Fair automated assessment of noncompliance in cargo ship networks

Bruin, G.J. de; Pereira Barata, A.; Herik, H.J. van den; Takes, F.W.; Veenman, C.J.

de Bruin et al. *EPJ Data Science* (2022) 11:13
https://doi.org/10.1140/epjds/s13688-022-00326-w

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Fair automated assessment of noncompliance in cargo ship networks

Gerrit Jan de Bruin[1,2,3]*, Antonio Pereira Barata[1,2], H. Jaap van den Herik[3], Frank W. Takes[1] and Cor J. Veenman[1,4]

*Correspondence:
g.j.de.bruin@liacs.leidenuniv.nl
[1] Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333 CA, Leiden, the Netherlands
[2] Human Environment and Transport Inspectorate, Rijnstraat, 2515 XP, the Hague, the Netherlands
Full list of author information is available at the end of the article

Abstract

Cargo ships navigating global waters are required to be sufficiently safe and compliant with international treaties. Governmental inspectorates currently assess in a rule-based manner whether a ship is potentially noncompliant and thus needs inspection. One of the dominant ship characteristics in this assessment is the 'colour' of the flag a ship is flying, where countries with a positive reputation have a so-called 'white flag'. The colour of a flag may disproportionately influence the inspector, causing more frequent and stricter inspections of ships flying a non-white flag, resulting in confirmation bias in historical inspection data.

In this paper, we propose an automated approach for the assessment of ship noncompliance, realising two important contributions. First, we reduce confirmation bias by using fair classifiers that decorrelate the flag from the risk classification returned by the model. Second, we extract mobility patterns from a cargo ship network, allowing us to derive meaningful features for ship classification. Crucially, these features model the behaviour of a ship, rather than its static properties. Our approach shows both a higher overall prediction performance and improved fairness with respect to the flag. Ultimately, this work enables inspectorates to better target noncompliant ships, thereby improving overall maritime safety and environmental protection.

Keywords: Data-driven inspection; Fair classification; Mobility patterns; Port state control; Cargo ship network; Ship risk profile; Preventing confirmation bias

1 Introduction

Maritime cargo transport is of major importance to global trade, often being the most cost-effective way to move goods from one place to another. It results in many ship movements across the world; around 80% of world merchandise is carried by sea [1]. However, maritime transport also comes with risks, such as labour exploitation, culpable ships accidents, and environmental pollution. Obviously, these risks need to be mitigated by shipowners. To ensure mutual trust that ships from all countries are adhering to international laws, port state control inspections are conducted when ships berth a port. There are two possible outcomes of an inspection; either the ship is found compliant, or there are particular deficiencies. These port state control inspections check for compliance with many regulations, including any deficiency that could lead to one of the aforementioned maritime risks. If severe enough, such deficiencies can lead to a detention, meaning that the ship is not allowed to depart the port before the deficiencies are rectified, or to a ban meaning that the ship is not allowed to enter certain ports any longer. In this work we aim to predict in an automated manner whether a ship will have a deficiency in port state control, and thus is potentially noncompliant, which we consider equivalent to a ship posing a high risk.

In recent years, governments have established more strict laws to mitigate the negative consequences of maritime transport. Members of the Paris Memorandum of Understanding[1] introduced a so-called 'new inspection regime' [2]. Arguably the biggest innovation in the renewed memorandum, is the introduction of a ship risk profile. It awards a score to each ship based on a weighted sum of six factors [2]. The factors used in the risk profile for a given ship are its: (1) type, (2) age, (3) commercially issued safety certificate, (4) owning company's performance, (5) historical misconducts and (6) the flag a ship is flying, or equivalently, the country of registration [3]. Ships with a low risk profile should be inspected every three years, while ships with a high risk profile should be inspected every six months. The new inspection regime, with its ship risk profile, allows inspectorates to focus on noncompliant ships. It also leads to efficient use of the inspection capacity and budget, as every unnecessary port state control inspection costs the inspectorates on average around $1,000 [4]. This money can then be allocated to find more noncompliant ships. In [5] it was estimated that a noncompliant ship saves on average around $400,000 on maintenance by not complying to regulations, whereas the loss of a ship can incur costs up to $67,000,000. Shipowners with a low risk profile can benefit by having a reduced inspection burden, saving precious turn-around time in the port.

From the six factors used in the current ship risk profile, the flag plays an important role [6, 7]. The flag is considered black, grey, or white, based on the detention ratio of the country over a three-year rolling period [8]. Fleets from countries on the black list were significantly more often detained over a three-year rolling period, compared to fleets from countries on the whitelist. We mention three drawbacks in considering the flag for the ship risk profile. First, there are ethical concerns. The use of the flag can be considered disparate treatment [9], because ships are intentionally treated differently based on membership of a privileged class, being the white flag. Second, there are opportunities for ships to change flags, opening up the possibility for noncompliant ships to 'hide' under a white flag [10]. Although changing flags does not necessarily improve compliance, the current new inspection regime would grant such a ship a lower risk profile. In an ideal situation, merely changing an administrative property of a ship should not change the assessment of the risk associated with that ship. Third, inspectors can use their own discretion (possibly leading to subjectivity) to decide how thorough an inspection is. Hence, ships flying a black flag could possibly be subjected to stricter inspections, resulting in a higher probability of finding a noncompliant issue [11, 12]. This potential greater focus on ships flying a black flag may mean that these ships are inspected more often and stricter, contributing to a confirmation bias in historical inspection data [10]. The potential danger of inspec-

[1] The following countries are part of the Paris Memorandum of Understanding: all European coastal countries, as well as Canada, Norway, Russia, and the United Kingdom.

tors' bias has been recognised and great efforts are made to harmonise the training of inspectors, thereby making the overall inspectorate system consistent [3]. Nevertheless, complete global harmonisation is not achieved yet [12].

In our study, we could choose to start ignoring the flag of a ship altogether to reduce the aforementioned confirmation bias, thus providing what in the literature [13] is known as *equal opportunity*. However, correlations between the other characteristics of a ship and its target exist, thus the classifier will indirectly learn to use the flag of a ship, resulting in *inequality of outcomes*. Considering all drawbacks of using the flag in risk prediction, we argue that it might be better to get equal outcomes and therefore investigate how we can decorrelate the flag with respect to the outcome of the automated prediction of noncompliance. We do so by employing a so-called fair classifier [14], that can classify whether a ship is noncompliant but prevents (to a specified extent) correlation between its output and the ship's flag. Such a fair classifier may reduce the confirmation bias and therefore improve overall fairness of the risk assessment, compared to the aforementioned ship risk profile, which we consider to be the baseline model.

We consider the actual behaviour of the ships for prediction of noncompliance, in contrast to using the aforementioned six factors of the ship risk profile. Ship behaviour has been used to find anomalous ships [15], which may be indicative of noncompliance. An example of behaviour of a ship that might be characteristic for noncompliance, is that a ship is sailing primarily on routes with a lot of competition, potentially leading to the cutting of costs at the expense of safety. While we do not know the fares on specific routes, our proposed classifier will still take relations between noncompliance and the sailed routes into account. In the current study, we derive a cargo ship network from data containing notifications of ships calling to a port. This data is available to all inspectorates that are member of the European Maritime Safety Agency.

In the network nodes are ports, and edges are ships that travel between ports. By considering the structural function each port has in the network, we extract mobility patterns for each ship. These mobility patterns are provided to the fair machine learning classifier, enabling automated assessment of the risk of ships based on their behaviour. The use of these mobility patterns is novel, since data on port calls has only recently become available to inspectorates.

Hence, our goal is to devise an automated and accurate assessment of ship noncompliance. We do so by answering two research questions. First, how can we obtain our goal using *behavioural* data? Second, how can we obtain our goal in a *fair* manner?

