Network analysis methods for smart inspection in the transport domain

Bruin, G.J. de

6

Fair automated assessment of noncompliance in cargo ship networks

International cargo ships must follow a plethora of safety standards and international treaties [147]. Governmental inspectorates currently assess a ship's compliance with the help of a rule-based process using the color (white, gray, or black) of a ship's flag as a dominant factor. The flag's color is determined yearly by considering the fraction of noncompliant ships of that flag [145]. The usage of the flag's color can lead to confirmation bias and unfair inspections. Rather than using static ship characteristics, we wish to utilize information about the actions of the ship, i.e., its behavior. This brings us to the following research question.

Research question 5: *How can ship behavior be utilized to enable smart inspection of cargo ships?*

We propose an approach for smart inspection (cf. Definition 2), and aim to realize two crucial contributions. First, we would like to reduce confirmation bias by using a fair model. Second, we aim to extract relevant mobility patterns from a cargo ship network (see Definition 13), allowing us to derive meaningful behavioral features for ship classification. Our approach will improve fairness at the cost of a limited performance loss. Thereby, it will enhance *maritime safety and protection* through smarter inspection targeting. In a general sense, this work demonstrates how network science can use behavioral data for smart inspection.

The current chapter corresponds to the following publication:
G. J. de Bruin, A. Pereira Barata, C. J. Veenman, H. J. van den Herik, and F. W. Takes. „Fair automated assessment of non-compliance in cargo ship networks." *EPJ Data Science* 11, 13 (2022). DOI: 10.1140/epjds/s13688-022-00326-w

6.1 Smart cargo ship inspection

Maritime cargo transport is essential to global trade, often being the most cost-effective way to move goods from one place to another. It results in many ship movements worldwide; around 80% of world merchandise is carried by sea [190]. However, we mentioned in Chapter 1 that maritime transport has risks, such as (1) labor exploitation, (2) culpable ship accidents, and (3) environmental pollution. These risks need to be mitigated by shipowners. Port State Control (PSC) inspections are conducted when ships berth in a port to ensure mutual trust between countries that *all* ships adhere to the same international laws. There are two possible outcomes of an inspection; either the ship is found fully compliant, or there are particular noncompliances. These PSC 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 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 specific ports any longer. In this research, we aim to predict 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 strict laws to mitigate the negative consequences of maritime transport. Members of the Paris Memorandum of Understanding (MoU)[1] introduced a so-called New Inspection Regime (NIR) [147]. Arguably the most significant 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 [147]. The six factors used in the risk profile for a given ship are [56] derived from (1) the type of a ship, (2) the age of a ship, (3) commercially issued safety certificates, (4) owning company's performance, (5) historical misconducts, and (6) the flag a ship is flying, or equivalently, the country of registration. Using the score, each ship is classified as low-risk or high-risk. Ships classified as low-risk should be inspected every three years, while ships classified as having a high-risk profile should be inspected every six months. With the ship risk profile, the NIR 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 $\$1,000$ [98]. In [210], it was estimated that a noncompliant ship saves, on average, around $\$400,000$ on maintenance by not complying with regulations, whereas the loss of a ship can incur costs up to $\$67,000,000$. Shipowners with a low-risk profile can benefit by reducing inspection burden, saving precious turn-around time in the port.

From the six factors used in the current ship risk profile, the flag plays a vital role [36, 166]. The flag is considered black, gray, or white based on the detention ratio of the

[1]The following countries are part of the Paris MoU: all European Union coastal countries, Canada, Norway, Russia, and the United Kingdom.

country over a three-year rolling period [145]. Fleets from countries on the blacklist were significantly more often detained over a three-year period than fleets from countries on the whitelist. We mention three drawbacks in considering the flag for the ship risk profile.

1. There are ethical concerns. The use of the flag can be considered disparate treatment [57] because ships are intentionally treated differently based on membership of a privileged class, being the white flag.
2. There are opportunities for ships to change flags, opening up the possibility for noncompliant ships to "hide" under a white flag [37]. Although changing flags does not necessarily improve compliance, the NIR would grant such a ship a lower risk profile. In an ideal situation, merely changing an administrative property of a ship should keep the assessment of the risk associated with that ship the same.
3. Inspectors can use their discretion (possibly leading to subjectivity) to decide how thorough an inspection is.

Hence, ships flying a black flag could be subjected to stricter inspections, resulting in a higher probability of finding a noncompliant issue [20, 67]. 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 [37]. The potential danger of inspectors' bias has been recognized, and great efforts are made to harmonize the training of inspectors, thereby making the overall inspectorate system consistent [56]. Nevertheless, complete global harmonization has yet to be achieved [67].

An option is to start ignoring a ship's flag altogether to reduce the confirmation bias mentioned earlier, thus providing what in the literature [75] is known as *equal opportunity*. However, correlations exist between the other characteristics of a ship and its target; thus, the classifier will indirectly learn to use the ship's flag, 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 model [96] (see Definition 5), 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 model reduces the confirmation bias and improves the overall fairness of the risk assessment.

Our contribution

Rather than using potentially unfair and biased static ship characteristics, we prefer to consider the ships' actual behavior for noncompliance prediction, explicitly moving away from the six factors used in the ship risk profile. Ship behavior has been used to find anomalous ships [141], which may indicate noncompliance.

An example of ship behavior potentially characteristic of noncompliance is a ship sailing primarily on routes with much competition. Such routes may lead to reduced profit margins and a greater push for owners to cut shipping costs at the expense of safety.

While we have yet to determine the fares on specific routes, our proposed classifier will still consider relations between noncompliance and the sailed routes.

In the current study, we derive a cargo ship network from data containing notifications of ships calling a port. In the cargo ship network, nodes are ports, and edges are ships that travel between ports. By considering each port's structural function in the network, we extract mobility patterns for each ship. These mobility patterns are provided to the fair machine learning model, enabling automated assessment of the risk of ships based on their behavior. Altogether, we have devised an accurate, automated, interpretable and fair assessment of ship noncompliance based on ship behavior, providing an answer to Research question 5. The data used in our approach is available to all members of the European Maritime Safety Agency, allowing each of them to apply our approach.

