Electronic Research Archive

http://www.aimspress.com/journal/ERA

ERA, 31(1): 251–272.
DOI: 10.3934/era.2023013
Received: 13 September 2022
Revised: 17 October 2022
Accepted: 19 October 2022
Published: 28 October 2022

Research article

Data analytics in transport: Does Simpson's paradox exist in rule of ship selection for port state control?

Simon Tian[1] and Xinyi Zhu[2,*]

[1] School of Economics and Management, Wuhan University, Wuhan 430072, China
[2] Sino-US Global Logistics Institute, Shanghai Jiao Tong University, Shanghai 200000, China

* **Correspondence:** Email: tianxuecheng@whu.edu.cn, zxy-logistics@sjtu.edu.cn.

Abstract: Although previous studies have applied artificial intelligence techniques to improve the accuracy and efficiency of ship selection in port state control (PSC) inspections, the new inspection regime (NIR) is still in effect and widely adopted by PSC authorities in the Tokyo Memorandum of Understanding to select ships for inspection. It considers seven features, and each candidate value of a certain feature is assigned a fixed weighting point. The sum of the weighting points of these seven features determines the risk level of a ship. The assumption behind the NIR is that ships with values attached with higher weighting points should have more deficiencies. However, this paper finds that Simpson's paradox may exist for this assumption; that is, the average number of deficiencies of ships with values attached with higher weighting points is lower than that of ships with values attached with lower weighting points. Therefore, this paper examines the plausibility of the NIR's weighted-sum method and further explores which feature flips the effect. Finally, we arrive at the conclusion that the features selected by NIR are coupled with each other, so we should not use a simple weighted-sum method to determine the risk level of a candidate ship. Based on the results, we further provide suggestions for PSC authorities with respect to the improvement of the ship selection scheme of NIR.

Keywords: transportation; port state control; ship selection scheme; Simpson's paradox; data analytics

1. Introduction

Oceangoing shipping transports a significant amount of different goods across the globe, which supports the development of the global economy [1–8]. In these years, sustainable shipping has also become an important issue because of the air emissions released by ships, which adversely impact the marine environment [9–13]. To guarantee maritime safety and protect the marine environment, PSC is an international regime to inspect foreign visiting ships. It is designed to ensure that foreign visiting ships are seaworthy and comply with required international conventions, such as the international convention for the safety of life at sea (SOLAS) and the international convention for the prevention of pollution from ships (MARPOL). When a ship is selected to be inspected, the port state control officer (PSCO) first conducts an initial inspection, including the first impression of the ship, certificate check, and walking around to check the overall ship conditions. During a PSC inspection, conditions that do not comply with the relevant conventions are denoted as deficiencies. When the deficiencies identified are too many or severe, the PSCO will detain the ship until these deficiencies are rectified.

To achieve uniform and efficient PSC inspections, regional Memorandum of Understandings (MoUs) on PSC are established through cooperation among their members. The goal of MoUs on PSC is to verify that the foreign visiting ships meet the international conventions' requirements through a harmonized system of PSC, which allows for information sharing [14]. By the end of 2018, nine MoUs on PSC had been signed worldwide. And the inspection results, including the deficiencies identified and the detention outcome, combined with ship information, are recorded in the database of MoUs on PSC.

One of the essential issues faced by PSC authorities is how to select ships for PSC inspections. Port states recognize that inspecting all foreign visiting ships would be impractical due to the resources it would take and unnecessary because not all ships are substandard. Therefore, port states started to select foreign visiting ships to inspect according to the features of ships. Taking Tokyo MoU as an example, it introduced a ship selection scheme in 2014, namely NIR, to evaluate the risk level of one ship, as shown in Table 1 [15]. It considers seven features related to the characteristics and historical inspection records of a ship, including ship type, ship age, ship flag performance, ship recognized organization (RO) performance, ship company performance, the number of deficiencies within the previous 36 months, and the number of detentions within the previous 36 months. Each candidate value of a certain feature is assigned a fixed weighting point, and a ship's risk level is determined by the sum of seven features' weighting points. Based on the total points, all ships are divided into three types: high-risk ship (HRS) (whose total weighting points is at least 4), standard-risk ship (SRS) (whose total weighting points is at most 3 and who does not meet all the criteria of low-risk ships), and low-risk ship (LRS) (whose total weighting points is at most 3 and who meets all the criteria for low-risk ships, including white ship flag performance, high ship RO performance, high ship company performance, 5 or fewer deficiencies within the previous 36 months, and no detention within the previous 36 months); this scheme is easy to understand and implement. Thanks to the implementation of the NIR, maritime security, pollution prevention, and working conditions have all been improved [16].