The structure of the remainder of this paper is as follows. We start with related work on the ship risk profile and ship risk classification in general. Then, we explain the data used in this work. Subsequently, we describe our methods used to answer the research questions. We then present the results, and end with a discussion and conclusions.

2 Related work

It is widely recognised that the introduction of the 'new inspection regime', and thereby the ship risk profile, has been beneficial to a reduction of the number of noncompliant ships [7, 12, 16, 17]. We remark that some weaknesses of the current ship risk profile have already been identified in literature [18–24]. We mention two of them, together with the solutions that were provided. We then continue with discussing related work on the cargo ship network.

The first weakness in the existing ship risk profile, which assesses risks based on a weighted sum of six characteristic ship factors, is that the weights are manually determined [25]. In doing so, the model ignores any interactions between the factors. Here we remark that more complex models may take into account more dependencies and correlations, thereby improving performance [25, 26]. To this end, machine learning classifiers have been introduced that can learn the weights automatically and do capture correlations between the factors. Gao *et al.* proposed to use a support vector machine and k-nearest neighbours pipeline to find high risk ships [25]. The support vector machine takes more complex (and non-linear) interactions into account and generalises well, while k-nearest neighbours makes the overall approach noise tolerant. Yan *et al.* acknowledged that only a small fraction of ships is detained, and therefore used a balanced random forest classifier to predict ship detentions [24].

The second weakness of the ship risk profile is that so far relatively static factors are used in the assessment of risk, meaning that the factors rarely change for a given ship. To remedy this weakness, datasets have been exploited that better reflect the current condition of a ship and hence may improve prediction. Xu *et al.* used web scraping techniques to gather more information from inspection reports [20]. Knapp and Franses use company inspections and data from other inspection regimes to enhance the ship risk profile [27]. Yan *et al.* proposed to add more historic information to the model, such as times of changing flag and casualties in the last five years [24]. Also, they suggested to make information exchange between different inspection regimes more coherent, such that deficiencies and detentions in other regions can be used as well [26]. An additional suggestion is to enrich the risk profile by incorporating more specific information, such as data pertaining whether the ship has been involved in an accident.

In this work we use port call data modelled as a cargo ship network. The first publication of such (global) cargo ship network was by Kaluza *et al.* [28]. They noted that the diameter of the network, the longest shortest path length, was smaller than expected for a random network with the same number of nodes and edges, with a value of only 8. Also, they found that the average topological distance between any two ports in the world was only 2.5. Likewise, Liu, Wang and Zhang found a diameter of only 7 and an average distance of 3.3 [29]. Peng *et al.* studied the robustness of the cargo ship network based on transponder data available in 2018 [30]. They differentiated between the different ship types (oil tankers, container, dry bulk) and reported the properties for each of the sub-networks derived for just those ships. No measure of the distances in the network was reported, but a density (of $\sim$0.02) similar to the first published cargo ship network was found. Finally, Van Veen analysed the cargo ship network as derived from data of port calls [31]. Although the data involved only journeys either departing or arriving at one of the members of the Paris MoU, a diameter of 7 was found and an average distance of 2.49, which is similar to the reported values of other works. In the data section, we compare the properties of these networks to those of the cargo ship network as derived by us. Ultimately, we predict noncompliance using a classifier supplied with mobility patterns extracted from the cargo ship network. Our approach thus addresses the two weaknesses observed in the ship risk profile that is currently used by inspectorates.

3 Data

The purpose of the paper is to classify in a fair manner the noncompliance of ships, using behavioural data. The data used in the paper stems from two sources; (1) port calls and (2) inspections.

The first data source, being the port calls, contains notifications of all cargo ships calling to a port. Our port call data contains only calls to a port participating in the Paris Memorandum of Understanding and is accompanied with the following five pieces of information: (1) the port it calls to, (2) the arrival date, (3) the duration that the ship is berthed, (4) the flag of the ship when it called, and (5) the ship risk profile (low, medium, high risk) computed at the time of entering the port. From this port call data, we can reconstruct journeys that took place. A journey of a ship goes from one departure port to an arrival port, and has an associated travel time.

The second data source, being the inspections, provides information about ships that had a deficiency. Also, we know if such a deficiency has led to a detention. Ships without deficiencies were assumed to be compliant, because every ship should be inspected at least every three years at one of the ports participating in the Paris MoU [3]. The inspection results are used as ground truth for our classifier.

Ships in these two datasets are linked together by means of the International Maritime Organisation number—a unique identifier used in the maritime sector. We select data from years that occur in data from both sources (2014–2018), resulting in over 3 million calls from 28,416 cargo ships to a port in one of the 30 countries. Most of them, 97.3% (27,647), did not change their flag during the years under consideration. From these ships, the total number of ships with a white, grey, or black flag are 26,300, 672, and 675, respectively. Because only a small proportion of ships is flying a black or grey flag, we take them together and refer to the group as non-white flags. As mentioned before, ships can easily and quickly change their flag to either a so-called flag of convenience or to a more trustworthy flag with a better reputation [32]. In the data, 2.7% (1347) of all ships changed their flag in 2014–2018. More details on the used data are presented in the Appendix. In the next section, we present our approach to the prediction of noncompliance.

4 Methods

We aim to create a machine learning classifier that performs fair automated assessment of the risk for each ship. To this end, two types of features are used as input to the classifier; *network features* and *temporal features*.

We start by explaining the construction of the cargo ship network. In the second part we explain our approach to feature engineering, dealing with both the network features and temporal features. In the third part, we discuss the classifier in the context of machine learning. We elucidate the fair random forest classifier and explain the performance measures and fairness measures.

4.1 Cargo ship network

To obtain the structural importance of each port, later used to characterise the behaviour of ships, we construct a cargo ship network. The edges of the directed weighted network are obtained by considering the journeys of all ships, linking a port to another port if at least one ship made a journey visiting those two ports immediately after each other. Edges are weighted according to how many such journeys exist between the two ports. Hence, each node of the network is a port.

Below, we explain the structural properties of the cargo ship network in terms of their density, diameter, average distance and clustering coefficient (for a definition of these elementary network measures, see [33]). These structural properties help understand whether our cargo ship network is fact similar to earlier constructed networks of the same type. The structural importance of each port is obtained by computation of the following twelve centrality measures:

- (1) in-degree, (2) out-degree, (3) degree,
- (4) in-strength, (5) out-strength, (6) strength,
- (7) closeness centrality and (8) weighted closeness centrality,
- (9) betweenness centrality and (10) weighted betweenness centrality,
- (11) eigenvector centrality and (12) weighted eigenvector centrality.

These centrality measures are used in the features provided to a machine learning classifier. Degree and strength capture (a) the number of routes and (b) the weighted number of routes connected to a port, respectively. The strength of a port is thus equal to the number of journeys towards a port. Closeness centrality is equal to the reciprocal of the average shortest path distance from a node to all other nodes [34]. A more central node is closer to all other nodes and hence has a high closeness centrality. The betweenness centrality is equal to the number of shortest paths between every pair of nodes that pass through to the node under consideration [35]. A node with high betweenness centrality is associated to playing an important role in the network; a disruption of this node will affect many shortest paths. The eigenvector centrality is determined using eigendecomposition of the adjacency matrix [36]. High values of the eigenvector means that the node is connected to many nodes that themselves also have a high eigenvector centrality value. With these centrality measures, the aim is to capture a diverse set of measures for the structural role of a port in the cargo ship network.

The training set (used to learn the classifier) and the test set (used to estimate the performance of the classifier) should be independent. To prevent that data used to construct the network is used in both training and testing, we work with a separate hold-out data to construct the network. Hence, we divide every ship $i \in I$ into one of the two disjoint sets (here, I denotes the set containing all ships). A 10% sample of all ships I is then used for network construction (I_{network}), where the remaining ships ($I_{\text{classification}}$) are used in the classification part.