The structure of this chapter is as follows. In Section 6.2, we provide related work on the ship risk profile and ship risk classification. Then, we explain the cargo shipping data used in this work in Section 6.3. Subsequently, we describe the research methodology in Section 6.4 We present the results of our proposed classifier in Section 6.5. A discussion of these results is provided in Section 6.6. Finally, conclusions are provided in Section 6.7.

6.2 Related work on ship risk profile

It is widely recognized that introducing the NIR, and thereby the ship risk profile has been beneficial to reducing the number of noncompliant ships [67, 166, 201, 202]. Nevertheless, some weaknesses have been identified [47, 48, 77, 176, 204–206]. We mention two of them, together with possible 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* [61]. In doing so, the model ignores any interactions between the factors. Here we remark that more complex models may consider more correlations, thereby improving performance [61, 207]. To this end, machine learning classifiers have been introduced that can automatically learn the weights and capture correlations between the factors. We provide two examples.

- A pipeline with a support vector machine and k-nearest neighbors have been used to find high-risk ships [61]. The support vector machine takes more complex (and non-linear) interactions into account and generalizes well, while k-nearest neighbors make the overall approach noise tolerant.
- A balanced random forest classifier has been used to predict ship detentions because only a tiny fraction of ships are detained [206].

The second weakness of the ship risk profile is that relatively static factors are used in risk assessment, meaning that the factors rarely change for a given ship. Indeed, many

datasets have been exploited that better reflect the current condition of a ship and hence will likely improve prediction. We mention four different datasets that have been used.

- Web scraping have been used to gather information from inspection reports [205].
- Company inspections have been used to enhance the ship risk profile [99].
- More historical information, such as times of changing flags and casualties in the last five years, have been proposed to add in the ship risk model [206].
- Information between different regimes should be more coherent, such that deficiencies and detentions in other regions can be used as well [99, 207].

The impact of the literature on our work is as follows. We read in the literature that it was strongly recommended to use additional data to come to a better prediction. We used port call data modeled as a cargo ship network. We mention the following four works on the cargo ship network, that have inspired us.

First, in 2010, the initial unveiling of a cargo ship network on a global scale was documented by Kaluza *et al.* [94]. According to their findings, the network had a *smaller diameter* (measuring 8) than expected for a randomly constructed network of equivalent size. Additionally, they discovered that the *average distance* separating any two ports across the globe was just 2.5.

Second, other researchers have found a diameter of only 7 and an average distance of 3.3 [113].

Third, the robustness of the cargo ship network has been studied by analyzing the transponder [151]. Different ship types were studied (oil tanker, container, dry bulk), and properties of these ship types have been reported for each sub-network derived from those ships. No measure of the distances in the network was reported, but a *density* (of ~ 0.02) similar to the first published cargo ship network was found.

Fourth, Van Veen (2020) analyzed the cargo ship network as derived from data of port calls [192]. Although the data was extracted only from journeys either departing or arriving at one of the Paris Memorandum of Understanding members, a diameter of 7 was found and an average distance of 2.49, similar to the reported values of other works.

In Section 6.3, we compare the properties of these networks to those of our cargo ship network. Ultimately, we *predict* noncompliance using a classifier with mobility patterns extracted from the cargo ship network (see Section 6.5). Our contribution is thus an approach that addresses the two weaknesses observed in the ship risk profile currently used by inspectorates: (1) manually adjusted weights and (2) relatively static factors.

6.3 Cargo shipping data

The chapter aims to classify ships' noncompliance using mobility data. The data used originates from two sources: (1) port calls (Subsection 6.3.1) and (2) inspections (Subsection 6.3.2). After collection, the port calls and inspections are merged (Subsection 6.3.3).

6.3.1 Port calls

The first data source, the port calls, contains notifications of cargo ships calling a port. The data contains only *calls to a port* participating in the Paris MoU and is accompanied by the following six pieces of information: (1) the International Maritime Organization (IMO) number — a unique identifier used in the maritime sector; (2) the port it calls to; (3) the date of arrival; (4) the duration that the ship is berthed; (5) the flag of the ship when it called; and (6) the ship risk profile (low, medium, high risk) computed when berthing. From this port call data, we reconstruct journeys that took place. A ship's journey goes from a departure port to an arrival port and has an associated travel time.

6.3.2 Inspections

The second data source, the inspections, provides information about ships with a deficiency. Also, we know whether such a deficiency has led to detention. Ships without deficiencies are assumed to be compliant because every ship should be inspected at least every three years at one of the ports participating in the Paris Memorandum of Understanding [56]. The inspection results are used as ground truth for our classifier. In Figure 6.1, we show the fraction of noncompliant ships that visit all countries. We observe that this fraction is very different across countries in Europe.

6.3.3 Merging port calls and inspections

Ships in these two datasets are linked using the IMO number. We select years occurring in data from both sources (2014–2018), resulting in over 3,000,000 calls from 28,416 cargo ships to a port in one of the thirty countries. Most of them, 97.3% (27,647 ships), did not change their flag during the years under consideration. Of these ships, the total number of ships with a white, gray, or black flag is 26,300, 672, and 675, respectively. Because only a tiny proportion of ships are flying a black or gray 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" (FOC) or a more trustworthy one with a better reputation [131]. In the data, 2.7% (1,347) of all ships changed their flag in 2014–2018. The distribution of flags over all countries is shown in Figure 6.2. We observe that most ships are registered in countries often identified as FOC, such as Panama and Liberia. Although difficult to observe, most ships are registered to Panama (2,904), Marshall islands (2,153), and Liberia (2,119), which are all known as typical FOC countries. In Figure 6.3, the fraction of noncompliant ships for each flag is shown. We observe that some black or gray flags are associated to a large fraction of noncompliant ships. The other way around, some of the white flag states have many noncompliant ships as well, such as the United States of America. Figures used in this section, can be downloaded at higher quality from our online repository [27].