Regarding the weighted-sum method of the NIR, the assumption is that ships with values attached with higher weighting points should have more deficiencies. For example, ships in high-risk types (i.e., chemical tankers, gas carriers, oil tankers, bulk carriers, etc.) are supposed to have higher numbers of deficiencies and thus have more chances to be inspected. However, by analyzing the dataset containing PSC inspection records that we collect (more information with respect to this dataset will be introduced in Section 3.2), we find that foreign visiting ships in low-risk ship types have a higher average number of deficiencies (4.02), while ships in high-risk ship types have a lower average number of deficiencies (3.77). A possible explanation for this finding is that those ships in high-risk types are registered under white flags to evade inspection, or the ages of these ships are young. Because the total weighting points of ships under white flags and at young ages are relatively low, even if those ships are in high-risk types, they are not likely to be inspected. Therefore, selected ships in high-risk types might have lower numbers of deficiencies compared with ships in low-risk types. This finding indicates that the NIR's weighted-sum method, which is based on the assumption that the values of the selected seven features are linear to the risk level of ships (i.e., the number of deficiencies) might not be reasonable. If the weighted-sum method does not consider the correlations of pairwise features among selected features, its effectiveness might be compromised. Therefore, the weighting method should not be established based on the linear total score of all considered features but consider a more comprehensive manner, such as considering the compound influence of pairwise features.

Table 1. The weighted-sum method of the NIR of Tokyo MoU.

Features	High-risk Value	Weighting Points	Low-risk Value	Weighting Points
Ship type	Chemical tanker, gas carrier, oil tanker, bulk carrier, passenger ship, container ship	1	Other types	0
Ship age	All types with age > 12y	1	All types with age ≤ 12y	0
Ship flag performance	Black	1	Grey/white	0
Ship RO performance	Low/very low	2	High/medium	0
Ship company performance	Low/very low/no inspections within the previous 36 months	2	High/medium	0
Deficiencies within the previous 36 months	Number of inspections which recorded over 5 deficiencies	Number of inspections which recorded over 5 deficiencies	Number of inspections which recorded below 5 deficiencies	0
Detentions within the previous 36 months	3 or more detentions	1	2 or fewer detentions	0

The paper aims to investigate the correlations of pairwise features among selected features of the NIR and further investigate the plausibility of the NIR's weighted-sum method. We require that when we classify ships according to the values of a certain feature, the values of the remaining features of ships should be identical. According to the NIR, the average number of deficiencies of ships in high-risk values of a certain feature is assumed to be higher than the average number of deficiencies of ships in low-risk values of that feature. However, if the relationship reverses (i.e., the average number of deficiencies for ships in high-risk values is lower than that for ships in low-risk values) when the values of remaining features are identical, a paradox with respect to the NIR appears, which is termed Simpson's paradox. If Simpson's paradox exists, we further explore which feature flips the effect. By investigating Simpson's paradox and analyzing the causes, we answer the question of whether it is reasonable to follow NIR's weighted-sum method to select ships for inspection.

The contributions of the paper are as follows. First, the methodology used in our research identifies possible paradoxes of the NIR by analyzing a PSC dataset. To the best of our knowledge, identifying Simpson's paradox by finding correlations of pairwise features among selected features of the NIR has not been considered in previous relevant research. Therefore, our research is the first one to examine the plausibility of the NIR. Second, based on the correlations of pairwise features revealed in this paper, we conclude that the values of selected features of NIR are nonlinear to the risk level of ships. Different from previous studies that propose machine learning (ML) models to invent a brand-new ship selection scheme that requires great technological transformations, we mainly focus on diagnosing the intrinsic issues with respect to the current NIR, leading to managerial insights and suggestions that are easier to operate and implement for efficient and effective ship selections in PSC.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 presents a detailed description of our methods and materials. Section 4 shows the results. Finally, Section 5 concludes this study.

2. Literature review

2.1. Studies on PSC inspection

A recent literature review classified the large body of literature on PSC into four main categories: targeted features influencing PSC inspection results, inspected ship selection scheme, PSC inspection effects, and suggestions for MoU management [17]. In this study, we focus on the literature about features influencing PSC inspection results and inspected ship selection schemes.

