message element positioning given in the NMEA encoding specification. Table 2 provides samples of the relevant splitting positions for an AIS message.

Element	Position
MMSI Number	from bit 8 for 30 bits
Longitude	from bit 61 for 28 bits
Latitude	from bit 89 for 27 bits
Course	from bit 116 for 12 bits
Heading	from bit 128 for 9 bits

Table 2: AIS message splitting positions

The data was finally split according to the message type and saved in intermediate storage. During saving, the data was partitioned according to the time stamp of the AIS message. The parquet storage format was used as it offered good compression, incremental addition of new data, most importantly, the ability to read only specified segments instead of the full dataset.

One important feature that was considered when processing the AIS data was the ability to extract the messages within a specified time window and for a specific area of interest. The area of interest was defined as a rectangular geometric shape specified by the latitudinal and longitudinal coordinates of the top-left and bottom-right corners. The data was read from the compressed parquet storage format using predicate pushdown based on the requested time window and filtered in a distributed manner based on the coordinates for each message. Then the data was collected into the driver node and exported in comma-separated values (csv) format.

All the above procedures were developed within the Apache Spark distributed computing engine and the data was stored in Apache Hadoop. For this project, only data from the waters around the UK was used, however the linear scalability of the distributed computation and storage allows the algorithm to be easily applied to data at a global scale.

3.6. Visualising AIS data

Once the AIS data has been decoded into a series of latitude and longitudinal pairs it can be plotted and superimposed onto a map or nautical chart. These plots can be used to gain an understanding of ship positions at a given point in time and ship movements across any given time window.

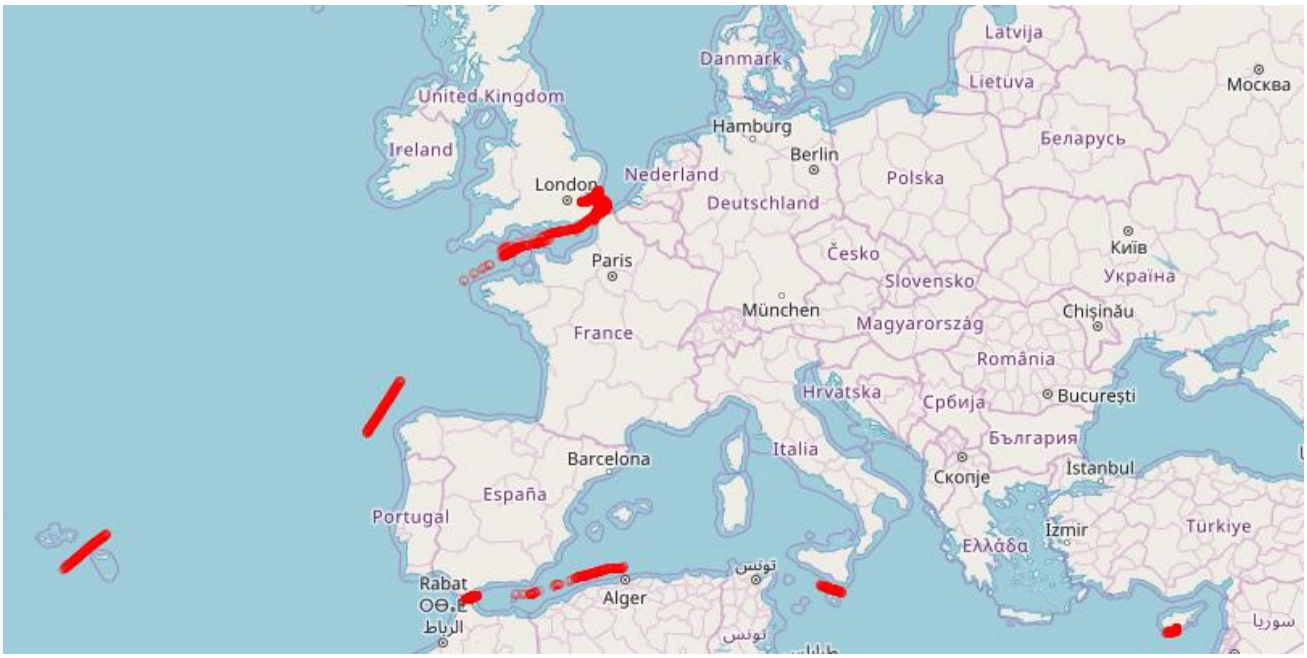


Figure 24: AIS base-station data for a single ship

Figure 24 gives the AIS track for a single Liquefied Natural Gas (LNG) tanker[6]. The diagram shows that ship travels through the Azores (bottom left of the chart) into the English Channel and docks in London Medway port. The ship then leaves port and once again passes through the English Channel, before heading south past the north-western tip of Spain. The ship turns east and then continues its journey through the Strait of Gibraltar on through the Mediterranean, finally docking in Cyprus.

It should be noted that the gaps in the base-station data relate to missing AIS coverage caused by the ship being outside the range of ground stations. In such cases, AIS coverage is maintained by using satellite tracking. As all UK ports and their surrounding waters are within range of at least one base station, AIS coverage is complete in the data used for this project.

One particular area of interest in the voyage of this tanker can be seen due east of London Medway port (Figure 25).

[6] The data for this was provided by Centre for Big Data Statistics at Statistics Netherlands (CBDS) and did not form part of the data extract used for the remainder of the project.

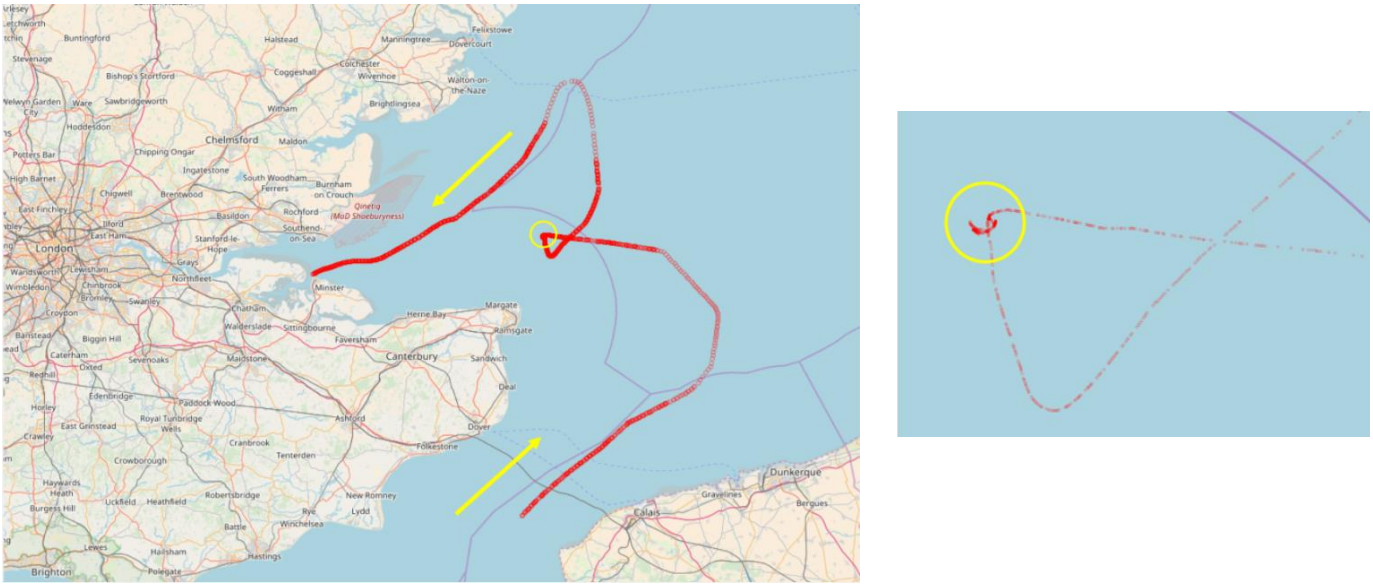


Figure 25: AIS base-station data for a single ship (enlarged view)

One may reasonably expect that the tanker would follow a smooth arching path starting north-east through the channel and turning onto a south-westerly bearing into port. However, the ship comes to a stop and remains stationary for several hours (circled), before heading away from port on a due north heading before turning through 180 degrees and travelling south-west into port. Conversation with domain expert have indicated that this behaviour is indicative of the ship waiting for cargo price to increase before it enters port and docks. Although the behaviour of the above tanker may appear counterintuitive in open sea, it becomes more consistent as it approaches port, as the ship enters and leaves in a more uniform manner. However, this consistency is not observed in all ports.

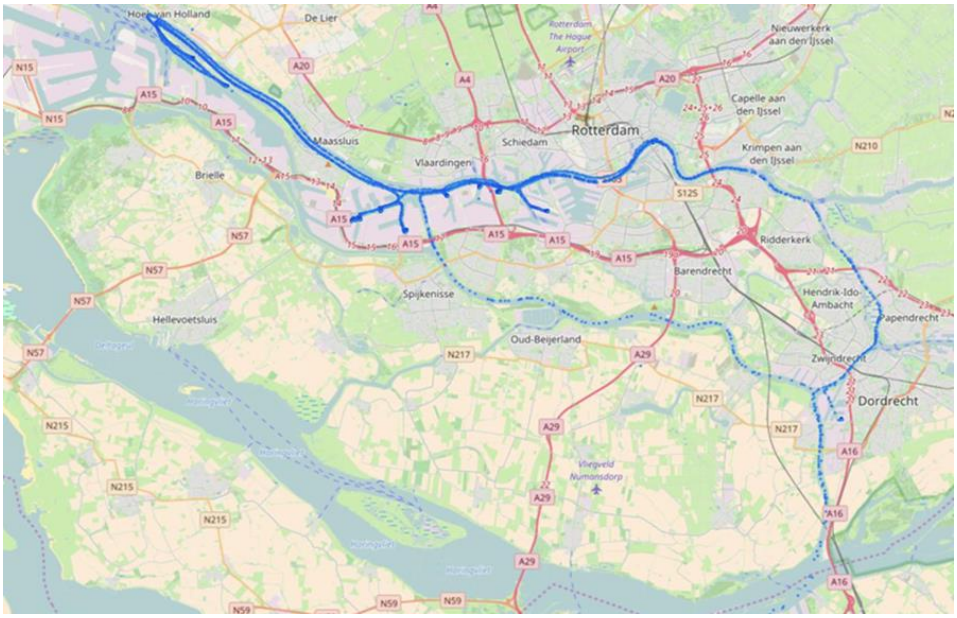


Figure 26: AIS base-station data for a single ship (port of Rotterdam)

Figure 26 shows the journey of a light general transporter in and around the port of Rotterdam[7]. Unlike the LNG tanker discussed earlier, the ship in this instance calls at several berths within the port, loading and unloading cargo throughout.

A final example is discussed below. The tracks relate to several journeys undertaken by a large 200,000 tonnes, 400m by 60m container ship within the port of Felixstowe. The highlighted area illustrates that on at least two occasions the tanker heads north before turning sharply through 180 degrees and heading south before docking at a berth it had previously travelled past (see Figure 27). On first inspection this behaviour may appear unexpected, however it is an accepted practice for a docking ship to manoeuvre against the prevailing inbound current caused by the flooding tide to aid the docking process.