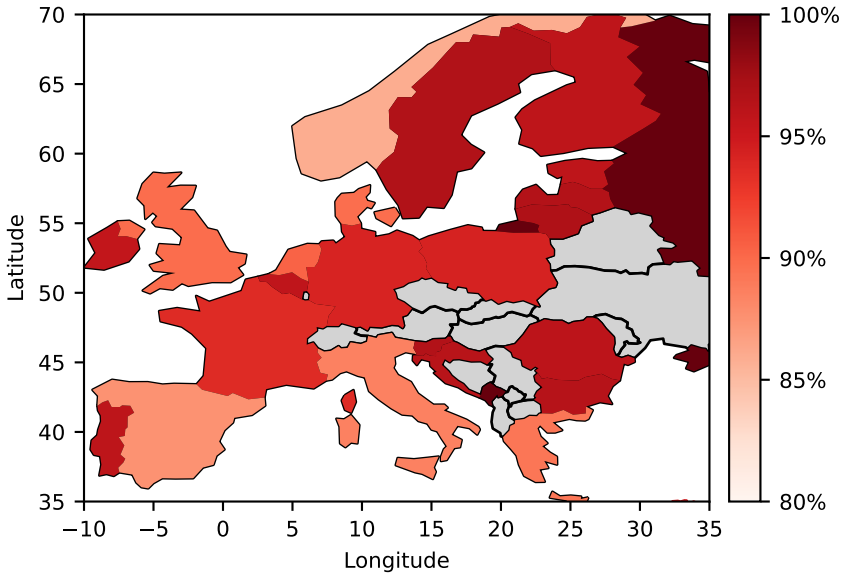

Figure 6.1: Fraction of ships being noncompliant per country. (Countries indicated in gray were not visited by a ship in the data.)

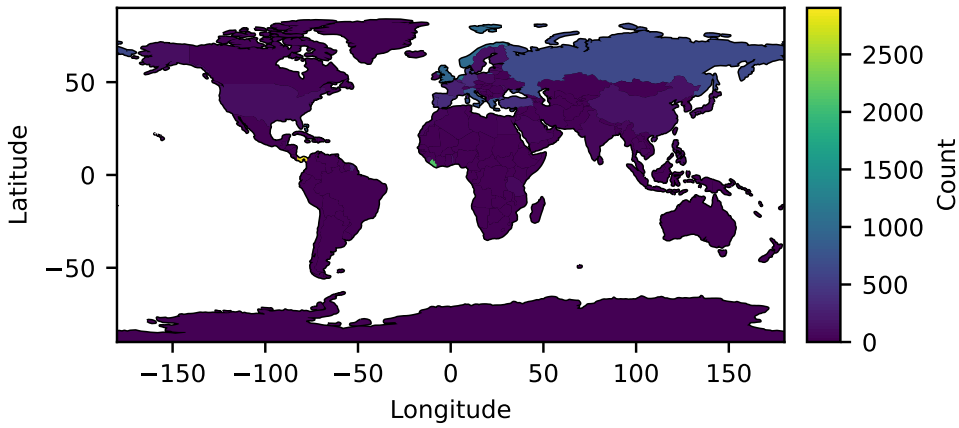

Figure 6.2: Number of ships registered to each country. (If multiple registrations for a single ship were observed, we use the most recent registration.)

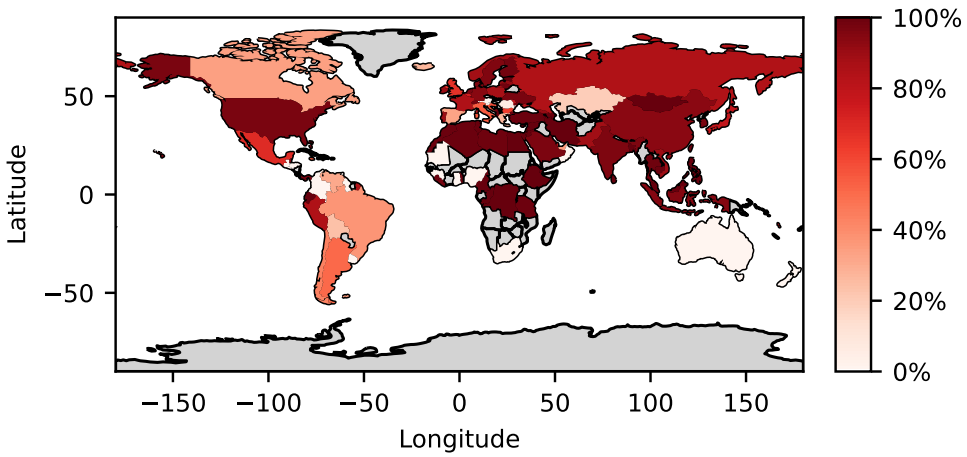

Figure 6.3: Fraction of ships for each flag state being noncompliant. (States without any ship registered to it, are indicated in gray.)

6.4 Chapter research methodology

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

In Subsection 6.4.1, we start by explaining the construction of the cargo ship network. We explain our approach to feature engineering, dealing with both the network and temporal features, in Subsection 6.4.2. Then, we discuss the classifier in machine learning in Subsection 6.4.3. We elucidate the fair model and explain the performance measures in Subsection 6.4.4. Finally, the fairness measures are explained in Subsection 6.4.5.

6.4.1 Cargo ship network

To obtain the structural importance of each port, we construct a cargo ship network. It is later used to characterize the behavior of ships. 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 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 Section 1.2). They help us understand whether our cargo ship network is, in fact, similar to earlier constructed networks of the same type. For each port, we obtain the following twelve structural importance 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 [60];
- (9) betweenness centrality and (10) weighted betweenness centrality [24, 59];
- (11) eigenvector centrality and (12) weighted eigenvector centrality [23].

These measures are used in Subsection 6.4.2 to engineer features that are provided to the machine learning classifier. We will now explain each of them.

- **Degree** of a node capture the number of routes (i.e., the number of edges connected to the node).
- **Strength** of a node capture the number of journeys connected to a port (i.e., the total weight of the edges connected to the node).
- **Closeness centrality** is equal to the reciprocal of the average shortest path distance from a node to all other nodes [60]. A more central node is closer to all other nodes and hence has a higher closeness centrality.
- **Betweenness centrality** is equal to the number of shortest paths between every pair of nodes that pass through to the node under consideration [59]. A node with high betweenness centrality is associated with playing an essential role in the network; disruption of this node will affect many shortest paths.
- **Eigenvector centrality** is determined using the eigendecomposition of the adjacency matrix [23]. High eigenvector values mean that the node is connected to many nodes with a high eigenvector centrality value.