4.2 Feature engineering

We have two types of features that describe the behaviour of ships in $I_{\text{classification}}$; network features and temporal features.

4.2.1 Network features

The network features aim to capture what type of ports a given ship visits, which can correlate with noncompliant behaviour. We obtain the network features in four steps.

Step 1. *Determination of centrality measures*. We characterise each journey of a given ship by the structural importance in the cargo ship network of both the departure and arrival port. If the port is observed in the cargo ship network, the twelve centrality measures (see Sect. 4.1) are determined. For each centrality measure, we combine the value obtained from the departure port and the value obtained from the arrival port using the four arithmetic operations separately (sum, multiplication, absolute difference, division). After this step, we have $12 \cdot 4 = 48$ values characterising each journey.

Step 2. *Binning*. To capture the distribution of the values obtained for each journey, we make a histogram of these centrality measures, by splitting each of the values obtained in the previous step into ten equal-width bins. The edges of all these bins are learned from the journeys of I_{network}, to prevent information leaking. After this step, we have $48 \cdot 10 = 480$ values for each journey.

Step 3. *Aggregation*. Now, the classifier is ultimately provided with information about the instances, the individual ships. Hence, we need to aggregate the information of each journey to a fixed set of values per ship. The 480 values, obtained from step 2, can then be aggregated for each ship by summation of all journeys. Thereafter, we normalise these values by dividing them by the total number of journeys. We use the total number of journeys as a separate feature, and add it to the list of features. The procedure of normalising allows the classifier to compare the distributions, regardless of the number of journeys of a ship. In this way, we obtain $480 + 1 = 481$ features.

Step 4. *Encoding the missingness*. In step 1 we explained that the centrality measures are only defined if the port was observed in the cargo ship network. Obviously, the information that a port is missing in the network is informative for the classifier. Hence, we will encode this missingness. We do so by two separate features. The first feature equals the number of journeys where only one port was unobserved. The second feature equals the number of journeys where both ports were unobserved. In the end, we thus have $481 + 2 = 483$ network features.

4.2.2 Temporal features

The temporal features are computed from the duration of a ship's journeys and port berths. Abnormal short or long ship berths or journeys may be indicative of noncompliance. For example, very short berths may lead to rushing through safety procedures while significantly longer berths may be indicative of problems with the port authorities. We calculate the temporal features from (a) the berth duration in ports and (b) the travel time of journeys. To preserve the estimated distribution of the berth durations and travel timing during aggregation, we first make a histogram of these values for each ship. The histogram is made by splitting for each ship all berth and journey durations into ten equal-width bins. To prevent information leaking, the boundaries of the bins are learned from the port calls and journeys occurring in I_{network}. In this way, $2 \cdot 10 = 20$ temporal features are obtained. For each ship, we sum all the values obtained for each ship of (a) the histogram of the berth duration and (b) the histogram of the journey duration and divide them by the total number of berths and journeys, respectively. In total, we have 483 network features and 20 temporal features, resulting in a total of 503 features describing each ship. We will represent the 503 features by a vector x_i for some ship i in the remainder of this paper.

4.3 Machine learning classifier

We employ a machine learning classifier to perform the automated assessment of noncompliance. The goal of the classifier is to learn for each ship $i \in I_{\text{classification}}$ from the feature vector $x_i \in X$ and target scalar $y_i \in Y$ a function $f : X \mapsto Z$ where $Z \in [0, 1]$ is a score. The positive instances, i.e., $y_i = 1$, indicate a noncompliant ship and the negative instances a compliant ship. From the introduction we may recall, that in search for a particular type of fairness our aim is to reduce the classifier's dependency on a sensitive attribute $s_i \in S$, where $s_i = 0$ marks a ship with a white flag (non-sensitive) and $s_i = 1$ otherwise.

We employ a fair random forest classifier [37], which is a modified random forest classifier. In brief, a random forest classifier works as follows. For every tree in the forest, a bootstrapped sample of the training data is taken. Then, a decision tree is grown, by recursively doing three steps: (1) select a sample from all features available; (2) optimise a criterion (commonly the information gain) calculated on each of the sampled features; and (3) split the node into two child nodes based on the outcome of the optimisation. The score of an instance is calculated as the fraction of positives in a child node. For further details, we refer the reader to [38].

Random forest classifiers have, like other tree learning algorithms, some beneficial properties. We mention two of them. The first property is that their good performance has been confirmed in different domains, even with minimal tuning [38]. The second property is that the criterion considered does not have to be differentiable, in contrast to many other classifiers. Both properties allow us to use a specifically designed criterion, called Splitting Criterion Area under the curve For Fairness (SCAFF) [37]. The criterion ensures both that different labels are separated and that the sensitive class remain mixed. It is defined as follows:

$$\mathrm{SCAFF}(Z, Y, S, \Theta) = (1 - \Theta) \cdot \mathrm{AUC_Y}(Z, Y) - \Theta \cdot \mathrm{AUC_S}(Z, S),$$

with $\mathrm{AUC_Y} \in [0, 1]$ marking the well-known Area Under the receiver operating characteristic Curve:

$$\mathrm{AUC_Y} = \frac{\sum_{i=1}^{s_+} \sum_{j=1}^{s_-} \sigma(Z_i, Z_j)}{s_+ \cdot s_-} \quad \text{with } \sigma(Z_i, Z_j) = \begin{cases} 1, & \text{if } Z_i > Z_j, \\ \frac{1}{2}, & \text{if } Z_i = Z_j, \\ 0, & \text{otherwise,} \end{cases}$$

where s_+ and s_- marks the number of positive and negative instances, respectively. An $\mathrm{AUC_Y}$ value of 0.5 suggests random classification while $\mathrm{AUC_Y} = 1$ indicates a perfect classifier. The $\mathrm{AUC_S}$ considers the sensitive attribute as positive class. It is defined as follows:

$$\mathrm{AUC_S}(Z, S) = \max\left(1 - \frac{\sum_{i=1}^{s_+} \sum_{j=1}^{s_-} \sigma(Z_i, Z_j)}{s_+ \cdot s_-}, \frac{\sum_{i=1}^{s_+} \sum_{j=1}^{s_-} \sigma(Z_i, Z_j)}{s_+ \cdot s_-}\right),$$

with $\sigma(Z_i, Z_j)$ defined exactly the same as for $\mathrm{AUC_Y}$. The measure is closely related to strong demographic parity [39]. For $\mathrm{AUC_S} = 0.5$, corresponding to a strong demographic parity of 0, the split in the node is made regardless of the values of the sensitive attributes, meaning equality of outcome. A value of $\mathrm{AUC_S} = 1$, corresponding to a strong demographic parity of 1, is the worst score possible, since in that case the classifier is able to predict the sensitive attribute perfectly. The orthogonality parameter, $\Theta \in [0, 1]$, allows to balance the performance-fairness trade-off [14]. The fair random forest classifier optimises for performance when $\Theta = 0$ and thus does not consider fairness. Hence, it corresponds in that case to the ordinary random forest classifier with an information gain criterion. In contrast, when provided a value of $\Theta = 1$, it optimises fairness and neglects performance, resulting in a random classifier. More details on the fair random forest classifier are given in [37].

4.4 Performance measures

The performance of the classifier can be determined both by threshold-dependent and threshold-free metrics. Scores equal to or above the threshold $t \in [0,1]$ are classified as positive ($\hat{y}_i = 1$) and values under the threshold are predicted negative ($\hat{y}_i = 0$). Threshold-free metrics have the advantage that they do not require this explicit cut-off point and instead consider the ranking imposed by the scores of the classifier. The three threshold-dependent performance metrics used by us are the precision, recall, and the harmonic mean of those two, the F_1 score. The threshold-free performance metric used in this work is the $\mathrm{AUC}_{\mathrm{Y}}$ (see previous section).