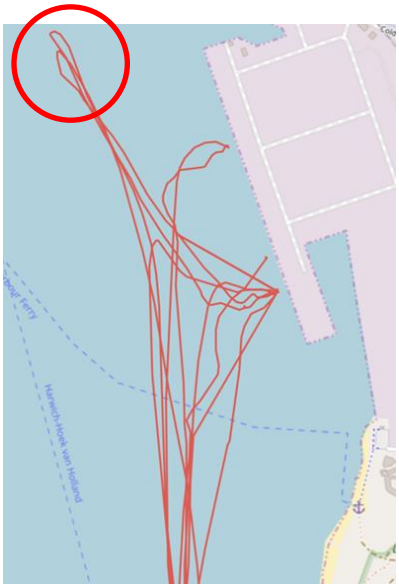


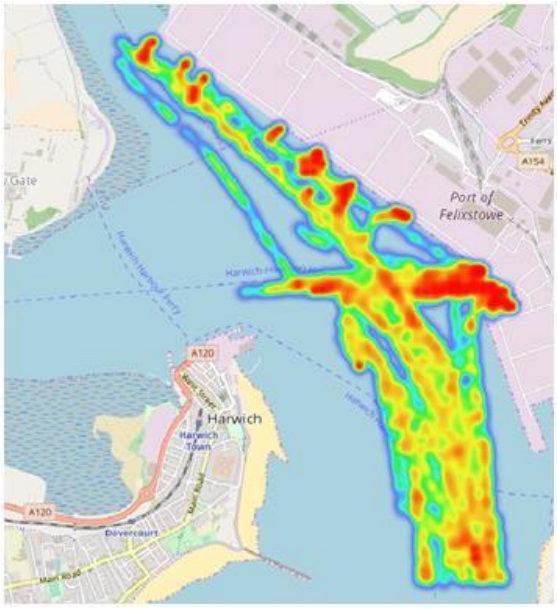
Figure 27: AIS base-station data for a container ship

To this point, AIS data has been used to explore the voyages of individual ships. However, the daily points taken from within an area can be combined to produce a heatmap of ship movements within a port. This can be used to indicate shipping lanes, port berths and port loading over time.

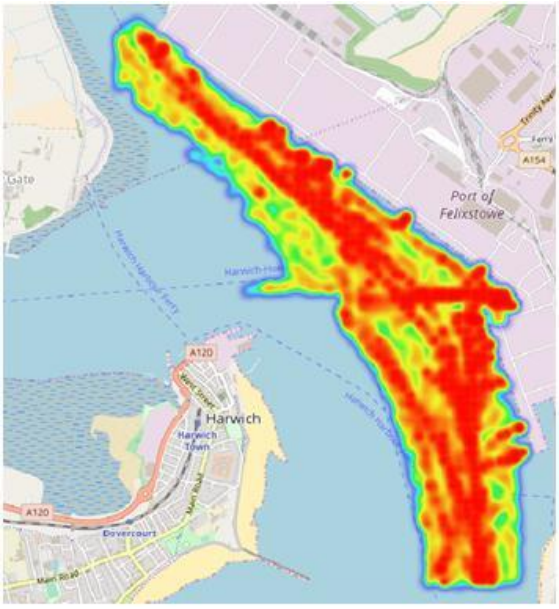
Figure 28 highlights notable features relating to Felixstowe (the busiest container port in the UK), including:

- the emergence of "hot" regions indicating unique berths within the northern edge of the port, most notably on 12 December and 1 January
- evidence that not all ships that enter Felixstowe go on to dock at the port, some ships turn east and head towards the international port of Harwich whilst others head north west onto the river Orwell and onto inland destinations
- port loading on Christmas day is considerably lower than on other days

[7] The data for this was provided by Centre for Big Data Statistics at Statistics Netherlands (CBDS) and was not part of the data extract used for the remainder of the project

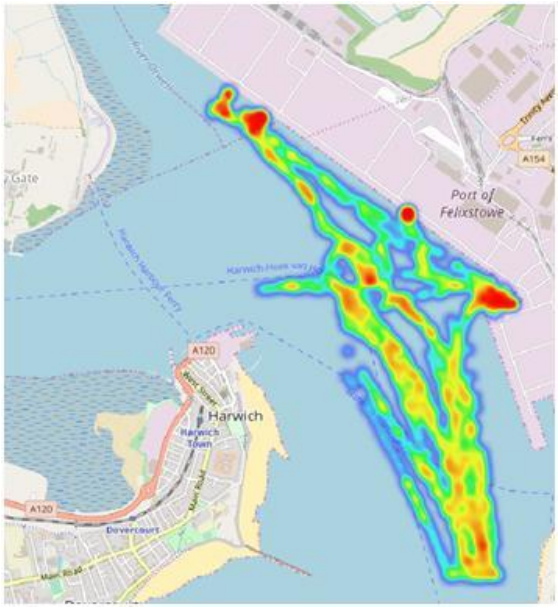


Monday 12 December 2016



Monday 19 December 2016

Sunday 25 December 2016



Sunday 1 January 2016

Figure 28: Port loading, Felixstowe (from green, through amber to red represents an increasing density of AIS points)

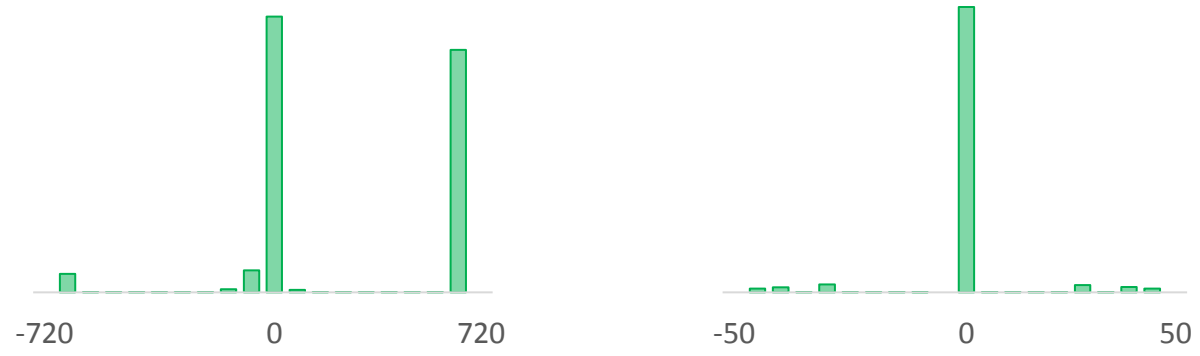
4. Classifying AIS data

The previous section presents examples of how the visualisation of AIS tracks within a given port and on the open sea can be used to support the high-level understanding of ship and port behaviour on a macro level, for example, identifying shipping lanes, holding areas, port berths, port loading and so on. However, this visually driven approach is of limited use in quantitatively classifying behaviour on an individual ship-by-ship basis. To address this, various machine learning techniques were applied to the AIS data to extract actionable insight. An unsupervised k-means classifier[8] was used to segment the behaviour of a ship into one of many intuitive states (transitioning through port, manoeuvring into dock, docked and so on). These behavioural states transform the AIS data to a more meaningful and interpretable representation which can be used to better understand the behaviour of a ship at any point on its voyage.

AIS data was extracted from the port of Felixstowe and its surrounding sea spanning a 12-month time window from 1 August 2016 to 31 July 2017. Several features within the AIS data were investigated as potential inputs to the segmentation, these including:

- Speed over Ground (SOG) (knots)
- acceleration (derived from SOG) (knots per second)
- Rate of Turn (ROT) (degrees per second)
- bearing (degrees)
- heading (degrees)

It is critical that any classification should be sufficiently robust so that it may be applied to the ships at any port or at any geographic area in open sea. Early investigations showed that a segmentation using either bearing or heading as features, although powerful from a classification perspective, would fail from the standpoint of robustness. Consider a segmentation trained upon the data from ships operating in and around the port of Southampton on the south coast of England. It follows that a ship heading into dock would predominantly be heading in a northerly direction. During training, the machine learning algorithm would learn that ships heading north would be entering port whilst ships heading south would be leaving port. If this pre-trained segmentation was then applied to a port on the north coast of Scotland where ships head south to enter port and north to leave port, the segmentation would clearly classify incorrectly. The impact of this could be mitigated to a degree by sampling AIS tracks from ports at different orientations; however, the problem would not be completely removed since UK ports do not cover every possible orientation.



Between -720 and 720 degrees per minute Between -50 and 50 degrees per minute

Figure 29: Distribution of ROT

[8] Kmeans clusters aims to find similar groups or segments within the data, with the number of groups represented by the parameters *k*. It works by iteratively assigning each data point to one of *k* groups based on the features that are provided. Data points are clustered based on feature similarity (MacQueen, 1967).

As speed over ground (SOG) and rate of turn (ROT) do not suffer from these drawbacks they were chosen as more suitable inputs to the segmentation. On closer inspection, the distribution of the ROT variable taken directly from the AIS sample did not follow the expected normal distribution - a large proportion of records are located at very large ROT values (see Figure 29).

A new ROT variable was derived directly from the latitudinal and longitudinal points as shown below.

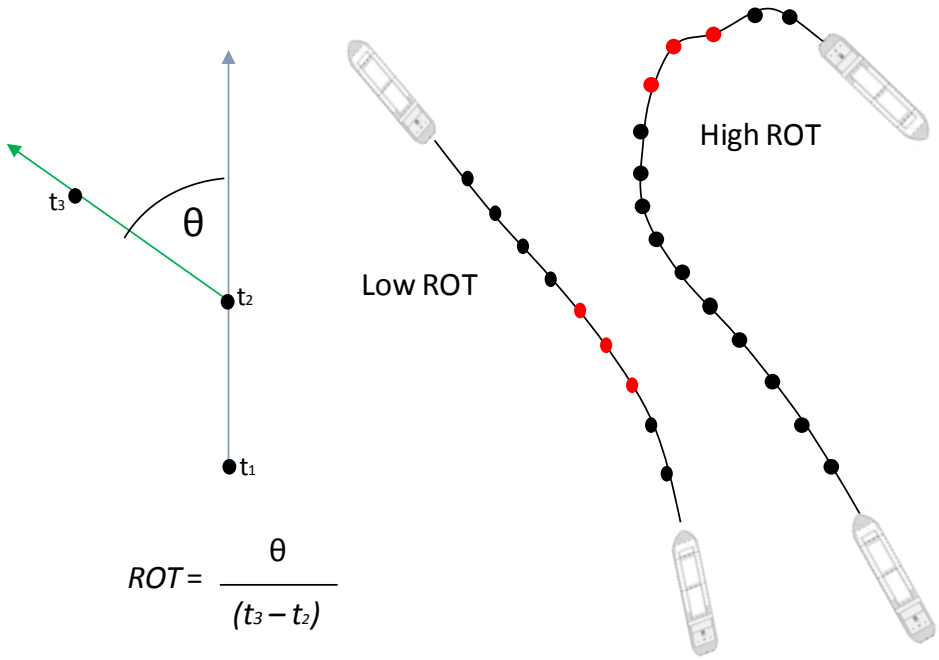


Figure 30: Calculating ROT

Figure 30 shows three latitudinal and longitudinal pairs that describe positions on the sea at times t_1, t_2 and t_3. The ROT at t_2 is given by firstly calculating the angle of turn θ at t_2 using standard trigonometric techniques. This angle is then divided by the time taken to travel between t_3 and t_2 to give ROT. The two ship tracks on the right-hand side illustrate the differences between high and low ROT.