With the latter three 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 train set (used to learn the classifier) and the test set (used to estimate the classifier's performance) should be independent. To prevent the data used to *construct* the network is also used in testing, we work with separate hold-out data to construct the network. Hence, we assign every ship $i \in I$ to 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 (later divided into train and test set by the cross-validation procedure, see Subsection 6.5.1).

6.4.2 Feature engineering

In $I_{\text{classification}}$, there are two different types of features that describe how ships behave: network features (see Subsection 6.4.2A) and temporal features (see Subsection 6.4.2B).

6.4.2A Network features

The network features aim to capture what type of ports a given ship visits. We obtain the network features in four steps.

Step 1. Determination of structural importance of each port. We characterize each ship's journey by the structural importance of the cargo ship network of both the departure and arrival ports. Only if the port is observed in the cargo ship network, the 12 structural importance measures (see Subsection 6.4.1) are determined. For each 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, and division). After this step, we have $12 \cdot 4 = 48$ values characterizing each journey.

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

Step 3. Aggregation. The model is ultimately provided with information about the individual ships' instances. 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. After that, we normalize 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. Normalization allows us to compare the distributions regardless of the number of journeys of a ship. In this way, we obtain $480 + 1 = 481$ features.

Step 4. Encode the missingness. In Step 1, we explained that the structural importance measures are only defined if the port was observed in the cargo ship network. The information that a port is missing in the network is informative for the classifier. Hence, we will encode this missingness, a common approach discussed in more detail in [138, 156]. We do so with two different 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.

6.4.2B Temporal features

The temporal features are computed from the duration of a ship's journeys and port berths. Anomalous short or long ship berths or journeys may be indicative of noncompliance. For example, short berths may lead to rushing through safety procedures, while significantly longer berths may indicate problems with the port authorities. We first make a histogram of each ship's observed journey and port berth duration values to preserve the estimated distribution of the berth durations and travel timing during aggregation. The histogram is made by splitting each ship's berth and journey durations into 10 equal-width bins. The boundaries of the bins are learned from (1) the port calls and (2) the journeys occurring in I_{network} to prevent information from leaking. In this way, $2 \cdot 10 = 20$ temporal features

are obtained. We sum all the values obtained for each ship of (1) the histogram of the berth duration and (2) the histogram of the journey duration and divide them by the total number of berths and journeys, respectively.

We have 483 network features and 20 temporal features, resulting in a total of 503 features describing each ship, represented by a vector x_i for some ship i.

6.4.3 Fair random forest classifier

We employ a machine learning model to perform the automated assessment of noncompliance. The goal of the model 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 $f\colon X \mapsto Z$ where $z_i \in Z$ is a score between 0 and 1. The positive instances, i.e., $y_i = 1$, indicate a noncompliant ship, and the negative instances a compliant ship. We may recall from Section 1.1 that in search of a particular type of fairness, we aim to reduce the classifier's dependency on sensitive features $s_i \in S$, 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* [157], which is a modified random forest classifier. In brief, a random forest classifier works as follows. A bootstrapped training data sample is taken for every tree in the forest. Then, a decision tree is grown by recursively doing three steps:

1. Select a sample from all features available.
2. Optimize a criterion (commonly the information gain) calculated on each sampled feature.
3. Split the node into two child nodes based on the optimization outcome.

For more details of the working of a random forest classifier, we refer the reader to [76].

Like other tree learning classifiers, random forest classifiers have some beneficial properties. We mention two of them. The first property is that their robust performance has been *confirmed* in different domains, meaning that a minimum of tuning is needed [76]. The second property is that the criterion considered does not have to be differentiable, in contrast to many other classifiers, allowing to introduce the SCAFF criterion (see later on). Both properties together allow us to use a specifically designed criterion, called Splitting Criterion Area under the curve for Fairness (SCAFF) [157]. The criterion ensures that different labels are separated and the sensitive class remains mixed. We first give the definition and then explain the formulas.

$$\text{SCAFF}(Z, Y, S, \Theta) = (1 - \Theta) \cdot \text{AUC}_{\text{Y}}(Z, Y) - \Theta \cdot \text{AUC}_{\text{S}}(Z, S),$$

with AUC_{Y} a value in the closed interval $[0, 1]$:

$$\text{AUC}_{\text{Y}}(Z, Y) = \frac{\sum_{i=1}^{y_+} \sum_{j=1}^{y_-} \sigma\left(Z_i, Z_j\right)}{y_+ \cdot y_-} \quad \text{with} \quad \sigma(Z_i, Z_j) = \begin{cases} 1, & \text{if } Z_i > Z_j \\ \frac{1}{2}, & \text{if } Z_i = Z_j \\ 0, & \text{otherwise} \end{cases},$$

where y_+ and y_- mark the number of positive and negative instances. An AUC_Y value of 0.5 suggests random classification while $\text{AUC}_\text{Y} = 1$ indicates a perfect classifier. The AUC_S considers the sensitive feature as the positive class. It is defined as follows:

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

with $\sigma\left(Z_i, Z_j\right)$ defined exactly the same as for AUC_Y. The measure is closely related to strong demographic parity [93]. For $\text{AUC}_\text{S} = 0.5$, corresponding to a strong demographic parity of 0, the split in the node is made regardless of the values of the sensitive features, meaning equality of outcome. A value of $\text{AUC}_\text{S} = 1$, corresponding to a strong demographic parity of 1, is the worst score possible since, in that case, the classifier can predict the sensitive feature perfectly. The orthogonality parameter, $\Theta \in [0, 1]$, allows to balance the performance-fairness trade-off [96].

At a value of $\Theta = 0$, the fair random forest classifier optimizes solely for performance and does not consider any fairness. Hence, it corresponds, in that case, to the ordinary random forest classifier. At a value of $\Theta = 1$, the classifier optimizes fairness and neglects any performance. We refer the reader for more details on the fair random forest classifier to [157].