4.5 Fairness measures

Similar to the performance measures, fairness with respect to the sensitive group can also be quantified by threshold-dependent and threshold-free metrics. We report the threshold-dependent precision and recall for the two groups, i.e., ships with a white flag and a non-white flag. A large difference between the two groups indicates an unfair outcome of the model.

Moreover, we use the threshold-dependent demographic parity and equalised odds [13]. These measures consider the difference for some performance measures between the two groups, i.e., ships with a white flag and a non-white flag. The demographic parity measure is the absolute difference between the positive prediction rates of the two groups, i.e., $|P(\hat{Y} = 1|S = 1) - P(\hat{Y} = 1|S = 0)| \leq \epsilon_{\text{parity}}$. Lower values of ϵ_{parity} indicate more equal outcomes and thus more fair predictions. The equalised odds metric is defined as follows:

$$\begin{aligned}
&\bigl|P(\hat{Y} = 1|S = 1, Y = 0) - P(\hat{Y} = 1|S = 0, Y = 0)\bigr| \leq \epsilon_{\text{odds}}, \\
&\bigl|P(\hat{Y} = 1|S = 1, Y = 1) - P(\hat{Y} = 1|S = 0, Y = 1)\bigr| \leq \epsilon_{\text{odds}}.
\end{aligned}$$

It means that the equalised odds measures the equality of opportunity in a supervised setting, with lower values for ϵ_{odds} implying more equal opportunity and thus more fair predictions.

The threshold-independent fairness measure used in this work, is the aforementioned $\mathrm{AUC}_{\mathrm{S}}$.

5 Results

The section starts with our experimental setup. Then, we continue with the analysis of the cargo ship network. Subsequently, we evaluate the performance of the baseline ship risk profile. After that is established, we report on the performance of the (non-fair) random forest classifier. We conclude by reporting the performance of the fair random forest classifiers.

5.1 Experimental setup

In this work, we use five-fold nested cross validation with stratified sampling [40]. The inner folds are used to select the best parameter set for that specific outer fold. The considered parameters are combinations of the selected values for the depth of each tree ($\{1, 2, \ldots, 10\}$) and the number of bins used in discretization of the values of the continuous variables (2 or 10). Hence, there are $10 \cdot 2 = 20$ candidate sets of parameters in each outer fold. The mean and standard deviation of the performance of the classifier are evaluated on the five outer folds using the selected parameter set. We report the outcome of this cross validation for 11 different values of the orthogonality parameter, $\Theta \in \{0, 0.1, 0.2, \ldots, 1\}$.

The code used in this research is publicly available [41]. It uses several open source Python packages. Specifically, scikit-learn [42], SciPy [43], and Pandas [44] are used for feature engineering and for measuring the performance of the baseline ship risk profile and the proposed classifier. The fair random forest is also open source [45], making extensive use of the CVXpy package for optimising SCAFF [46]. For the analysis on the cargo ship network we used the NetworkX package [47]. The C++ library teexGraph was used to determine the diameter of the network [48]. The packages used for visualisation, and all other dependencies and supportive software versions, can be found at [41].

5.2 Cargo ship network

A quite 'overwhelming' visualisation of the cargo ship network is shown in Fig. 1. Still, we only show ports in Europe, because we are interested in predicting risk for ships that arrive in Europe. From the figure, we can learn the following. First, we see a large component connecting virtually all ports. Second, we observe that only a few ports have a high strength, as indicated by the yellow colour, of which (1) Puttgarden (Germany), (2) Rotterdam (Netherlands), and (3) Algeciras (Spain) have the highest strength. Third, two different types of ports can be distinguised; (1) ports that are well-connected (e.g. ports in Germany, Netherlands, and Belgium), and (2) ports that are more in the periphery of the network (e.g. Iceland and the Azores). Fourth, we see that some ports are connected by thick lines, indicating an edge with a high weight. The nodes that are connected by these edges are likely to have a high weighted betweenness centrality, because failure of such node would cause to have other shortest paths running through edges with less weight.

In Table 1 we provide numeric information on sizes, relations, and distances. We show nine common properties of the cargo ship network of our work in the second column. In the third through sixth column we provide values for the properties of four similar cargo

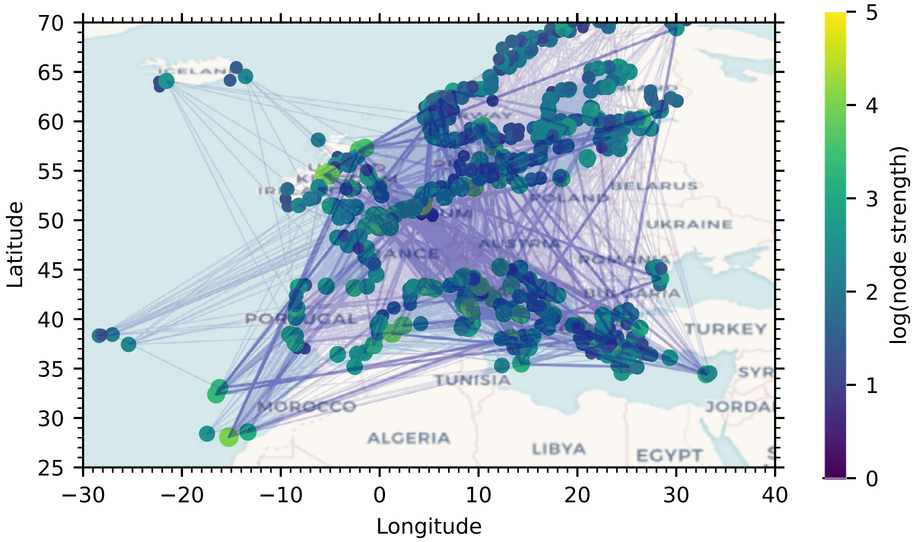


Figure 1 The considered cargo ship network. Nodes are coloured by their strength. Thicker edges mark more busy routes. Generated using OpenStreetMap data

Table 1 General properties of the considered cargo ship network used in this work, compared to others reported in literature. GC denotes the giant component (largest connected component). In case of directed networks we report on the largest strongly connected component

Work	This work	[31]	[30]	[29]	[28]
Directed	Yes	Yes	No	No	No
Number of nodes	1459	728	1488	439	951
Number of nodes in GC	1445	726	–	–	935
Number of routes	28,653	18,142	17,135	2331	36,328
Number of routes in GC	28,638	18,140	–	–	–
Density in GC	0.027	0.03	0.015	0.019	0.08
Diameter in GC	6	7	–	7	8
Average distance in GC	2.63	2.49	2.99	3.290	2.5
Clustering coefficient in GC	0.48	0.58	0.55	0.396	0.49

networks observed in literature [28–31]. We compare these properties in an attempt to better understand whether our 10% sample used to compute port features is representative. From Table 1 we see that even though very different numbers of nodes and edges are reported in these works, the measures such as density, diameter, average distance, and clustering coefficient are similar. Hence, we may conclude that the constructed cargo ship network can be used to sensibly extract mobility patterns for our ship compliance classifier.

5.3 Performance of the baseline ship risk profile

The confusion matrices are shown for the white and non-white flags separately in Fig. 2. Together with Table 2, where we show the calculated performance and fairness measures, they provide information on the performance of the baseline ship risk profile. Ships having a medium risk are predicted as compliant. We observe that virtually no ship flying a non-white flag gets a low risk profile. The majority of the ships (90%) are classified as medium risk. Of these ships, only a small fraction (22%) is compliant. From the ships with a high risk profile, only a tiny fraction (4%) is compliant, resulting in a high precision for the baseline model. However the recall is quite low as many ships with a medium ship risk profile are noncompliant. Interestingly, ships with a white flag having a low or medium risk profile are noncompliant more frequently than ships with a non-white flag. This also results in a low value of the AUC_Y value of only 0.543 ± 0.006. Hence, we may conclude that, at least using the data from 2014–2018, we cannot predict compliance with the baseline ship risk profile.