For targeted features influencing PSC inspection results, several studies arrived at the conclusion that generic features, including ship age, ship flag, and ship type, were main determinants of ship deficiencies and detention [18–20]. As for non-generic features, Knapp and Franses [21] claimed that inspection areas and different backgrounds of inspectors would influence the inspection results. Ravira and Piniella [22] and Graziano et al. [23] both concluded that the professional profile of PSC inspectors might affect inspection results. These papers all used statistical models to analyze data and find out the determinant features of PSC inspection results.

For ship selection scheme, relevant studies used ML models to select ships to be inspected or predict the number of deficiencies and detentions of foreign visiting ships. Xu et al. [24] introduced a risk assessment system based on a support vector machine (SVM) to classify foreign visiting ships as either high-risk or low-risk according to the target factors. Yang et al. [25] combined the Bayesian network model with the game model between PSC port authorities and ship owners to present an optimal PSC inspection scheme. In addition, several studies developed new ship selection models to predict the deficiencies and detentions of ships. Wang et al. [26] proposed a BN model to predict the number of ship deficiencies and compared it with the current NIR's ship selection scheme in the Tokyo MoU, demonstrating the superiority of the BN model. Based on the static risk factors adopted by the NIR, Dinis et al. [27] developed a BN-based ship risk assessment model and conducted a quantitative assessment of the predictive validity of the model using historical PSC inspection records. Yan et al. [28] proposed a random forest-based model to predict the probability of ship detention. In a recent study, Yan and Wang [29] further proposed an anomaly detection model for ship detention prediction. These studies used ML models for ship selection in PSC inspections, which can identify substandard ships more efficiently and accurately.

2.2. Studies on the Simpson's paradox in operations management

Simpson first described the paradox in 1951 [30]. It is a statistical phenomenon that causes a potential bias in certain data analyses. The paradox occurs when a relationship between two variables reverses when a third variable, called a confounding variable, is introduced. The literature on Simpson's paradox has focused on explaining the phenomenon, specifying its magnitude [31,32], the conditions where it vanishes [33], and its frequency [34]. The implications of Simpson's paradox on managerial decision-making have been considered in operations management. Sunder [35] considered the paradox in the context of the allocation of indirect costs in the logistics system, which sharpened our intuition through deriving new rules of thumb. Mehrez et al. [36] discussed the paradox in the case of efficiency measures for firms or decision-making units. This research reminded us to exercise caution when developing models to deal with different technologies. Curley and Browne [37] observed the paradox in the background of on-time rate for delivery companies, where the judged relationship between two variables (e.g., company and performance) differs depending on whether that relationship is viewed within subcategories of a third variable (e.g., package size) or in the aggregate. Melumad and Ziv [38] looked at the relationship between product quality and increased production and found that under certain conditions, each individual firm's average quality decreased while the overall market average quality increased.

In summary, relevant studies have proposed ML models to improve the accuracy and efficiency of ship selection for PSC. Nevertheless, the NIR (i.e., weighted-sum method) is still in effect for Tokyo MoU. Although most existing studies propose new methods, they do not investigate the internal reasons for the drawbacks of the current weighted-sum method of the NIR. Therefore, our research studies the correlations between selected features of the NIR and investigates whether there are paradoxes with respect to the NIR, aiming to diagnose for the current scheme and provide suggestions for PSC authorities.

3. Methods and materials

3.1. Methods

In this article, we aim to investigate the correlations of pairwise features among selected features of the NIR by studying whether ships with high-risk values of a certain feature have more deficiencies. To achieve this aim, we compare the average number of deficiencies of two categories divided by a splitting value of a certain feature. To examine the effect of a certain feature, we require that when we classify ships according to the values of a certain feature, the values of the remaining features of ships in these two subcategories should be identical. For example, ship age and ship flag performance are two features that affect the overall points of a visiting ship. If we first divide ships into two categories according to their ship flag performance, the total number of deficiencies, the total number of ships, and the average number of deficiencies, of the two categories are shown in Table 2. To examine the effect of ship flag performance on the number of deficiencies, we then require that the ships in these two categories have an identical range of ship age (i.e., above 12 or below 12). Therefore, by further stratifying the data, because we can divide ship age into two different range levels, we could get the corrected data, namely four subcategories, as shown in Table 3. Then pairwise comparisons of ships under the identical range level of age but with different values of the ship flag performance are conducted.

Table 2. The total number of deficiencies, the total number of ships, and the average number of deficiencies divided by ship flag performance.