6.4.4 Performance measures

The classifier's performance can be determined 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 as negative ($\hat{y}_i = 0$). Threshold-free metrics have the advantage that they do not require this explicit cut-off and instead consider the ranking imposed by the scores of the classifier. The three threshold-dependent performance metrics are (1) precision, (2) recall, and (3) the harmonic mean of those two, the F_1 score. The threshold-free performance metric used in this work is the AUC_Y (see the previous section).

6.4.5 Fairness measures

Similar to the performance measures, fairness with respect to the sensitive group can also be quantified by two metrics: *threshold-dependent* and *threshold-free* metrics.

First, we report on the *threshold-dependent* metrics by (1) the precision and (2) the recall for the following two groups: (a) ships with a white flag and (b) ships with a non-white flag. A significant difference between these two groups indicates an unfair outcome of the model, which we aim to avoid.

Moreover, we report also on the *threshold-dependent* metrics by (3) demographic parity and (4) equalized odds [75]. These latter two measures consider the difference

in performance measures between the two groups, i.e., ships with a white flag and a non-white flag.

The *demographic* parity measure, denoted as ϵ_{parity}, sets an accepted maximum on the absolute difference between the positive prediction rates of the two groups. It is mathematically represented as $\left|P\big(\hat{Y}=1|S=1\big)-P\big(\hat{Y}=1|S=0\big)\right| \leq \epsilon_{\text{parity}}$. Lower values of ϵ_{parity} signify more similar outcomes to the sensitive and non-sensitive groups, indicating fairer predictions.

The *equalized odds* metric, denoted as ϵ_{odds}, imposes a maximum accepted difference on the equality of opportunity in a supervised learning setting. It is expressed as

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

Reduced values for ϵ_{odds} suggest greater equality of opportunity for the sensitive and non-sensitive groups, thus more fair predictions.

Finally, we have also reported on the *threshold-free* fairness measures (denoted by AUC_S), for which we refer to Subsection 6.4.3.

6.5 Results

The section starts with our experimental setup in Subsection 6.5.1. Then, we continue analyzing the cargo ship network in Subsection 6.5.2. In Subsection 6.5.3, we evaluate the baseline ship risk profile performance. Subsequently, we report on the performance of the non-fair random forest classifier in Subsection 6.5.4 and the fair random forest classifier (announced in Subsection 6.4.3 as our preferred choice) in Subsection 6.5.5. In Subsection 6.5.6, we report on the effects of the orthogonality parameter. Finally, in Subsection 6.5.7, we describe the effects of the threshold quantile in combination with the orthogonality parameter.

6.5.1 Experimental setup

In our experimental setup, we use five-fold nested cross-validation with stratified sampling [39]. The inner folds select the best parameter set for that specific outer fold. The considered parameters are all combinations of the selected values for the depth of each tree ($\{1,2,\ldots,10\}$) and the number of bins (10 or 2) used in discretization for continuous variables. Hence, there are $10\cdot 2=20$ candidate sets of parameters in each outer fold. The mean and standard deviation of the classifier's performance is 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 [27]. It uses several open-source Python packages. Specifically, scikit-learn [149], SciPy [193], and Pandas [117] 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 open source as well [152], making extensive use of the CVXpy package for optimizing SCAFF [2]. For analyzing the cargo ship network, we used the NetworkX package [72]. The C++ library teexGraph was used to determine the diameter of the network [185]. The packages used for visualization and all other dependencies and supportive software versions can be found at [27].

6.5.2 Cargo ship network

A quite "overwhelming" visualization of the cargo ship network obtained is shown in Figure 6.4. Still, we only show ports in Europe because we are interested in predicting the risk for ships that arrive in Europe. From the figure, we can learn the following four properties.

1. A GC connects virtually all ports.
2. Only a few ports have high strength, as indicated by the yellow color, of which (1) Puttgarden (Germany), (2) Rotterdam (Netherlands), and (3) Algeciras (Spain) have the highest strength.
3. Two different types of ports can be distinguished: (1) ports that are well-connected (e.g., ports in Germany, Netherlands, and Belgium), and (2) ports that are more in the network's periphery (e.g., Iceland and the Azores).
4. Some ports are connected by thick lines, indicating an edge with a high weight. The nodes connected by these edges are likely to have a high weighted betweenness centrality because the failure of such nodes would cause other shortest paths to run through edges with less weight.

In Table 6.1, we provide numeric information on sizes, relations, and distances. In the first column, we show our work's nine common properties of cargo ship networks. In the second column through the sixth column, we provide values for the properties of our network and four similar cargo networks observed in literature [94, 113, 151, 192]. We compare these properties to understand whether our 10% sample used to compute port features is representative. From Table 6.1, we see that although very different numbers of nodes and edges are reported in these works, the measures such as *density*, *diameter*, and *clustering coefficient* are similar. Hence, we may conclude that the constructed cargo ship network can extract mobility patterns for our ship compliance classifier in a sensible way.

6.5.3 Performance of the baseline ship risk profile

The confusion matrices for the baseline ship risk profile are shown separately for the white and non-white flags in Figure 6.5. Together with Table 6.2, where we show the *calculated performance* and *fairness measures*, they provide information on the performance of the baseline ship risk profile. We remark that low or medium risk ships are predicted as compliant.