The model is quite unfair. In particular, we observe a large difference in the F_1 metric for the white and non-white group, resulting in high values for $\epsilon_{\mathrm{parity}}$ and ϵ_{odds}. There is a strong correlation between the sensitive attribute, i.e., the ship flag, and the scores of the model with $\mathrm{AUC}_\mathrm{S} = 0.672 \pm 0.010$.

5.4 Random forest classifier

The confusion matrices of the random forest classifier are shown in Fig. 2, and in Table 2 we report the performance and fairness metrics. We observe that more ships are predicted correctly compared to the baseline model. The recall is higher, meaning that many actual positives are predicted as such. This comes with decreased precision, indicating that more compliant ships are predicted as noncompliant. However, the harmonic mean of the two, i.e., the F_1 measure, is higher than the baseline model, indicating that the random forest

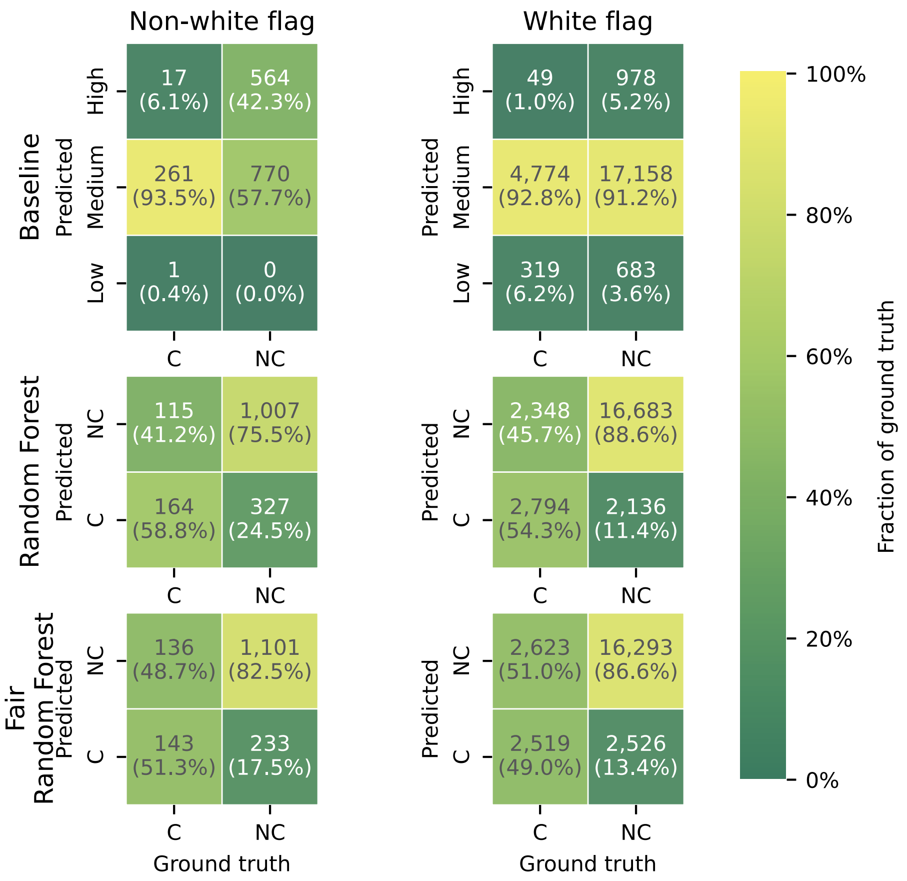


Figure 2 Confusion matrices for both the white and non-white flagged ships. The baseline model is the ship risk profile currently in use. C and NC mark compliant and noncompliant ships, respectively The percentages (and colour coding) are stratified based on the ground truth

Table 2 Performance (precision, recall, F_1) and fairness (demographic parity and equalised odds) measures for the different models

Measure	Baseline	Random forest	Fair random forest
precision (non-white)	97.1%	89.8%	89.0%
precision (white)	95.2%	87.7%	86.1%
recall (non-white)	42.3%	75.5%	82.5%
recall (white)	5.2%	88.6%	86.6%
F_1 (non-white)	58.9%	82.0%	85.6%
F_1 (white)	9.9%	88.2%	86.4%
ϵ_{parity}	0.317	0.099	0.023
ϵ_{odds}	0.371	0.132	0.040
AUC_Y	0.543 ± 0.006	0.814 ± 0.004	0.776 ± 0.008
AUC_S	0.672 ± 0.010	0.627 ± 0.014	0.538 ± 0.011

classifier outperforms the baseline model. This finding is supported by the AUC_Y measure, showing a value of 0.814 ± 0.004. This implies that that we can accurately assess ship noncompliance in an automated fashion with a random forest classifier, using behavioural data. This answers our first research question.

The confusion matrices show that ships with a white flag are predicted to be noncompliant more often than ships with a non-white flag. The difference in frequency also results in a higher recall for ships with a white flag. All in all, the prediction is much more fair compared to the baseline model. Since the random forest classifier does not use the flag as a feature. It means that using only behavioural data makes the model more fair.

5.5 Fair random forest classifier

Below we list our observations. First, from the confusion matrices in Fig. 2 and the performance and fairness metrics in Table 2 we observe that the fair random forest classifier has more comparable true positive and true negative rates amongst ships flying a white and non-white flag, with only a slight cost in predictive performance on the target. Second, in terms of the F_1 measure, we observe that the performance drops only for the ships flying a white flag, such that the difference between the two groups becomes very small. Third, we observed that the demographic parity and equalised odds measures decrease when using a fair random forest classifier, suggesting that the classifier was able to improve fairness.

Before drawing any conclusion, we show the effect of the orthogonality parameter in more detail (see Fig. 3). The top left figure shows that the AUC_Y measure is only weakly influenced for a broad range of values for the orthogonality parameter (Θ), meaning that overall, we can reliably ensure equality of outcome while maintaining acceptable performance. An orthogonality value of 0.7 appears to give the best trade-off between performance and fairness in our work, with a performance of $\mathrm{AUC}_\mathrm{Y} = 0.776 \pm 0.008$ and fairness

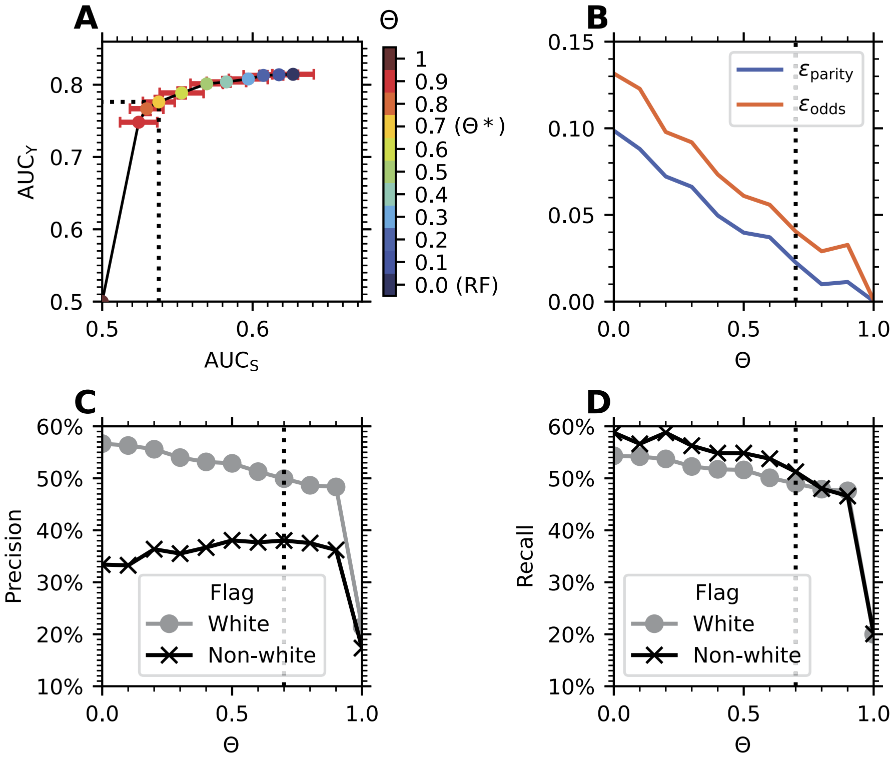


Figure 3 (**A**) The performance of the fair random forest classifier for different values of the orthogonality. (**B**) The fairness performance measured in demographic parity and equalised odds for different values of Θ. (**C**)–(**D**) The performance measured in precision (**C**) and recall (**D**) for different values of Θ, separated for ships flying a white and non-white flag