A kmeans clustering algorithm was applied to training data containing SOG and ROT pairs taken in isolation at every point along the track of each ship (Figure 31). In this instance, if the voyage of a single ship contains 1,000 AIS points all 1,000 points were used for training purposes.

To remove noise created by smaller ships such as tugs and ferries, AIS data was extracted for the following ship types:

- container ship
- general cargo ship
- chemical and oil product tankers
- cargo ship
- bulk carrier

This reduced the number of available AIS data points from 10.1 to 2.3 million, corresponding to 772 unique ships.

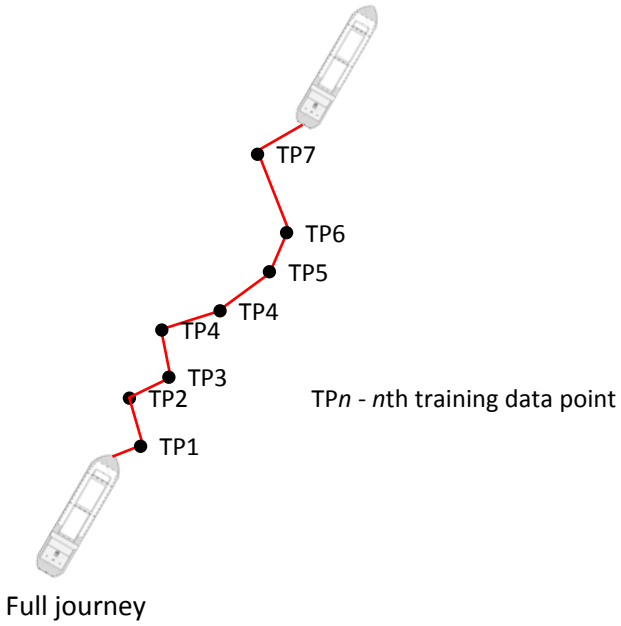


Figure 31: Training data for a single ship
(Simplified for clarify, in reality each ship journey contains many more AIS data points)

The approach identified several unique segments. However, most identified segments were noisy with very few containing unique or intuitive behaviours. As this approach treats each AIS data point in isolation, with no consideration given to the relationship between points over time, it is overly sensitive to noise within the AIS data, caused by atmospheric effects, obsolete equipment or on-board interference. A more robust approach was developed that added a time-based aggregated component to the SOG and ROT fields that reduces the sensitivity of the segmentation to AIS signal noise. For each journey, a random slice of AIS data was taken during the voyage (see Figure 32). Several different time windows we investigated ranging from one to 10 minutes. A two-minute slice was found to give the best balance between sensitivity and robustness.

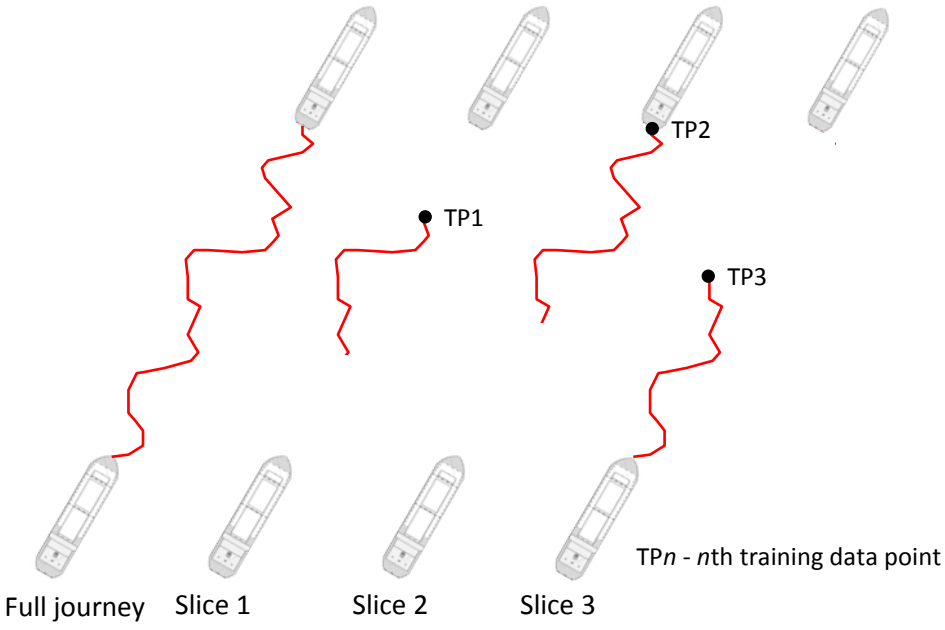


Figure 32: Three random journey slices taken from the path of a single ship

Histograms of the SOG and ROT values were then created and converted into two state vectors, **SOG** and **ROT.** Essentially, each state vector represents a historical distribution of the individual SOG and ROT values across the preceding two minutes. The bin boundaries for the SOG and ROT state vectors were selected to ensure approximately equal quantile population density across the entire training dataset. The final boundaries are given in Table 3,

Quantile	Range
1	SOG = 0
2	0 < SOG ≤ 1
3	1 < SOG ≤ 2
4	2 < SOG ≤ 3
5	3 < SOG ≤ 5
6	5 < SOG ≤ 10
7	SOG > 10

Quantile	Range
1	ROT ≤ 0.1
2	0.1 < ROT ≤ 0.6
3	0.6 < ROT ≤ 2.5
4	2.5 < ROT ≤ 8
5	ROT > 8

Table 3: Bin boundaries for SOG and ROT

Figure 33 gives the ROT and SOG state vectors for a given point (TP1) on a hypothetical voyage. In this example the ROT distribution assumes an approximate log normal distribution whilst the SOG distribution is exponentially distributed, it can therefore be concluded that in the preceding two minutes this ship is generally travelling at lower speeds and with limited rate of turn.

One disadvantage of this approach stems from the fact that the data is aggregated over all the AIS points within a two-minute window. If only one slice is taken from each ship voyage then the resulting training data-set will be significantly smaller than that generated in the previous approach. To overcome this, several random slices are taken from each voyage to maintain approximately consistent training dataset sizes.

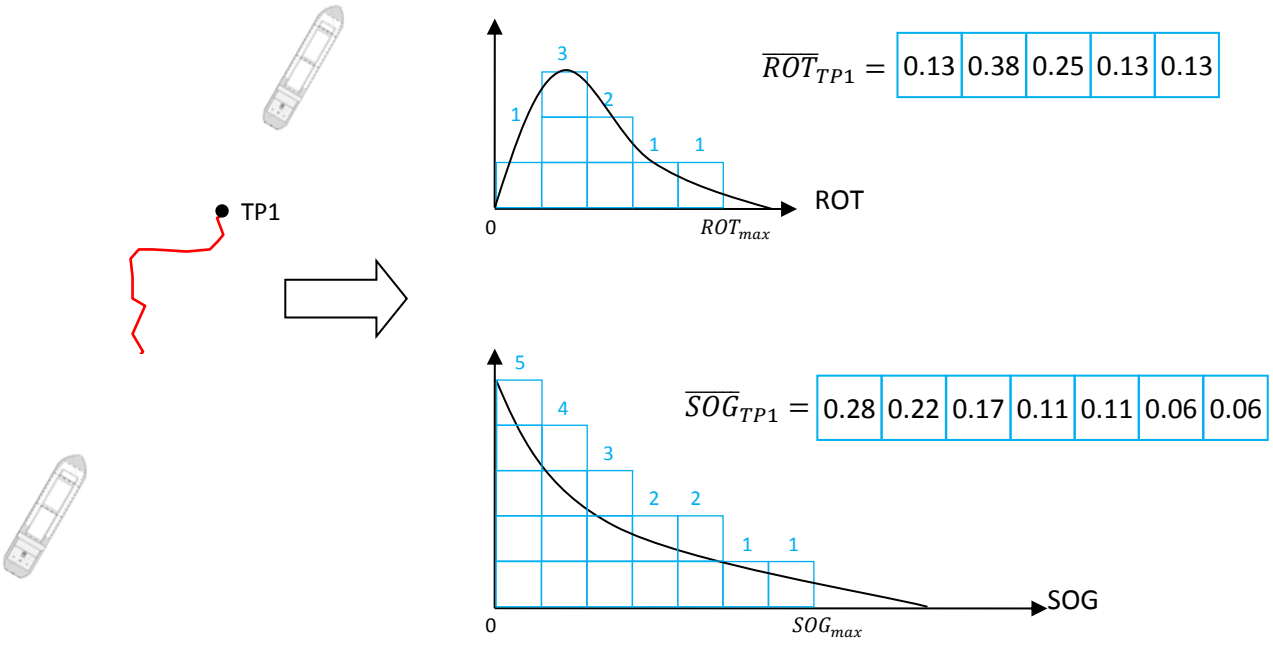


Figure 33: The SOG and ROT state vectors for a given journey slice

As with the previous approach, a k-means unsupervised algorithm was applied to the training data, the number of generated segments (*k*) was set to an arbitrarily large value of eight. Segments that were similar based upon their centroids as defined by the ROT and SOG state vectors were merged. This gave a final set of six unique segments. These segments were each assigned descriptive name to aid interpretation. Each of the six segments fall into the following three higher-level behavioural groups:

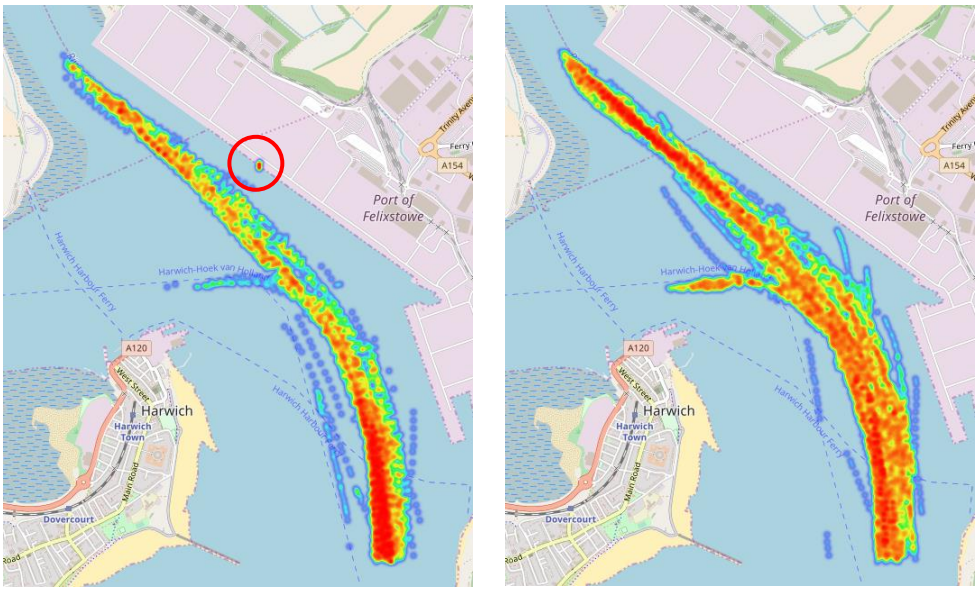
- transitional behaviour
- docking behaviour
- docked behaviour

The following section discusses each of the six segments in turn. In each case, a table indicates where each segment over indexes (green) within the SOG and ROT input state vectors. A heat map then gives all the AIS points that fall into each of the segments (each percentage value gives the number of training points that fall into that segment).