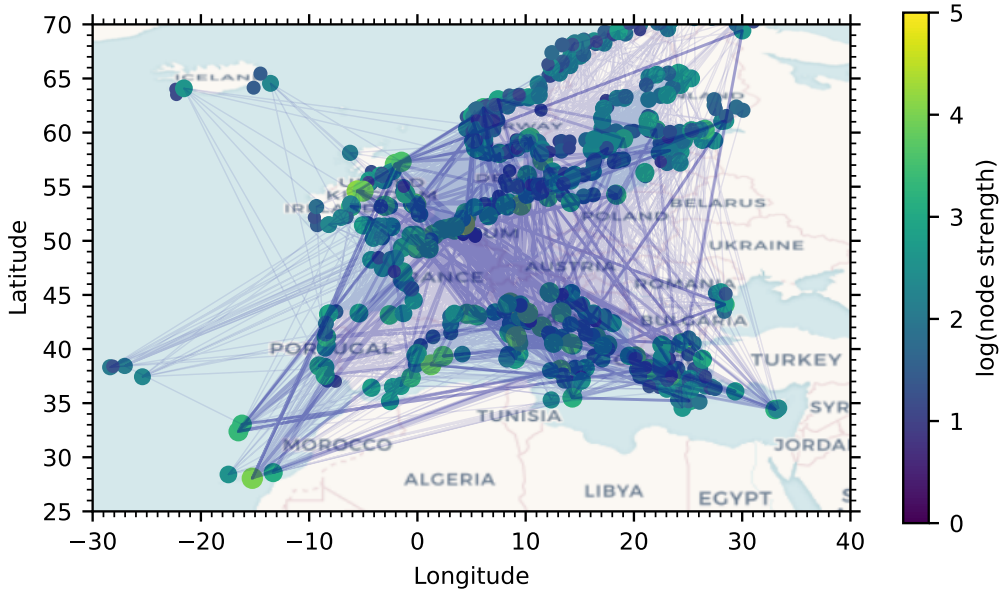

Figure 6.4: The considered cargo ship network. (Nodes are colored by their strength. Thicker edges mark busy routes. The figure is generated using OpenStreetMap data.)

Table 6.1: Summary statistics of considered cargo ship networks.

Property	*This work*	[192]	[151]	[113]	[94]
Directed	Yes	Yes	No	No	No
Number of nodes	1,459	728	1,488	439	951
Number of nodes in GC	1,445	726	–	–	935
Number of routes	28,653	18,142	17,135	2,331	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

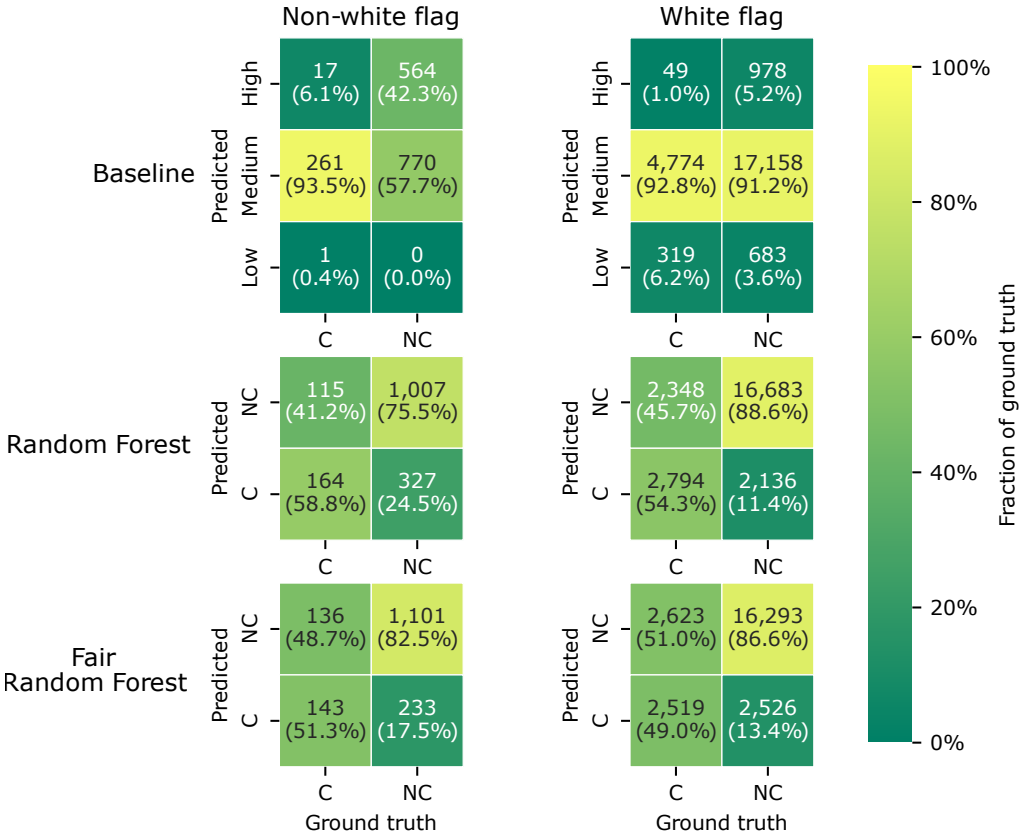

Figure 6.5: Confusion matrices (shown for both the white and non-white flagged ships). (Baseline model is the ship risk profile currently in use. C and NC mark Compliant and NonCompliant ships, respectively. The percentages [and color coding] are stratified based on the ground truth.)

Table 6.2: Performance (precision, recall, and F_1, and AUC_Y) and fairness (demographic parity and equalized odds, and AUC_S) 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

Below, we make four observations from Figure 6.5 and Table 6.2.

First, we note that virtually no ship flying a non-white flag gets a low-risk profile (see left upper corner), indicating that the baseline model uses the flag to a large extent.

Second, most ships (90%) are classified as medium risk (see baseline predicted medium). Only a small fraction, $(261 + 4774)/(261 + 770 + 4774 + 17158) = 22\%$, is compliant.

Third, a smaller fraction, $(17 + 49)/(17 + 49 + 564 + 978) = 4\%$, is compliant from the ships with a high-risk profile. It results in high precision for the baseline model. However, the recall is relatively low as many ships with a medium risk profile are also noncompliant.

Fourth, unexpectedly, ships with a white flag with a low or medium risk profile are more noncompliant than ships with a non-white flag. It also results in a low value of the AUC_Y value of only 0.543 ± 0.006 (see Table 6.2). Hence, we may conclude that using the data from 2014–2018, we cannot predict compliance with the baseline ship risk profile. It follows that the model is *quite unfair*. In particular, we observe a significant difference in the F_1 metric for the white and non-white group, resulting in high values for ϵ_{parity} and ϵ_{odds} (see Table 6.2). There is a strong correlation between the sensitive feature, i.e., the ship flag, and the scores of the model with $\text{AUC}_\text{S} = 0.672 \pm 0.010$ (see Table 6.2).