of $\text{AUC}_S = 0.538 \pm 0.011$. The performance can be further improved (although slightly, to $\text{AUC}_Y = 0.814$), but only at decreased equality of outcome and vice versa.

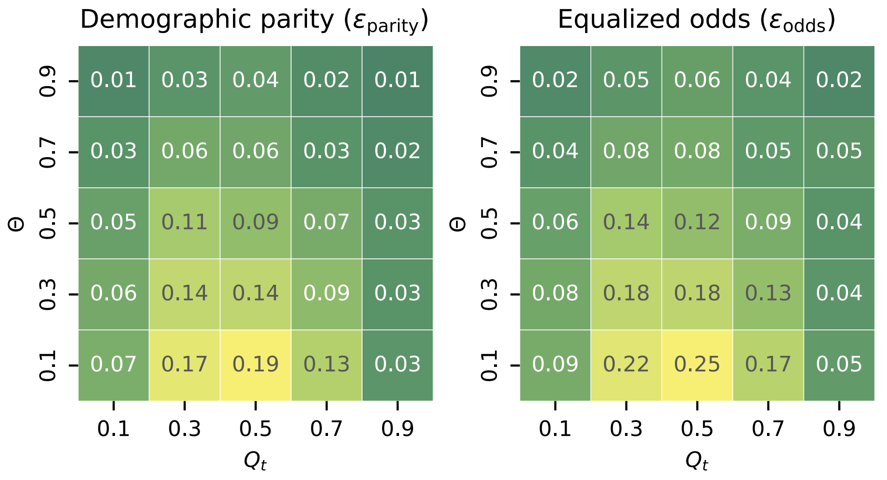
Then we will have a closer investigation of Fig. 3(B), where the two fairness measures decrease monotonically at increasing orthogonality values. At the extreme value of $\Theta = 1$ they are zero, but at this value the predictive performance is also very low, as can be observed in Fig. 3(A).

Subsequently, in Figs. 3(C) and 3(D) we observe that the precision and recall for ships flying a white and non-white flag have only small differences for larger values of the orthogonality. The precision of the ships flying a non-white flag increases slightly at higher values of the orthogonality, at the cost of precision for ships with a white flag. To calculate these values, the threshold was set to $t = 0.34$ in such a way that $P(Z \geq t)$ is equal to $P(Y = 1)$. This threshold is also used to calculate the confusion matrix shown in Fig. 2.

Here we remark that the threshold t is important, as it determines how many ships are selected as being noncompliant. Higher values of the threshold result in fewer ships that are predicted as being noncompliant. Therefore, we define the threshold quantile Q_t in such a way that $P(z \geq t)$ equals the threshold quantile.

Finally, in Fig. 4 we show the effect of the orthogonality and the threshold quantile on the selected threshold-dependent fairness measures. We observe that high values of the orthogonality yield a fair prediction for all values of the threshold, even when the threshold quantile is set to a high value, such that most ships are predicted to be compliant. For lower values of the orthogonality, we observe that the fairness of the model is worst when the threshold quantile is near 0.5. This result is expected, as at other values of the threshold quantile the performance for both groups is low, leading to a small difference between the groups. Even at these 'bad' choices for the orthogonality and threshold quantile, the values of the demographic parity measure and the equalised odds measure are still lower than observed for the baseline ship risk profile.

From all these observations and results we may conclude that the fair random forest classifier is effective in reducing bias towards the flag of a ship, for wide ranges of the used threshold and orthogonality. This answers the second research question.



Θ \ Q_t (Demographic parity)	0.1	0.3	0.5	0.7	0.9
0.9	0.01	0.03	0.04	0.02	0.01
0.7	0.03	0.06	0.06	0.03	0.02
0.5	0.05	0.11	0.09	0.07	0.03
0.3	0.06	0.14	0.14	0.09	0.03
0.1	0.07	0.17	0.19	0.13	0.03

Θ \ Q_t (Equalized odds)	0.1	0.3	0.5	0.7	0.9
0.9	0.02	0.05	0.06	0.04	0.02
0.7	0.04	0.08	0.08	0.05	0.05
0.5	0.06	0.14	0.12	0.09	0.04
0.3	0.08	0.18	0.18	0.13	0.04
0.1	0.09	0.22	0.25	0.17	0.05

Figure 4 Fairness measures evaluated on the proposed classifier, as function of the threshold quantile Q_t and orthogonality Θ

6 Discussion

In this section we discuss two limitations of our proposed classifier. The first limiation concerns the ground truth. It might be biased towards the flag, as well as towards the inspector's background [12]. The problem is that different inspectorates judge compliance differently for similar ships. The difference in judging leads to inequality between ports and so-called port-shopping. Port-shopping denotes a situation when a noncompliant ship decides to go to another port, solely because the inspection regime there is more favourable to noncompliant ships. In this way the ship yields a lower risk profile. Port-shopping influences our model since the ground truth data is unjustly positive for such noncompliant ships. One of the reasons for the existence of the Paris MoU is to avoid this kind of competition between ports [2]. Hence, in future work, the country of inspection could also be added as a sensitive attribute, which can reduce correlation between the inspectorate and the inspection outcome.

The second limitation of this work is conceptually related to Goodhart's law, commonly formulated as: "When a measure becomes a target, it ceases to be a good measure" [49]. This is applicable to any ship risk model because ships have incentive to get a low-risk profile. In the baseline ship risk model, a better risk profile could be achieved by changing an administrative property of the ship. In our proposed classifier ships would need to change their behaviour to get a better score, which is substantially harder to achieve than merely changing administrative properties.

7 Conclusions

The aim of the present research was to devise an automated, accurate and fair ship risk classification approach. The study has led to two conclusions. First, by using a fair random forest classifier, we can offset the confirmation bias present in historical inspection data. Experimental results indicate that the disparate impact and equalised odds measures improve significantly, regardless of the chosen parameters, meaning that the constructed classifier works well. Second, we may conclude that the performance of our approach provided with behavioural data is $\mathrm{AUC}_{\mathrm{Y}} = 0.776 \pm 0.008$, clearly improving on the $\mathrm{AUC}_{\mathrm{Y}} = 0.543 \pm 0.006$ of the ship risk profile currently in use. Hence our work is supportive for global efforts to minimise risks associated to maritime transport by conducting more targeted inspections. More generally, we have shown how ubiquitous mobility information can be used to perform inspection in a better and more fair way. The devised approach may also be applicable in other inspection applications broader than port state control.