Transitional segments

Ships within the two transitional segments (Figure 34) do not dock within the port and are not in the process of docking. Instead they use the port to transition into other inland areas. In the case of Felixstowe this is either heading east to the port of Harwich or north-west to other inland destinations.

	Speed over Ground (SOG)							Rate of Turn (ROT)				
	0	0-1	1-2	2-3	3-5	5-10	10+	0.1	0.1-0.6	0.6-2.5	2.5-8	8+
Border phase												
General phase												



Border phase
(4% of training dataset)

General phase
(57% of training dataset)

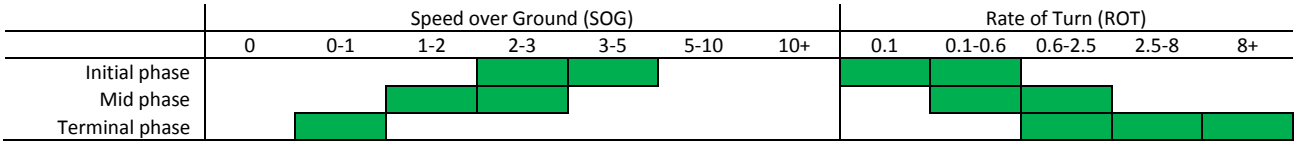
Figure 34: Transitional segments, (from green, through amber to red represents an increasing density of AIS points)

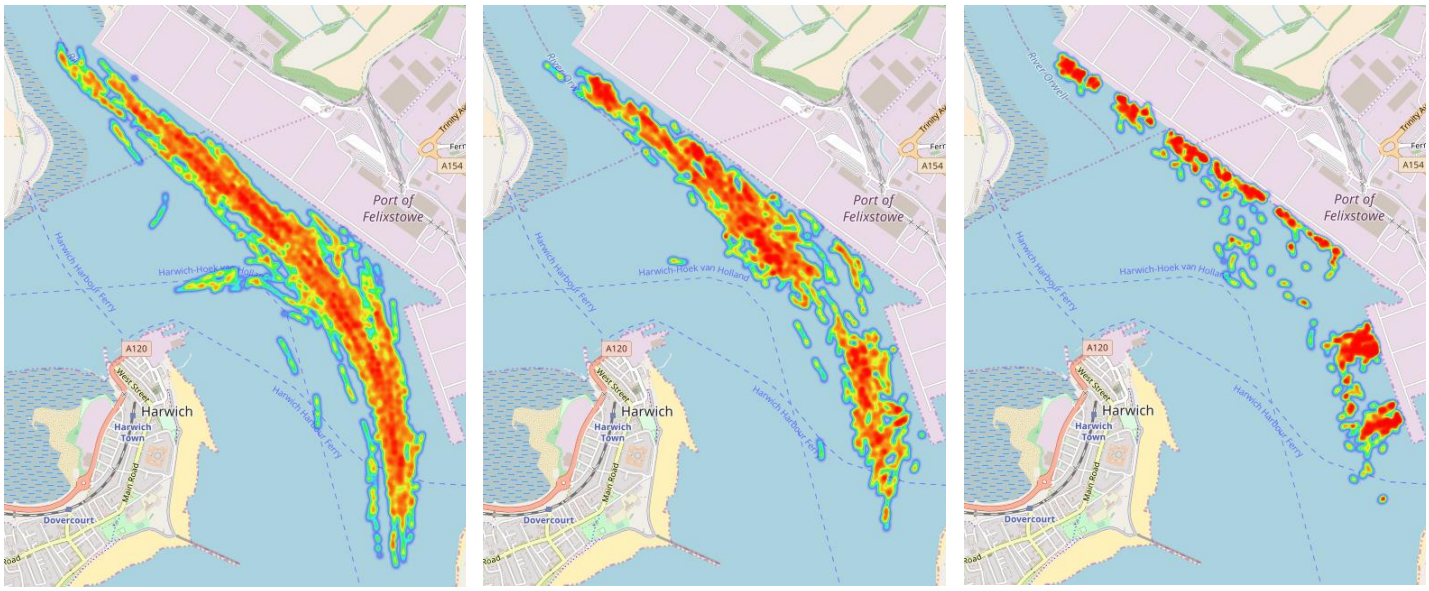
The border phase segment has the highest speed of all segments, its ROT is typically low indicating that ships in this segment are not manoeuvring to any significant degree. The heatmap shows that there is a higher density of points located around the southern border of the port. It is suggested that this segment represents ships that are entering port and slowing down from their open-water cruising speed to the speed limit within the port. The general phase segment is closely related to the border phase segment. Although the ROT distributions are similar, the speed of the ships in this segment is significantly lower. This coupled with the heatmap, which shows no noticeable increase in density, suggests that these ships have passed the port boundary and are observing the local speed limit whilst they transition through the port.

One interesting feature of the border phase plot illustrates the noise that is sometimes present in the AIS data. The circled point relates to the AIS reading from a ship that is positioned against the northern harbour wall. However, the SOG reading for this point indicates that the ship is travelling at over 10 knots. This is clearly impossible and is likely to be caused by noise within the AIS data, however the machine learning approach is robust enough to handle this and other data anomalies.

Docking segments

The docking segments (Figure 35) classify ships that have initiated the process of docking into one of the harbour berths. There are three docking segments, each one relates to a unique phase of the docking process.





Initial phase
(10% of training dataset)

Mid phase
(5% of training dataset)

Terminal phase
(13% of training dataset)

Figure 35: Docking segments, (from green, through amber to red represents an increasing density of AIS points)

The speed of ships within the initial phase segment has dropped below that of the transitional phase ships. Their AIS points on the heat map indicate that they are turning towards the harbour berths. Ships classified within the mid-phase segment have decreased their speed further and are also starting to manoeuvre into dock (indicated by increasing ROT values). The final docking segment, terminal phase, is characterised by very slow speeds (less that one knot) and high ROT values. Ships in this segment are in the final stages of docking and are turning into their intended berth at very slow speeds and higher rates of turn. The heatmap for this illustrates the vicinity of the ships to the harbour wall.

Docked segment

The final segment relates to ships that have reached their destination (Figure 36). Ships within this segment have virtually no speed, the high rates of turn relate to the very final stages of the docking process where final adjustments are made to the ship position within the port. This heatmap for the segment confirms that the ships are all docked or very close to being so.

	Speed over Ground (SOG)							Rate of Turn (ROT)				
	0	0-1	1-2	2-3	3-5	5-10	10+	0.1	0.1-0.6	0.6-2.5	2.5-8	8+
Docked												

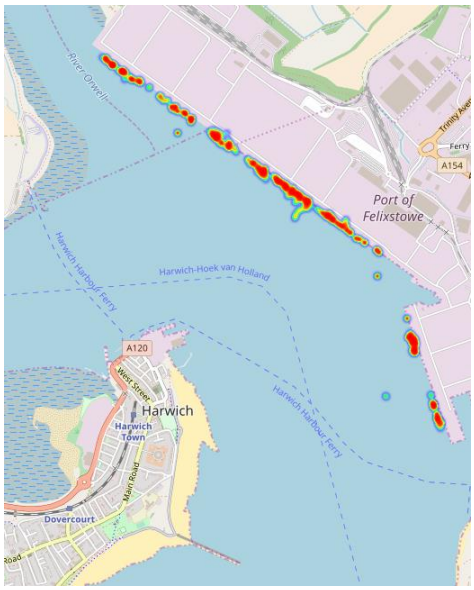


Figure 36: Docked segment (12% of training dataset)

In an attempt to further distinguish between ships entering and leaving port and dock, the segmentation was expanded to include a deceleration feature. This feature was pre-processed in exactly the same way as the SOG and ROT features. It was found that the acceleration feature dominated the segmentation and washed out the contribution of both the ROT and SOG features resulting in its removal for subsequent analysis.

Recall that the inputs to the segmentation algorithm are two state vectors consisting of a total of 12 values. The centroid of each of the six segments can be defined within 12-dimension space. An informative exercise is to visualise the six segment centroids in two-dimensional space so that the spatial separation of each segment may be explored. This is achieved by applying t-distributed Stochastic Neighbour Embedding (t-SNE)[9]. Figure 37 gives the result of applying t-SNE to the six segment centroids.

[9] t-SNE is a dimensionality reduction technique that is especially suited to reducing high-dimensional data into a space of two or three dimensions (van der Maaten et. al. 2008)

The resulting chart suggests that segments within both the transition and docking segments share the same localised space within the remapped axis. However, the interpretation of t-SNE results must be treated with caution as the remapped axis do not have any interpretable meaning.