6.5.4 Performance of the random forest classifier

The confusion matrices of the random forest classifier are also shown in Figure 6.5. In Table 6.2 we report the performance and fairness metrics (column 3). Below, we make five observations. *First*, we observe that more ships are predicted correctly compared to the baseline model. *Second*, the recall is higher, meaning many actual positives are predicted. The table also shows decreased precision, indicating that many compliant ships are predicted as noncompliant. *Third*, the harmonic mean of the recall and precision, the F_1 measure, is higher than in the baseline model, indicating that the random forest classifier outperforms the baseline model. *Fourth*, the AUC_Y measure, shows a high value of 0.814 ± 0.004, supporting also that the random forest classifier outperforms the baseline model. It implies that we accurately can assess the ship noncompliance in an automated fashion with a random forest classifier using behavioral data. *Fifth*, the confusion matrices (Figure 6.5) 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 results in a higher recall for ships with a white flag.

Finally, we remark that the prediction by the random forest classifier is much more fair compared to the baseline model. In conclusion we remark that the random forest classifier does not use the flag as a feature, meaning that using only behavioral data thus makes the model more fair.

6.5.5 Performance of the fair random forest classifier

The confusion matrices of the fair random forest classifier are also shown in Figure 6.5. In Table 6.2 we report the performance and fairness metrics (column 4). Below we list our three observations. *First*, from the confusion matrices in Figure 6.5 and the performance and fairness metrics in Table 6.2, we observe that the fair random forest classifier has comparable true positive and true negative rates amongst ships flying a white and non-white flag, with only a small cost in predictive performance. *Second*, the F_1 performance measure drops only for the ships flying a white flag, so the difference between the two groups becomes minimal. *Third*, the demographic parity and equalized odds measures decrease when using a fair random forest classifier, suggesting that the classifier improved fairness.

6.5.6 The effect of the orthogonality parameter

Before drawing any conclusion, we show the effect of the orthogonality parameter (Θ) in more detail (see Figure 6.6). Below we list our six observations. *First*, the top left figure (Figure 6.6A) shows that the AUC_Y measure is only weakly influenced by a broad range of values for the orthogonality parameter, meaning that overall, we can reliably ensure equality of outcome while maintaining acceptable performance. *Second*, an orthogonality value of 0.7 appears to give the best trade-off between performance and fairness in our work, with a performance of $\text{AUC}_\text{Y} = 0.776 \pm 0.008$ and fairness of $\text{AUC}_\text{S} = 0.538 \pm 0.011$. *Third*, the performance can be further improved (although slightly, to $\text{AUC}_\text{Y} = 0.814$), but only at decreased equality of outcome and vice versa.

Then we will closely investigate Figure 6.6B, where the two fairness measures decrease monotonically at increasing orthogonality values. *Fourth*, we make one observation that the extreme value of $\Theta = 1$, they are zero, but at this value, the predictive performance is also deficient, as can be observed in Figure 6.6A.

Subsequently, in Figure 6.6C and Figure 6.6D, we make two observations. *Fifth*, we observe that the precision and recall for ships flying a white and non-white flag have only minor 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 vessels with a white flag. *Sixth*, the threshold was set to $t = 0.34$ so that $P(Z \geq t)$ equals $P(Y = 1)$. This threshold is also used to calculate the confusion matrix shown in Figure 6.5.

In conclusion, we remark that the threshold t is essential, as it determines how many ships are noncompliant. Higher threshold values result in fewer ships that are predicted as noncompliant. Therefore, we define the threshold quantile Q_t so that $P(z \geq t)$ equals the threshold quantile.

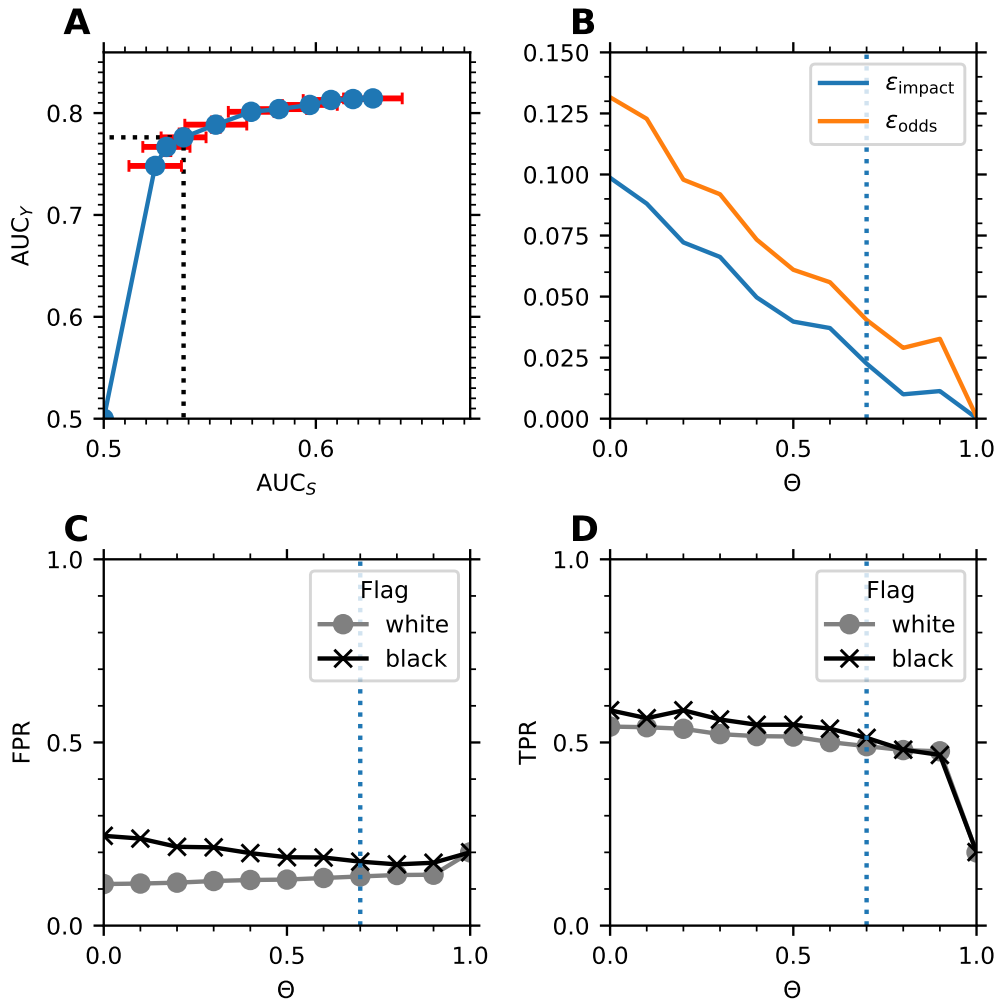

Figure 6.6: Performance and fairness of proposed ship selection classifier: (**A**) The performance of the fair random forest classifier for different values of the orthogonality. (**B**) The fairness performance is measured in demographic parity and equalized odds for different values of Θ. (**C**)–(**D**) The performance measured in precision (**C**) and recall (**D**) for different values of Θ, separated for ships flying a white and non-white flag.