A natural progression of this work is to determine with domain experts what behaviour is often associated with high risk and subsequently reduce this risky behaviour. A second possible direction for future work is to consider higher-order effects in the cargo ship network [50]. The construction of a higher-order network allows for a more accurate representation of the complex underlying system, that in turn may enable more accurate network analysis results. It has been shown that relations up to the fifth-order may be relevant in cargo shipping networks [50]. Finally, we may conclude that, the temporal aspect of the network can be exploited to obtain a better, more accurate centrality measure of the true, time-aware structural importance of the ports [51], therewith potentially resulting in an even better performing classifier for the task at hand.

Appendix

This appendix consists of three figures providing more insight into the data used in this work. First, in Fig. A1 the fraction of noncompliant ships that visit all countries is shown. We remind the reader that a ship is noncompliant if at least one deficiency has been found during 2014–2018 (see Data section). We observe that this number is very different across countries in Europe.

Second, in Fig. A2 the number of ships registered to each country is shown. Although difficult to observe, most ships are registered to Panama (2,904), Marshall islands (2,153), and Liberia (2,119). These flag states are typically known as 'flags of convenience', which is explained in the Data section.

Finally, in Fig. A3, for each flag, the fraction of the noncompliant ships is shown. It is true that some non-white flags are associated to a large fraction of noncompliant ships. The other way around, some of the white flag states have many noncompliant ships, such as the

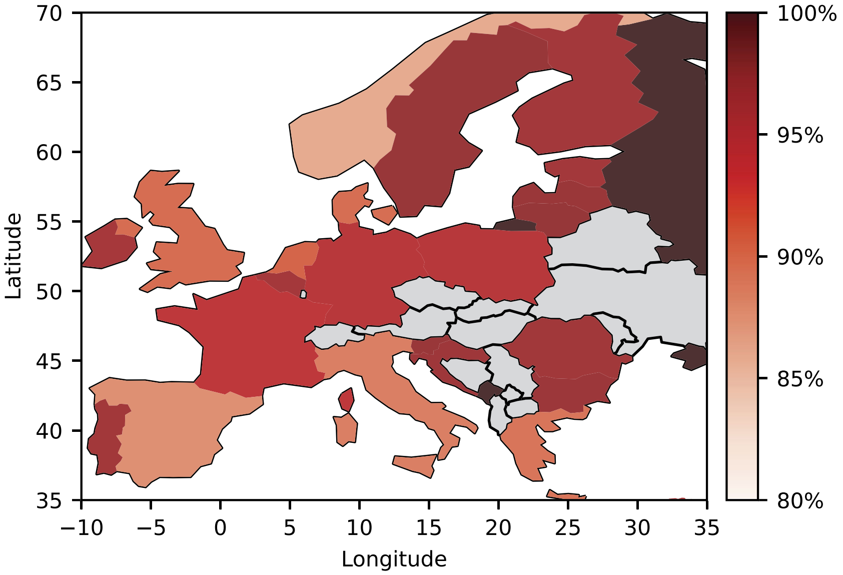


Figure A1 Fraction of noncompliant ships per country. Grey countries were not visited by a ship in the data

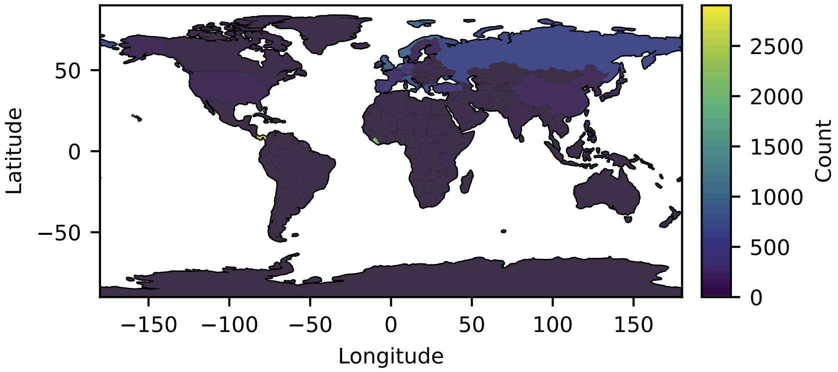


Figure A2 Number of ships registered to each country. If multiple registrations for a single ship were observed, we use the most recent registration

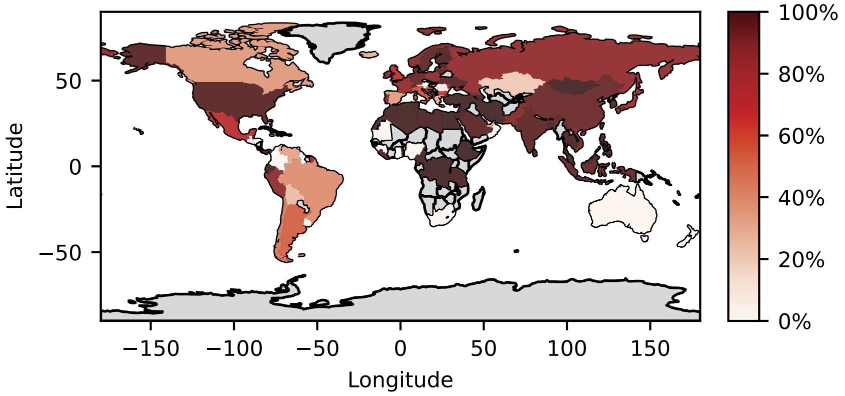


Figure A3 The fraction of noncompliant ships for each flag state. Countries without any ships registered to it, are shown in grey

United States of America. In our online repository [41], these figures can be downloaded at a higher resolution.

Acknowledgements
We thank Paul Merckx for data acquisition and curation and Jasper van Vliet for fruitful and critical discussions.

Funding
This work was partially funded by the Human Environment and Transport Inspectorate and the Dutch Ministry of Infrastructure and Water Management.

Abbreviations
AUC, Area Under the receiver operating characteristic Curve; SCAFF, Splitting Criterion AUC For Fairness.

Availability of data and materials
The data used in this research is not publicly available due to restrictions from the European Maritime Safety Agency. The code is publicly available [41].

Declarations

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors contributed to the study conception and design. Formal analysis and methodology were performed by GJdB and APB. Supervision was done by HJvdH, FWT, CJV. GJdB wrote the original draft. All authors read and approved the final manuscript.

Author details
[1]Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333 CA, Leiden, the Netherlands. [2]Human Environment and Transport Inspectorate, Rijnstraat, 2515 XP, the Hague, the Netherlands. [3]Leiden Centre of Data Science, Schouwburgstraat 2, 2511 VA, the Hague, the Netherlands. [4]Data Science Department, TNO, Anna van Buerenplein 1, 2595 DA, the Hague, the Netherlands.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 15 December 2021 Accepted: 24 February 2022 Published online: 12 March 2022

References
1. United Nations Conference on Trade and Development (2020) Review of maritime transport. United Nations Publications, New York.
2. Paris MoU (2021) Paris memorandum of understanding on port state control. https://www.parismou.org/sites/default/files/Paris%20MoU%20including%2043rd%20amendment%20final.pdf. Accessed 25 Nov 2021