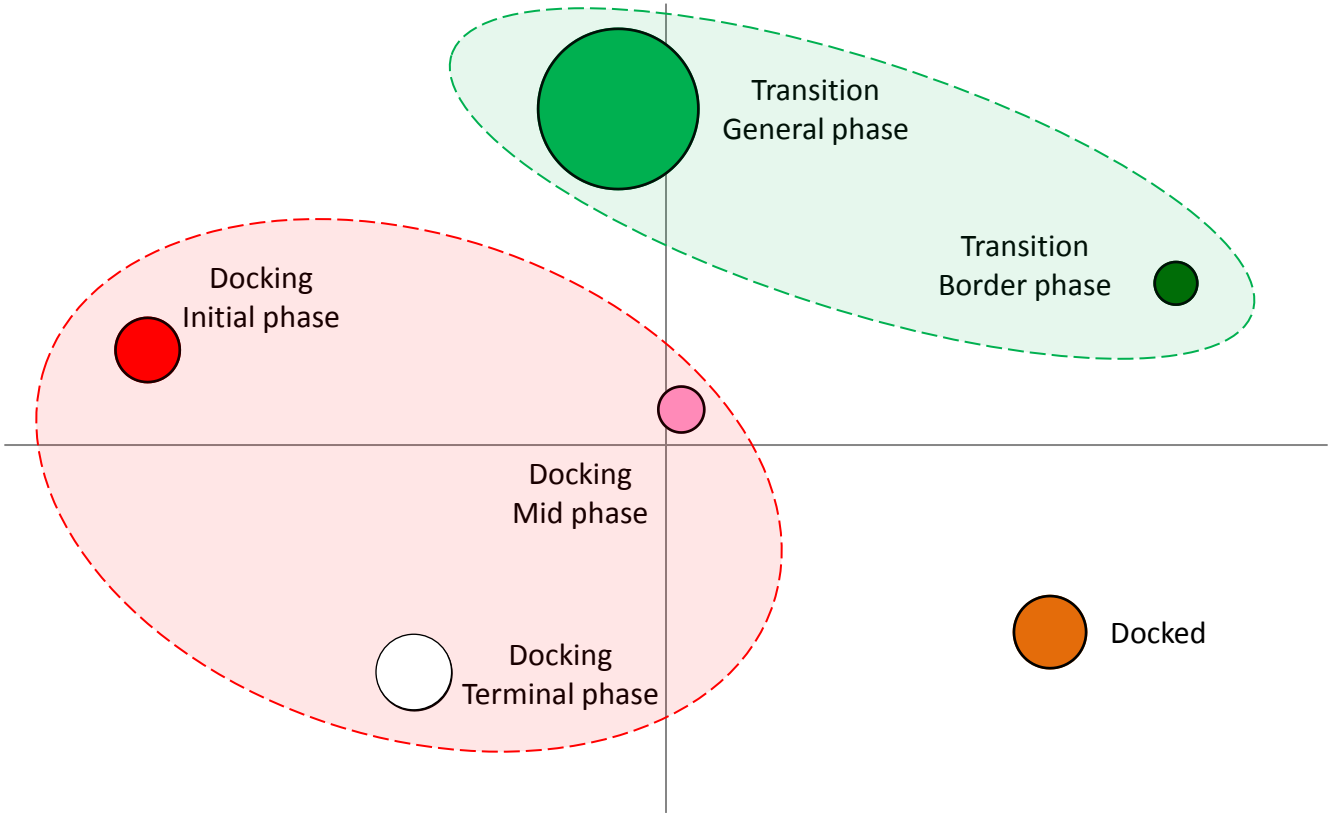


Figure 37: T-SNE plot of segment centroids mapped to two dimensions; size relates to the number of examples in each segment. As axes generated by t-SNE have no interpretable meaning, the results of these charts are indicative rather than definitive

T-SNE can also be used to illustrate the temporal order to the segments. With reference to the discussion of the six segments in an earlier section, it is reasonable to suggest that a ship entering port with the intention to dock will pass through each of the segments in a specific order. The ship will start by transitioning into the port and then pass into the general transition segment. It will then begin the docking procedure, which will consist of movement though each of the three docking modes, initial, mid phase and onto terminal phase (Figure 38). The ship will then move to the docked segment. The expected order through the segments is therefore:

- transitional border phase
- transitional general phase
- docking initial phase
- docking mid phase
- docking terminal phase
- docked

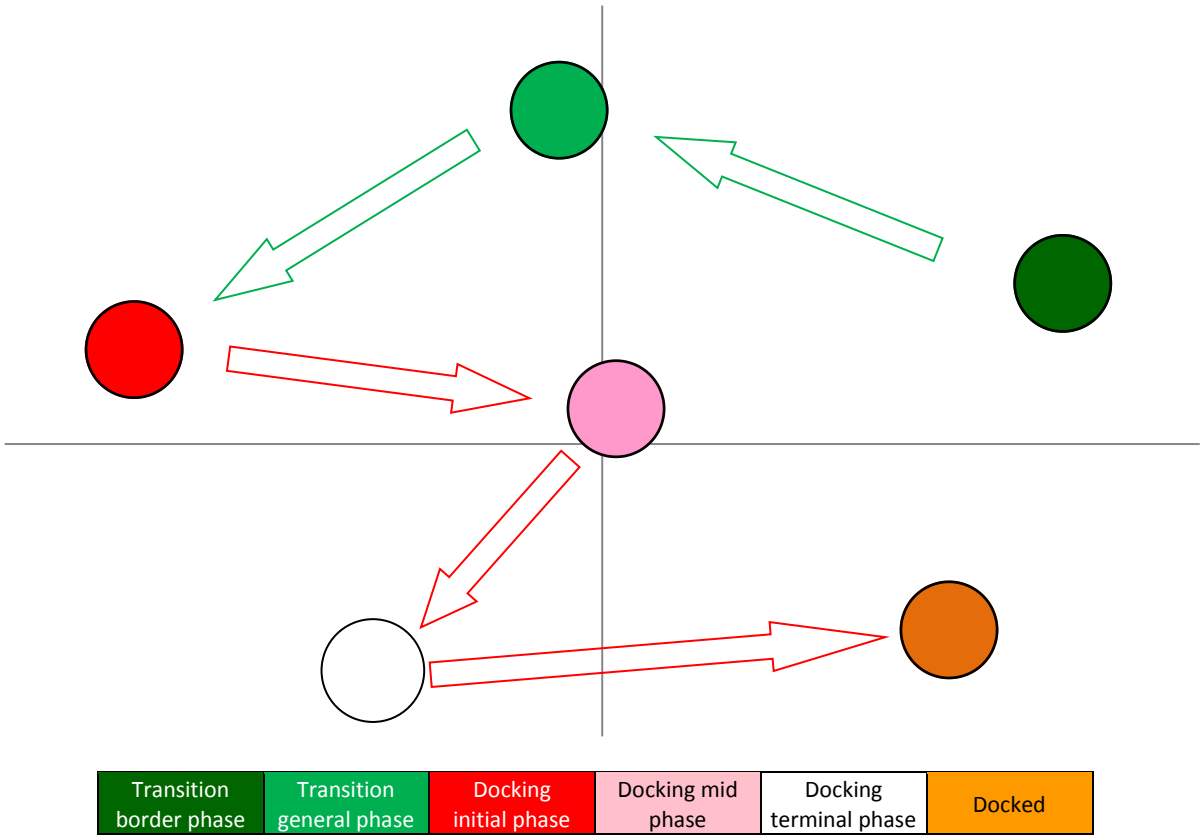


Figure 38: T-SNE plot of segment drivers mapped to two dimensions, showing the temporal progression through the segments for a ship entering port and going on to dock

4.1. Voyage classification

Once the centroids that define each of the six segments have been generated, it is a simple process to classify the behaviour of a ship at any point in its voyage by applying the segmentation to the preceding two minutes of AIS data.

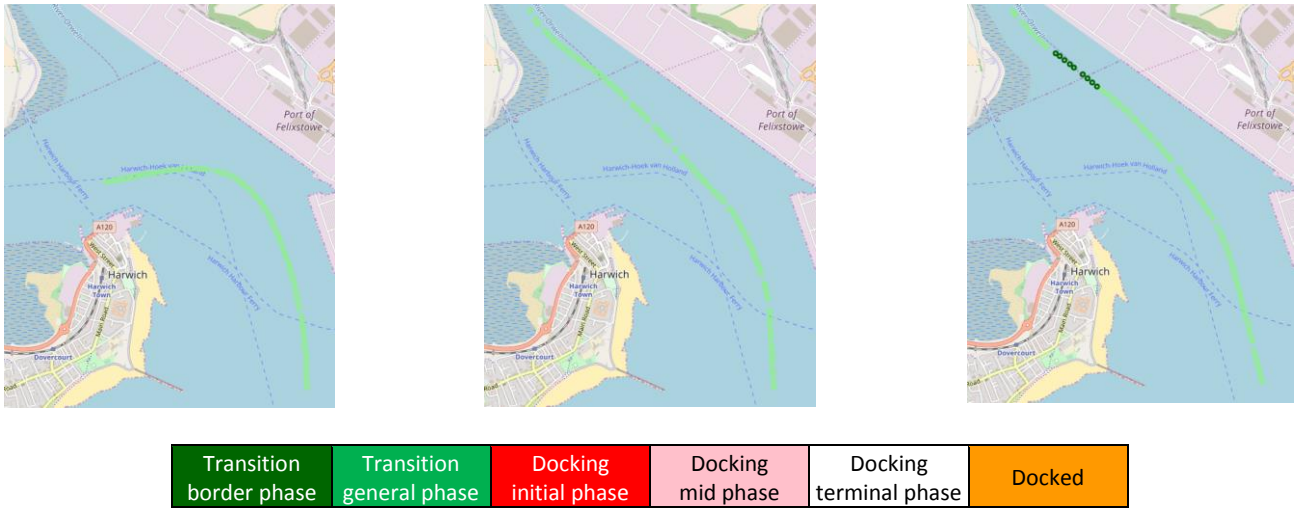


Figure 39: Behavioural classifications for ships passing through port, (from green, through amber to red represents an increasing density of AIS points)

Figure 39 gives the classified tracks of three ships that transition through Felixstowe. In the first two charts, both ships transition through the port with one turning east and the other heading north-west. In both cases the behaviour of the ship is classified as transitional general phase throughout the journey. In the third chart the ship again travels north west through the port; however, in this case there is a small period where the ship accelerates, moves into the transitional border phase before slowing down and returning to the transitional general phase.

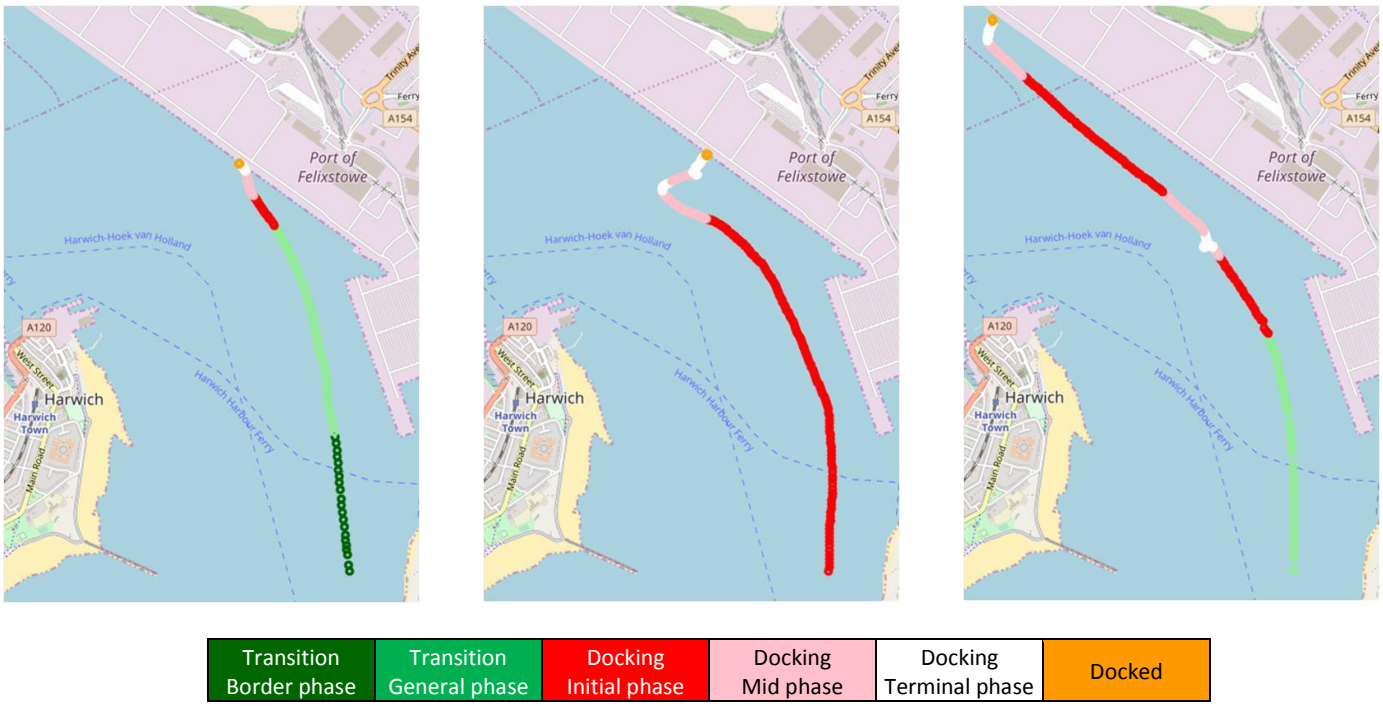


Figure 40: Behavioural classifications for docking ships, (from green, through amber to red represents an increasing density of AIS points)