6.5.7 The effect of the orthogonality and threshold quantile together

Finally, in Figure 6.7, we show the effect of the orthogonality and the threshold quantile on the selected threshold-dependent fairness measures. Below we list our three observations. *First*, 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. *Second*, for lower values of the orthogonality, we observe that the model's fairness is worst when the threshold quantile is near 0.5. This result is expected (see below). *Third*, at other values of the threshold quantile, the performance for both groups is low, leading to a slight difference between the groups. Even at these "bad" choices for the orthogonality and threshold quantile, the values of the demographic parity measure and the equalized odds measure are still lower than observed for the baseline ship risk profile.

In conclusion, from these results, we may state that the fair random forest classifier *effectively reduces bias* towards a ship's flag for wide ranges of the used threshold and orthogonality. It answers Research question 5.

6.6 Discussion on limitations

This section discusses two limitations of our proposed classifier.

First, the ground truth might be biased toward the flag and the inspector's background [67]. The problem is that different inspectorates assess compliance differently for similar ships. The difference in assessment leads to inequality between ports and

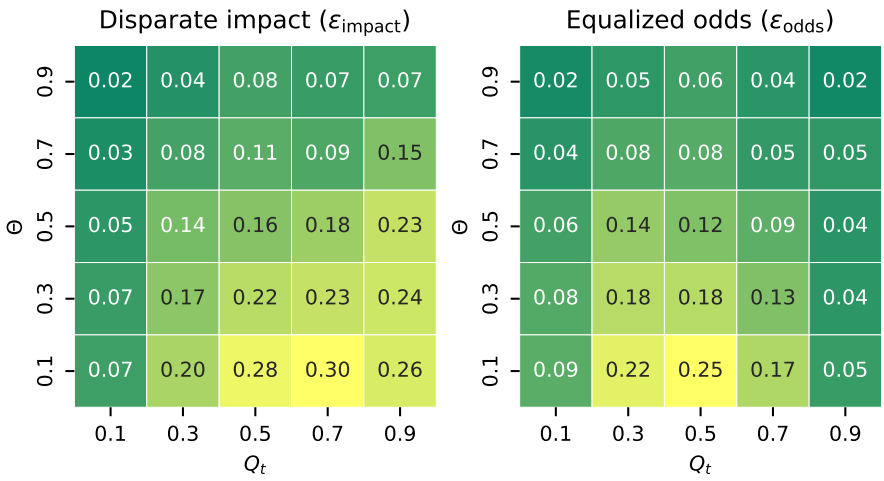

Figure 6.7: Fairness measures evaluated on the proposed classifier (as a function of the threshold quantile Q_t and orthogonality Θ).

so-called port-shopping. Port-shopping means that a noncompliant ship decides to go to another port solely because the inspection regime favors noncompliant vessels. In this way, the ship yields a lower risk profile. Port-shopping seriously influences our model since the ground truth data is unjustly positive for such noncompliant ships. The mission of the Paris MoU is to avoid this kind of competition between ports [147]. Hence, as a remedy, the inspection country could be added as a sensitive feature in future work, reducing the correlation between the inspectorate and the inspection outcome.

Second, we consider Goodhart's law, commonly formulated as: "When a measure becomes a target, it ceases to be a good measure" [181]. It applies to any ship risk model because ships are incentivized to get a *low-risk* profile. In the baseline ship risk model, a better risk profile could be achieved by *changing the administrative property* of the ship. In our fair random forest classifier endowed with orthogonality and threshold quantile setting, ships would need to change their behavior to get a better score, which is more complicated than merely changing administrative properties.

6.7 Chapter conclusions

The present research answers Research question 5: "How can ship behavior be utilized to enable smart inspection of cargo ships?" We devised an *accurate*, *automated*, *fair*, and *interpretable* assessment of ship risk, enabling smart inspection of cargo ships. This study has led to two conclusions.

Conclusion 1: We can offset the confirmation bias in historical inspection data using a fair random forest classifier. Experimental results indicated that the disparate impact and equalized odds measures improve significantly the assessment. This is regardless of chosen parameters, meaning that the constructed classifier works well.

Conclusion 2: The performance of our approach provided with behavioral data is $\text{AUC}_\text{Y} = 0.776 \pm 0.008$, which improves on the $\text{AUC}_\text{Y} = 0.543 \pm 0.006$ of the ship risk profile currently in use.

All in all, our *final conclusion* is that our work will support global efforts to minimize risks associated with maritime transport by conducting more targeted inspections. More generally, we have shown how ubiquitous mobility information can perform inspections to be better and more fair than so far. Finally, we believe that the devised approach may apply to inspection applications broader than port state control.

Chapter outlook

Below we provide four directions of future research.

First, a natural continuation of this work is to (with the help of domain experts) determine (1) what behavior is often associated with high risk, and subsequently (2) how we can reduce riskful behavior.

Second, a direction for future work is to consider higher-order effects in the cargo ship network [170]. Building a higher-order network allows for a more accurate representation of the underlying complex system, which 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 [170].

Third, we may investigate to what extent the temporal aspect of the network can be exploited to obtain a better, more accurate centrality measure that captures the true, time-aware structural importance of the ports [172].

Fourth, we may investigate to what extent the research under the third direction will result in an even better-performing classifier for the task at hand.