3. European Union (2009) Directive 2009/16/EC of the European Parliament and of the Council of 23 April 2009 on port state control. Off J Eur Union L131:57–100
4. Knapp S (2007) The econometrics of maritime safety. PhD thesis, Erasmus University Rotterdam. http://hdl.handle.net/1765/7913. Accessed 28 Dec 2021
5. Yang Z, Yang Z, Yin J, Qu Z (2018) A risk-based game model for rational inspections in port state control. Transp Res, Part E, Logist Transp Rev 118:477–495. https://doi.org/10.1016/j.tre.2018.08.001
6. Cariou P, Meijia MQ Jr, Wolff F-C (2007) An econometric analysis of deficiencies noted in port state control inspections. Marit Policy Manag 34(3):243–258. https://doi.org/10.1080/03088830701343047
7. Rodríguez E, Piniella F (2012) The new inspection regime of the Paris MoU on port state control: improvement of the system. J Marit Res 9(1):9–16
8. Paris MoU (2020) Current flag performance list. https://www.parismou.org/detentions-banning/white-grey-and-black-list. Accessed 25 Nov 2021
9. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 259–268. https://doi.org/10.1145/2783258.2783311
10. Cariou P, Wolff F-C (2011) Do port state control inspections influence flag- and class-hopping phenomena in shipping? J Transp Econ Policy 45(2):155–177
11. Bloor M, Datta R, Gilinskiy Y, Horlick-Jones T (2006) Unicorn among the cedars: on the possibility of effective 'smart regulation' of the globalized shipping industry. Soc Leg Stud 15(4):534–551. https://doi.org/10.1177/0964663906069546
12. Graziano A, Cariou P, Wolff F-C, Mejia MQ, Schröder-Hinrichs J-U (2018) Port state control inspections in the European Union: do inspector's number and background matter? Mar Policy 88:230–241. https://doi.org/10.1016/j.marpol.2017.11.031
13. Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Proceedings of the 30th conference on neural information processing systems (NIPS). Advances in neural information processing systems, pp 3315–3323
14. Kleinberg JM, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. http://arxiv.org/abs/1609.05807v2. Accessed 17 Jan 2022
15. Pallotta G, Vespe M, Bryan K (2013) Vessel pattern knowledge discovery from AIS data: a framework for anomaly detection and route prediction. Entropy 15(6):2218–2245. https://doi.org/10.3390/e15062218
16. Xiao Y, Wang G, Lin K-C, Qi G, Li KX (2020) The effectiveness of the new inspection regime for port state control: application of the Tokyo MoU. Mar Policy 115:103857. https://doi.org/10.1016/j.marpol.2020.103857
17. Xiao Y, Qi G, Jin M, Yuen KF, Chen Z, Li KX (2021) Efficiency of port state control inspection regimes: a comparative study. Transp Policy 106:165–172. https://doi.org/10.1016/j.tranpol.2021.04.003
18. Degré T (2007) The use of risk concept to characterize and select high risk vessels for ship inspections. WMU J Marit Aff 6:37–49. https://doi.org/10.1007/BF03195088
19. Xu R-F, Lu Q, Li KX, Zheng H-S (2007) A risk assessment system for improving port state control inspection. In: Proceedings of the 6th international conference on machine learning and cybernetics, pp 818–823. https://doi.org/10.1109/ICMLC.2007.4370255
20. Xu R, Lu Q, Li K, Li W (2007) Web mining for improving risk assessment in port state control inspection. In: Proceedings of the 2007 international conference on natural language processing and knowledge engineering, pp 427–434. https://doi.org/10.1109/NLPKE.2007.4368066
21. Degré T (2008) From black–grey–white detention-based lists of flags to black–grey–white casualty-based lists of categories of vessels? J Navig 61(3):485–497. https://doi.org/10.1017/S0373463308004773
22. Heij C, Knapp S (2019) Shipping inspections, detentions, and incidents: an empirical analysis of risk dimensions. Marit Policy Manag 46(7):866–883. https://doi.org/10.1080/03088839.2019.1647362
23. Wang S, Ran Yan XQ (2019) Development of a non-parametric classifier: effective identification, algorithm, and applications in port state control for maritime transportation. Transp Res, Part B, Methodol 128:129–157. https://doi.org/10.1016/j.trb.2019.07.017
24. Yan R, Wang S, Peng C (2021) An artificial intelligence model considering data imbalance for ship selection in port state control based on detention probabilities. J Comput Sci 48:101257. https://doi.org/10.1016/j.jocs.2020.101257
25. Gao Z, Lu G, Liu M, Cui M (2008) A novel risk assessment system for port state control inspection. In: Proceedings of the 2008 IEEE international conference on intelligence and security informatics (ISI), pp 242–244. https://doi.org/10.1109/ISI.2008.4565068
26. Yan R, Wang S, Peng C (2021) Ship selection in port state control: status and perspectives. Marit Policy Manag. https://doi.org/10.1080/03088839.2021.1889067
27. Knapp S, Franses PH (2008) Econometric analysis to differentiate effects of various ship safety inspections. Mar Policy 32(4):653–662. https://doi.org/10.1016/j.marpol.2007.11.006
28. Kaluza P, Kölzsch A, Gastner MT, Blasius B (2010) The complex network of global cargo ship movements. J R Soc Interface 7:1093–1103. https://doi.org/10.1098/rsif.2009.0495
29. Liu C, Wang J, Zhang H (2018) Spatial heterogeneity of ports in the global maritime network detected by weighted ego network analysis. Marit Policy Manag 45(1):89–104. https://doi.org/10.1080/03088839.2017.1345019
30. Peng P, Cheng S, Chen J, Liao M, Wu L, Liu X, Lu F (2018) A fine-grained perspective on the robustness of global cargo ship transportation networks. J Geogr Sci 28(7):881–889. https://doi.org/10.1007/s11442-018-1511-z
31. van Veen N (2020) The complex network of ship movements in Europe. Master's thesis. https://www.gerritjandebruin.nl/attachments/nathalie.pdf. Accessed 28 Dec 2021
32. NGO Ship Breaking Platform Breaking platform: flags of convenience. https://shipbreakingplatform.org/issues-of-interest/focs/. Accessed 28 Dec 2021
33. Newman MEJ (2018) Networks, 2nd edn. Oxford University Press, Oxford
34. Freeman LC (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239
35. Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40(1):35–41. https://doi.org/10.2307/3033543

36. Bonacich P (1987) Power and centrality: a family of measures. Am J Sociol 92(5):1170–1182. https://doi.org/10.1086/228631
37. Pereira Barata A, Takes FW, van den Herik HJ, Veenman CJ (2021) Fair tree classifier using strong demographic parity. https://arxiv.org/abs/2110.09295v3
38. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2nd edn. Springer, New York
39. Jiang R, Pacchiano A, Stepleton T, Jiang H, Chiappa S (2020) Wasserstein fair classification. In: Proceedings of the 35th uncertainty in artificial intelligence conference (UAI). Proceedings of machine learning research, pp 862–872
40. Cawley GC, Talbot NL (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 11:2079–2107
41. de Bruin GJ (2020) Fair automated assessment of noncompliance in cargo ship networks. https://doi.org/10.5281/zenodo.5727084
42. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
43. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 1.0 Contributors (2020) Scipy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2
44. McKinney W (2010) Data structures for statistical computing in Python. In: Proceedings of the 9th Python in science conference (SciPy), pp 56–61
45. Pereira Barata A (2021) Fair tree classifier using strong demographic parity. https://doi.org/10.5281/zenodo.5718556
46. Agrawal A, Verschueren R, Diamond S, Boyd S (2018) A rewriting system for convex optimization problems. J Control Decis 5(1):42–60. https://doi.org/10.1080/23307706.2017.1397554
47. Hagberg A, Swart P, S Chult D (2008) Exploring network structure, dynamics, and function using NetworkX. https://www.osti.gov/biblio/960616. Accessed 28 Dec 2021
48. Takes FW, Kosters WA (2011) Determining the diameter of small world networks. In: Proceedings of the 20th ACM international conference on information and knowledge management (CIKM), pp 1191–1196. https://doi.org/10.1145/2063576.2063748
49. Strathern M (1997) 'Improving ratings': audit in the British university system. Eur Rev 5(3):305–321
50. Saebi M, Xu J, Kaplan LM, Ribeiro B, Chawla NV (2020) Efficient modeling of higher-order dependencies in networks: from algorithm to application for anomaly detection. EPJ Data Sci 9:15. https://doi.org/10.1140/epjds/s13688-020-00233-y
51. Scholtes I, Wider N, Garas A (2016) Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities. Eur Phys J B 89:61. https://doi.org/10.1140/epjb/e2016-60663-0