More interesting results are observed with ships that dock within the port (see Figure 40). The ship in the left most chart follows the expected progression through the behavioural segments, namely transition, declaration through the docking phases before finally docking. The ship in the second chart is initially classified into the docking initial phase, which suggests that this ship enters port at a much lower speed and is slowing down. The ship then decelerates further, increases its rate of turn and moves through the docking mid and terminal phases. At this point, the ship accelerates and moves back into the docking mid phase segment before decelerating through the docking terminal phase and onto the docked segment. It is suggested that this behaviour is indicative of the ship manoeuvring into the prevailing current to aid the docking procedure. In the final chart the ship moves through the segments slowing down and increasing its manoeuvrability before entering the docking terminal phase segment. It then accelerates and enters the initial docking segment before decelerating, passing through the remaining docking segments and onto the docked segment. This behaviour is less clearly understood. However, it may be caused by the ship reducing speed to wait for an available berth before redirecting to a different berth.

5. Predicting delays

The unsupervised approach detailed above is used to classify the behaviour of a ship into a number of a behavioural states that change throughout the ship's voyage and specifically in and around port. The next area of investigation focused on predicting the likelihood that a ship would be delayed and arrive at its intended destination sometime after its estimated time of arrival. The objective of this study was not to develop an optimal model that could be used in a practical sense to predict delays, instead development followed a proof-of-concept approach, where the capability to predict delays was demonstrated and the various features of the prediction identified.

Unlike the previous example where ships were classified using an unsupervised approach (k-means), delays were predicted using supervised machine learning. A series of binary target fields were derived that indicated whether a ship was delayed or not. To do this, both the Automatic Identification System (AIS) and Consolidated European Reporting System (CERS) datasets were used and a means by which they could be joined was developed.

The AIS data provides a set of GPS locations covering the time window the AIS equipment was operational onboard the ship. However, these points are not organised into separate voyages. By joining the AIS and CERS datasets, one dataset was generated combining the information on voyages and arrival times plus the information which can be used to derive features relating to ship behaviour from their GPS locations.

The common identifier between the AIS and CERS data is Maritime Mobile Service Identity (MMSI, the unique identifier of each ship). To merge the two datasets together they were first inner joined on MMSI, the datasets were then filtered by restricting each timestamp within the AIS to its closest estimated time of arrival (ETA) or estimate time of departure (ETD) (both ETA and ETD were considered since the AIS data contained both the inbound and outbound portions of the journey). Records were retained where the ETA or ETD fell within 24 hours of the timestamp. As it is possible for more than one ETA or ETD to meet these criteria, the ETA or ETD closest to the timestamp was selected. After merging the complete dataset included 727 voyages relating to 235 unique ships.

Delays were calculated by subtracting the ETA from the Actual Time of Arrival (ATA), both these fields are contained within the CERS data. One feature of the data was that ETA may be updated throughout the journey of a ship its crew. Consequently, the last ETA is updated to match the ATA, giving the impression that the ship is not delayed. To overcome this the ETA from the message nearest to 24 hrs before the ATA was used. This aligns with the requirement that a CERS report must be made at least 24 hours in advance of arrival or departure. As this information is not available in the CERS data download, an automatic process was written to extract this information (with permission) from individual CERS messages.

As the threshold at which the length of delay becomes operationally critical differs for different situations, five binary target fields were created, each relating to different delay thresholds - these were 15, 30, 60, 90 and 120 minutes.

To explore all the factors that may contribute to delays, several features were taken and derived from both the AIS and CERS datasets. At each point on a ship's journey the following features were available:

Time and seasonality:

- minute of the hour
- hour of the day
- day of the week
- week of the year
- month of the year

Ship type:
- gross tonnage
- ship type
- hazmat cargo flag
- previous delay flag

Ship dynamics:
- SOG
- ROT
- acceleration
- distance from last port of call

Ship classification
- one of six segment classifications calculated at every AIS point during this voyage of the ship

Local loading
- distance to nearest ship
- number of ships within 10m, 50m, 100m, 500m, 1000m, 1000m+

Port separation
- distance between previous and intended ports

Port loading
- number of ships within port boundary
- number of ships within port boundary by ship type
- number of ships within port boundary by segment classification

Weather in port (average for day unless stated, taken from National Center for Environmental Information)
- temperature
- dew point
- sea level pressure
- station pressure
- visibility
- wind speed
- maximum wind speed
- maximum and minimum temperature
- fog, rain, drizzle, snow, ice indicator

To mitigate the effects of extreme (large or small) delay rates and to compare predictive performance across models predicting the five different delay definitions (15, 30, 60, 90 and 120 minutes), the development samples were balanced to contain equal numbers of delayed and non-delayed ships[10]. A 20% test dataset was randomly sampled and stratified on the target delayed field so that balanced outcomes were maintained within both training and test datasets. The test dataset was used to independently test the performance of each model. Training and test performance were compared to ensure overfitting was not present.

[10] It is noted that before any practical deployment of the model, the output from a classifier trained upon balanced outcomes should be adjusted to match expected delay rates

Several ensemble-based machine learning classifiers were investigated including Random Forest, AdaBoost and Gradient Boosted Trees[11]; the performance across all was broadly comparable. As a slight performance increase was obtained with an XGBoost algorithm[12] using the default hyperparameters (maximum tree depth of two, 100 independent estimators and a learning rate of 0.05), this approach was used as the preferred approach. To explore the relationship between model performance and delay time, an XGBoost model was trained on each of the target fields. Results are shown in Table 4

	Sample size	
Delay threshold	Training	Test
15 minutes	148,416	37,104
30 minutes	168,794	42,198
60 minutes	141,928	35,258
90 minutes	121,588	30,398
120 minutes	102,300	25,576

Table 4: Training and test dataset sizes. A delay threshold of 15 minutes indicates that a ship is deemed as being delayed if it arrives 15 minutes after its estimated arrival time

Figure 41 gives the results of training. It can be seen that with the exception of the 15-minute delay case, model performance increases with increasing delay threshold. This result is expected when one considers that larger thresholds are indicative of more extreme delays and therefore more discrimination should be evident between the features of delayed and non-delayed ships, making the predictive task a simpler one.

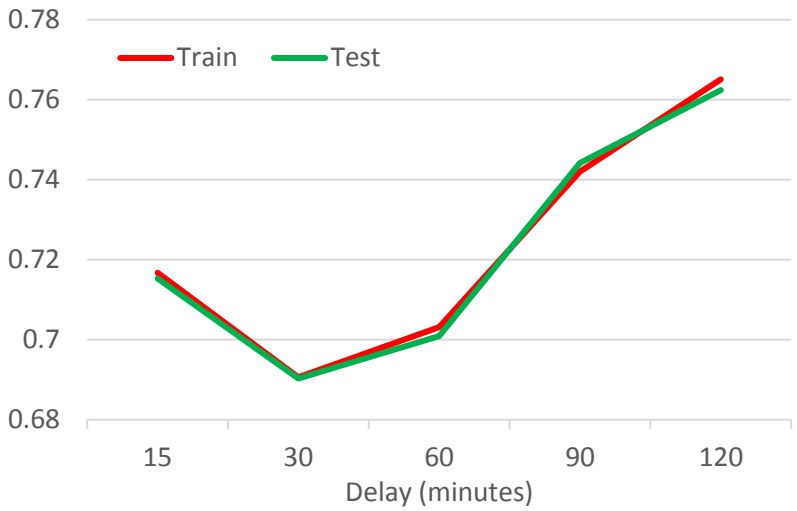


Figure 41: Test and training dataset accuracy when predicting delays of 15, 30, 60, 90 and 120 minutes. Accuracy is defined as the number of correct classifications divided by the total number of cases. Accuracy for a random classifier upon a balanced sample classifier is 0.5

[11] Random forests (Ho, 1995), AdaBoost (Freund, 1999) and Gradient boosted trees (Friedman, 1999) are all ensemble based supervised learning algorithms where a series of weak learnings are combined to form a single ensemble of weak classifiers. The ensemble of weak classifiers effectively decreases the variance of the model without increasing the bias.

[12] XGboost is a development of gradient boosting specifically developed to operate within a distributed parallel process environment (Chen et. al., 2016)

Focusing on the model predicting the most punitive definition of delay, a 15-minute threshold; precision was measured as 69% indicating that when the model predicts a delay it is correct in 69% of cases. Recall of 80% indicates that when a ship was delayed the model correctly predicts this delay in 80% of cases.

The standardised contribution of each feature in the model is shown in Table 5. Features selected during training fell into the weather, seasonality, port loading, local loading and ship type groups. The most powerful feature in the model was "distance between previous and intended port" with a standardised importance of 16%, whilst the weakest feature "day of the week" had an importance of under 1%.

Feature	Contribution
Distance between previous and intended port	0.164
Hour of the day	0.160
Maximum daily temperature	0.137
Sea level pressure	0.127
Gross tonnage	0.100
Visibility	0.040
Hazardous cargo on board	0.040
Wind speed	0.030
Minimum daily temperature	0.027
Port loading (tugs)	0.017
Port loading (cargo ships)	0.010
Distance to nearest ship	0.007
Day of the week	0.003

Table 5: Standardised feature importance by feature

Surprisingly, ship-dynamics did not feature in the final model. This may be a result of the fact that AIS data was limited to the area in and around port (the furthest AIS point being 2.5km from the port datum). Consequently, behavioural changes that are indicative of a delay (such as a positive acceleration or a higher than expected speed) are not captured within the AIS data used to train the model.

Additional insights can be uncovered by exploring the relationship between the target field expressed as delay rate and each feature on a univariate basis. This approach does not account for the combinatorial effect of features that will be captured by the machine learning approach. For example, the weather in port is of less importance for a ship that is sailing several thousands of kilometres from the port in question and only increases in importance when the ship approaches port. Consequently, weather becomes more predictive when considered on a multivariate basis and in combination with the "distance to port" and "current speed" features.

Figure 42 gives the results of such an analysis for a selection of noteworthy model features. Highlights include:

- ship delays are less likely to occur in the early hours of the morning and during the late afternoon and evening
- there is an increased likelihood of a delay occurring at the start of the working day and during the lunchtime window
- delays are generally more likely to occur in poorer weather conditions; this is clearly shown in both the temperature and visibility charts and specifically for temperatures below 12 degrees and visibility of less than 13 miles
- seasonality is also indicative of delays with ships arriving in the months of August, September and October less likely to suffer delays, with ships arriving in February, March, April and May more likely, it is suggested that seasonality may be related to weather
- if a ship is not carrying hazardous cargo it is less likely to be delayed

Finally, the likelihood of being delayed increases linearly with the number of tugs operating within the harbour. It is suggested that as many larger ships require tugs this feature is indicative of the number of ships manoeuvring into berth and therefore the availability or otherwise of suitable berths. It is likely that this feature is related to the segment based on port loading features, specifically those relating to the docked and docking segments.

Figure 42: Univariate analysis for a selection of notable model features giving delay rates relative to background balanced delay rate (0.5)

The proof-of-concept model discussed above has demonstrated that the development of a more powerful model predicting delays is a feasible and achievable objective. It is suggested that if additional data covering all major UK ports and the surrounding waters were extracted, more powerful machine learning approaches such as deep learning could be applied. These approaches would more accurately capture nonlinearities within the data and further increase model performance. However, there is concern regarding the accuracy and robustness of the delays flags used as a target within the training data (derived from the ETA field within the CERS dataset), consequently this should be investigated and if necessary addressed as a priority in any future work. In addition, more emphasis should be placed upon the development and investigation of a wider and more predictive suite of features, including:

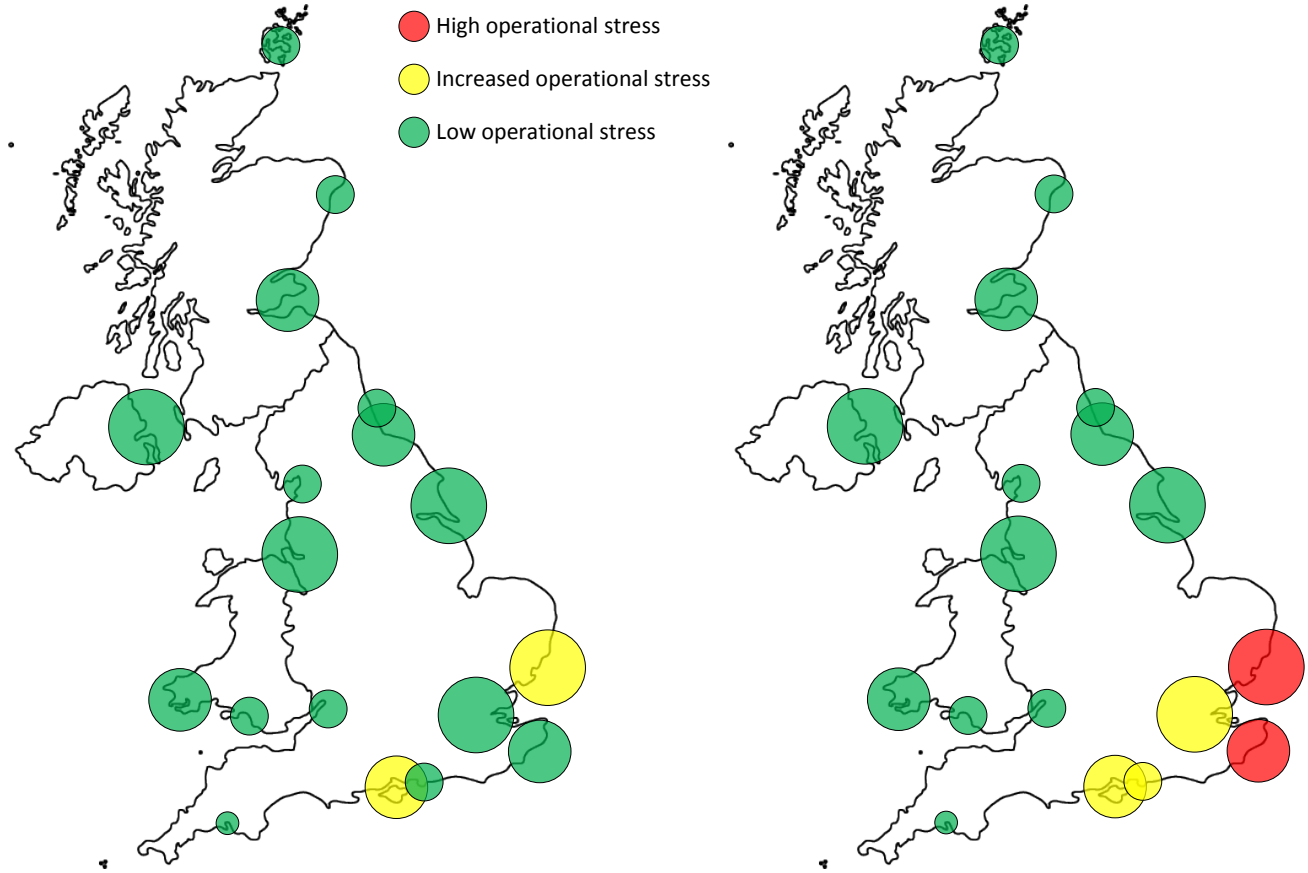
- weather data at a more granular resolution (at least hourly)
- AIS data covering a larger geographic area (open sea and not just the area immediately surrounding each port)
- AIS data taken from satellites that fills base-station coverage gaps
- deviation from expected behavioural footprints
- changes in ship behaviour taken over time
- dynamic interactions between ships
- cargo pricing (most notably for liquid bulk such as LNG, crude oil and other oil products)
- port opening data (staffing, port capacity and so on)
- breaking down hazardous materials by the type of material
- tidal information

6. Future work

Once model performance has been optimised, enabling the model to be used to predict delays in a practical sense, there are several areas where it could be applied to understand port characteristics, operation and utilisation. Some potential areas of interest are discussed below.

Early indicators of GDP

Gross Domestic Product (GDP) is a measure of the value of all goods and services produced by a country in a given period. It is essentially a measure of the economic performance of a country. GDP figures are calculated and released on a quarterly basis resulting in a degree of latency between releases. The Consolidated European Reporting System and Consolidated European Reporting System (CERS) data could be used to explore, understand and capture these relationships between GDP and freight transport volumes. Supervised machine learning techniques could then be applied to produce early indicators of GDP and support GDP based decision-making in the period between formal quarterly releases. More work needs to be undertaken to understand what is being carried onboard ships as this information would add significant weight to this analysis.



Time *t*
Felixstowe and Southampton are subject of medium levels of operational stress

Time *t + 1*.
Increased operational stress in the south west indicates demand is likely to outstrip capacity in the near future

Figure 43: Hypothetical plot showing a simulated scenario of operational stress at time *t* and *t + 1*
Size represents predicted volume of freight at port
Note: These charts are illustrative and are not intended to illustrate actual operational characteristics

Port loading simulation

Consider a snapshot of ship position and dynamics as defined by AIS data, in and around the UK ports at a given point in time. It follows that this snapshot effectively defines the shipping state across UK waters. A delays model could be used to predict the likelihood of each ship arriving late into its intended port. This data could be combined with ETA and aggregated to port level to give an indication of port loading into the future. Information relating to each port (staffing levels, capacity, operational efficiency and so on) could then be combined with port loading to give a measure of operational stress[13] at each port at that point in time. Standard simulation algorithms could then be applied to determine port loading into the future (see Figure 43). The next step would be to explore and simulate the knock-on impact of increased port loading upon local infrastructure such as the road network, provision of goods and environmental impacts.

Scenario planning

Simulation can be used to estimate and understand what will happen in the near future. Scenario planning uses "what-if" analysis to explore and understand what could happen if the dynamics that define the operating state of the system are changed or modified. For example, a stress analysis would explore the effects upon port loading, road utilisation, supply of goods and environmental impact under the following scenarios:

- port loading and unloading capacity is reduced or increased
- port staffing volumes are increased, for example, changed to favour weekends or late evenings
- resources are taken from one port and allocated to another
- ships are rerouted to different ports
- ships carrying hazardous materials are given priority in port
- oil prices change or are subjected to large daily fluctuations
- export and import levels change due to Brexit
- temporary closures of ports due to unforeseen factors (industrial action, accidents, terrorism)
- impact of global warming and the subsequent increases in sea levels and changes in tidal patterns

Scenario planning could be used to support the decision-making process by being incorporated into an online tool that would allow the user to quickly explore user-defined scenarios (including those detailed above) in a near real-time and zero-risk environment. Various shipping states taken either historically, in real time, or estimated into the future could be generated and the impact explored and quantified in each case.

Port operation optimisation

A further area of interest lies with the application of search-based heuristics such as genetic algorithms. These could be used to optimise a defined set of parameters that define the design or operation of a port so as to minimise or maximise one or more operational parameters of that port. Examples may include:

- minimise time at berth
- minimise the number of delayed ships
- maximise number of ships processed in a day
- minimise total time ships wait for a dock in a given time window
- minimise bottlenecks at certain times of the day
- minimise environmental impact
- minimise road loading during rush hour
- minimise operations stress across all ports
- maximise utilisation of road network on weekends

[13] Operational stress is defined as the difference between demand and capacity

To ensure that each optimised solution remains feasible from an operational perspective, optimisation should be constrained so as not to violate any one of many business based rules. Contextual examples may include:

- no ships may dock at certain times of the day
- no more than a maximum number of ships can be docked at any time of the day
- demand cannot exceed available resource at any point
- a minimum number of berths must be occupied throughout the day
- loading must be broadly uniform across all berths within the port

As with simulation and scenario planning, optimisation could be deployed within an integrated tool, further augmenting the power of the manual decision support and operational domain exploration.

7. Potential applications outside the maritime industry

The work detailed in this report has been developed with the maritime industry, specifically freight ships and tankers, in mind. However, the tools and techniques discussed here may easily be applied to a broader set of applications that are based upon the movement and relationship between individual entities. These may include:

Aircraft (military and civil) - Understand the movements of aircraft in and around airports and within designated airways.

Lorries / haulage - Explore the behaviour and environmental impact of traffic on the UK road network.

Crowd dynamics. - Simulate the response of crowds to a number of scenarios in and around open spaces, buildings and arenas. Optimise building design to minimise the risk and impact of potentially dangerous situations (for instance crowd density during emergency evacuations).

Movement of money - Classify the movement of money within, into and out of the UK, identify and predict criminal behaviour patterns. Quantify the financial impact of Brexit.

Internet traffic - Understand the movement of information around the internet. Predict future load patterns and data bottlenecks. Optimise server configuration to satisfy future demand.

Professional sports - Analyse the behaviour of players during games, identify successful tactical responses to opposition. Understand the relationship between individual behaviour and team position. Predict the likelihood of injury and identify early indicators of impending injury.

8. Summary

This report has outlined the work undertaken by the Data Science Campus to explore the operation, utilisation and relationships between ports in the UK at a macro level and the behaviour and operational characteristics of ships at a micro level.

Two data sources have been investigated, the Consolidated European Reporting System (CERS) and the Automatic Identification System (AIS). A high-level analysis of the CERS data identified several pieces of notable insight. There are a handful of ports within the UK that are more likely to either load or unload hazardous materials when compared with UK ports in general. Other ports have clear and specific links with other ports, for instance ships leaving Belfast generally travel to other UK destinations whereas ships leaving Felixstowe are far more likely to travel to ports in the EU. It has been shown that AIS can be extracted, decoded and stored within a Hadoop Distributed File System (HDFS) environment. This data can then be processed and the behaviour of individual ships visualised by overlaying on charts.

The rate of turn variable in the AIS data was found to be erroneous and unsuitable for use; this necessitated the development of a replacement variable, which appears to have worked well. In addition, the AIS data was found to contain a degree of noise; this sensitivity was reduced by aggregating the speed and manoeuvrability of a ship over a window of two minutes and then converting to a state vector. This approach significantly improved the robustness of the segmentation based classifier. When the classifier was applied to a selection of ships on their journey into port, the journey was decomposed into a series of distinct, unique and robust phases.

Several machine learning approaches were developed to predict the likelihood a ship would be delayed in arriving at port. XGBoost was found to give marginally better performance and was used as the preferred proof-of-concept algorithm. Comparable performance within both the training and development samples suggests the algorithm trained well and overfitting was not present to any significant degree. A univariate analysis of the more notable model features gave additional insight into the relationship between weather, seasonality, port loading, ship behaviour and delays. The delays model demonstrated that the development of a more powerful model predicting delays is a feasible and achievable objective. It is suggested that if additional data covering all major UK ports and the surrounding waters were extracted, more powerful machine learning approaches such as deep learning could be applied. These approaches would more accurately capture nonlinearities within the data and further increase model performance.

Once model performance has been optimised, there are several areas where the model may be deployed: port loading simulation, scenario planning and port operation optimisation.

Although the work detailed in this report has been developed with the maritime industry in mind, it is suggested that the tools and techniques identified and developed here may easily be applied to a far broader set of applications that are based upon the movement and relationship between individual entities and covering areas as diverse as crowd dynamics, internet traffic and the movement of financial funds.

Code repository

All relevant code relating to this project is available in the Github repository.

Acknowledgements

The Data Science Campus wishes to thank the staff of both the Maritime and Coastguard Agency and the Centre for Big Data Statistics at Statistics Netherlands for their support and assistance through this course of this project.

The Data Science Campus at the ONS

The Data Science Campus applies data science, and builds skills, for public good across the UK and internationally. We work at the frontier of data science and Artificial Intelligence (AI) – building skills and applying tools, methods and practices – to create new understanding and improve decision-making for public good.

The goals of ONS's Data Science Campus are to investigate the use of new data sources, including administrative data and big data for public good and to help build data science capability for the benefit of the UK. A new generation of tools and technologies is being used to exploit the growth and availability of these new data sources and innovative methods to provide rich informed measurement and analyses on the economy, the global environment and wider society.

The Data Science Campus was established within the Office for National Statistics (ONS) in 2017 with a core of well-qualified professionals, involving a strong network of third party participants in the mission of the Campus. We have set up a series of data projects that provide insight into key policy themes. We created new learning and development pathways in data science at a range of different levels from Level 4 Apprenticeships to providing support for PhDs and post-doctoral projects. We are located at ONS's Newport site in South Wales and have smaller unit in ONS's London and Titchfield office.

More information can be found at the Data Science Campus website.

References

Bartelmaos S, Abed-Meraim K, and Attallah S (2005), 'Fast algorithms for minor component analysis', in Statistical Signal Processing, IEEE/SP 13th Workshop on. IEEE, 2005, pages 239 to 244

Best R A and Norton J (1997), 'A new model and efficient tracker for a target with curvilinear motion', Aerospace and Electronic Systems, IEEE Transactions on, volume 33, number 3, pages 1030 to 1037

Blaich M, Rosenfelder M, Schuster M, Bittel O, and Reuter J, 'Fast grid based collision avoidance for vessels using a search algorithm', in Methods and Models in Automation and Robotics (MMAR), 2012 17th International Conference on. IEEE, 2012, pages 385 to 390

Bomberger N, Rhodes B J, Seibert M, Waxman A M, 'Associative learning of vessel motion patterns for maritime situation awareness', in Information Fusion, 2006 9th International Conference on. IEEE, 2006, pages 1 to 8

Cabrera F, Molina N, Tichavska M, Arana V, 'Design of a low-cost prototype of automatic identification system (AIS) receiver', in: 2015 1st URSI Atl. Radio Sci. Conf. (URSI AT-RASC), IEEE, 2015: page 1

Fagerholt K, Heimdal S, and Loktu A (2000), 'Shortest path in the presence of obstacles: An application to ocean shipping', Journal of the operational research society, pages 683 to 688

Grewal M S (2011), 'Kalman filtering', Springer

Guo X-R, Wang F-H, Du D-F, and Guo X-l, 'An improved neural network based fuzzy self-adaptive Kalman filter and its application in cone picking robot', in Machine Learning and Cybernetics, 2009 International Conference on, volume 1, IEEE, 2009, pages 573 to 577

Hamilton J D (1994), 'Time series analysis', Princeton university press Princeton, volume 2

Handayani D O D, Sediono W, Shah A, 'Anomaly detection in vessel tracking using support vector machines (svms)', in Advanced Computer Science Applications and Technologies (ACSAT), 2013 International Conference on. IEEE, 2013, pages 213 to 217

Hart P E, Nilsson N J, and Raphael B (1968), 'A formal basis for the heuristic determination of minimum cost paths', Systems Science and Cybernetics, IEEE Transactions on, volume 4, number 2, pages 100 to 107

Haykin S (2004), 'Neural Networks: A comprehensive foundation', Neural Networks, volume 2, number 2004

Holst A and Ekman J (2003), 'Anomaly detection in vessel motion, internal report Saab Systems', Sweden, Tech. Rep.

Hornauer S, Hahn A, Blaich M, and Reuter J (2015), 'Trajectory planning with negotiation for maritime collision avoidance', TransNav: International Journal on Marine Navigation and Safety of Sea Transportation, volume 9

Hu J, Zhang L (2011), 'Research and realization of AIS ship dynamic monitoring system based on SOA', International Conference on Electric Information and Control Engineering, pages 4370 to 4373

Johansson F and Falkman G, 'Detection of vessel anomalies – a Bayesian network approach', in Intelligent Sensors, Sensor Networks and Information, 2007. ISS-NIP 2007. 3rd International Conference on. IEEE, 2007, pages 395 to 400

Khan A, Bil C, and Marion K E, 'Ship motion prediction for launch and recovery of air vehicles', in OCEANS, 2005. Proceedings of MTS/IEEE. IEEE, 2005, pages 2795 to 2801

Kraiman J B, Arouh S L, Webb M L (2002), 'Automated anomaly detection processor', in AeroSense, International Society for Optics and Photonics, 217, pages 128 to 137

Laxhammar R (2008), 'Anomaly detection for sea surveillance', in Information Fusion, 2008 11th International Conference on. IEEE, pages 1 to 8

Lazarowska A (2014), 'Safe ship control method with the use of ant colony optimization', in Solid State Phenomena, volume 210. Trans. Tech. Publ., pages 234 to 244

Li X R and Jilkov V P (2003), 'Survey of manoeuvring target tracking. part i. dynamic models, 'Aerospace and Electronic Systems', IEEE Transactions on, volume 39, number 4, pages 1333 to 1364

Lisowski J (2001), 'Determining the optimal ship trajectory in collision situation', in Proceedings of the IX International Scientific and Technical Conference on Marine Traffic Engineering, Szczecin

Osekowska E, Axelsson S. and Carlsson B (2013), 'Potential fields in maritime anomaly detection', in Proceedings of the 3rd International Conference on Models and Technologies for Intelligent Transportation Systems

Peng D and Yi Z (2006), 'A new algorithm for sequential minor component analysis', Int. J. Comput. Intell. Res, volume 2, number 2, pages 207 to 215

Pershitz R (1973), 'Ships manoeuvrability and control', Leningrad, USSR: Sudostroenie

Rasmussen C E (2006), Gaussian processes for machine learning

Rhodes B J, Bomberger N, Seibert M, Waxman A M, 'Maritime situation monitoring and awareness using learning mechanisms', in Military Communications Conference, 2005. MILCOM 2005. IEEE, 2005, pages 646 to 652

Stateczny A and Kazimierski W (2011), 'Multisensor tracking of marine targets-decentralized fusion of Kalman and neural filters', International Journal of Electronics and Telecommunications, volume 57, number 1, pages 65 to 70

Szlapczynski R and Szlapczynska J (2012), 'On evolutionary computing in multi-ship trajectory planning', Applied Intelligence, volume 37, number 2, pages 155 to 174

Szlapczynski R (2006), 'A new method of ship routing on raster grids, with turn penalties and collision avoidance', Journal of Navigation, volume 59, number 01, pages 27 to 42

Szlapczynski R (2013), 'Evolutionary sets of safe ship trajectories within traffic separation schemes', Journal of Navigation, volume 66, number 01, pages 65 to 81

Szlapczynski R (2011), 'Evolutionary sets of safe ship trajectories: a new approach to collision avoidance', Journal of Navigation, volume 64, number 01, pages 169 to 181

Tam C and Bucknall R (2010), 'Path-planning algorithm for ships in close-range encounters', Journal of marine science and technology, volume 15, number 4, pages 395 to 407

Tsou M-C and Hsueh C-K (2010), 'The study of ship collision avoidance route planning by ant colony algorithm, Journal of Marine Science and Technology', volume 18, number 5, pages 746 to 756

Tsou M-C, Kao S-L, and Su C-M (2010), 'Decision support from genetic algorithms for ship collision avoidance route planning and alerts', Journal of Navigation, volume 63, number 01, pages 167 to 182

Tu E, Zhang G, Rachmawati L, Rajabally E, Huang G (2016), 'Exploiting AIS Data for Intelligent Maritime Navigation: A Comprehensive Survey', IEEE Transactions on Intelligent Transportation Systems, pages 10 to 1109

Ying S, Shi C, and Yang S (2007), 'Ship route designing for collision avoidance based on Bayesian genetic algorithm', in Control and Automation. ICCA 2007. IEEE International Conference on. IEEE, 2007, pages 1807 to 1811

Zhou B and Shi A, 'Lssvm and hybrid particle swarm optimization for ship motion prediction', in Intelligent Control and Information Processing (ICICIP), 2010 International Conference on. IEEE, 2010, pages 183 to 186