Ship flag performance	Total deficiencies	Total ships	Average deficiencies
Black	$q_1 + q_3$	$p_1 + p_3$	$\frac{q_1 + q_3}{p_1 + p_3}$
Grey/white	$q_2 + q_4$	$p_2 + p_4$	$\frac{q_2 + q_4}{p_2 + p_4}$

Table 3. The total number of deficiencies, the total number of ships, and the average number of deficiencies divided by ship age and ship flag performance.

Ship age	Ship flag performance	Total deficiencies	Total ships	Average deficiencies
>12	Black	q_1	p_1	$\frac{q_1}{p_1}$
	Grey/white	q_2	p_2	$\frac{q_2}{p_2}$
⩽12	Black	q_3	p_3	$\frac{q_3}{p_3}$
	Grey/white	q_4	p_4	$\frac{q_4}{p_4}$

As shown in Tables 2 and 3, assume that we obtain the following results:

$$\frac{q_1 + q_3}{p_1 + p_3} > \frac{q_2 + q_4}{p_2 + p_4}, \tag{1}$$

$$\frac{q_1}{p_1} < \frac{q_2}{p_2}, \tag{2}$$

$$\frac{q_3}{p_3} < \frac{q_4}{p_4}, \tag{3}$$

where Eq (1) indicates that the average number of deficiencies of ships under the black flag state is higher than that of ships under the white flag and grey flag when we do not require that the ships in each category have an identical range of ship age. However, Eqs (2) and (3) indicate that the average number of deficiencies of ships under the black flag states appear to be smaller than that of ships under the white flag and grey flag when we require that the ships in a subcategory should age below 12 or above. It means the relationship between the average number of deficiencies of ships and ship flag performance reverses after we divide the dataset into four subcategories by introducing a confounding feature ship age. The phenomenon observed is generally termed Simpson's paradox. In this paradox, we assume that the ship flag is the basic categorical feature, the average number of deficiencies is the outcome, and the ship age is the introduced categorical confounding feature that causes the paradox. The reason for this Simpson's paradox may be that the ages of ships under the black flag state are younger, so selected ships under the black flag state have lower average numbers of deficiencies.

Since the NIR considers seven features, the influence of ship selection features on the ship conditions (deficiencies and detentions) is complex, and the categorical confounding features that may cause the paradox might be a combination of the other features. Therefore, we consider situations where the ships in the two subcategories only have different values in one basic categorical feature, and the values of the remaining introduced categorical confounding features are identical. The classification of values of seven features according to the NIR is shown in Table 4. Then, we choose one feature as the basic categorical feature, group ships with identical values of remaining features into different subcategories, and conduct the analysis with subcategory data. The data analysis procedures for calculating the average number of deficiencies are shown in **Algorithm 1**, where the notation used is listed in Table 5.

Table 4. Classification of seven features.

Feature	High-risk value	Low-risk value
Type of ship	Chemical tanker, Gas Carrier, Oil tanker, Bulk carrier, passenger ship, and container ship	Other types
Ship age	>12	⩽12
Ship flag performance	Black	Grey/white
Ship company performance	Low/very low	High/medium
Ship RO performance	Low/very low	High/medium
Number of deficiencies within previous 36 months	>5	⩽5
Number of deficiencies within previous 36 months	⩾3	<3

Table 5. Notation.

Notation	Meaning
$Q=\left\{q_1,\dots,q_j,\dots,q_J\right\}$	The set of J features.
$Q_j'=Q\setminus\left\{q_j\right\}$	The set of features in Q except for feature q_j.
$D=\left\{(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)\right\}$	The dataset, where $x_i=\left(x_i^{q_1},x_i^{q_2},\dots,x_i^{q_k}\right)$ is a vector with J feature values, and y_i is the number of deficiencies corresponding to x_i.
M	The number of subcategories for a basic categorical feature.
m	The index of a subcategory, where $m\in\{1,2,\dots,M\}$.
I_j^m	The set of ships with identical values of features in Q_j' and different values of the feature q_j, where $m\in\{1,2,\dots,M\}$ and $M=2^{\lvert Q_j'\rvert}$.
a_{q_j}	The classification value of a numerical (categorical) feature q_j in the NIR.
$Y_{m1}^{q_j},Y_{m2}^{q_j}$	The total number of deficiencies in two subcategories divided by feature q_j.
$n_{m1}^{q_j},n_{m2}^{q_j}$	The total number of ships in two subcategories divided by feature q_j.
$avg_{m1}^{q_j},avg_{m2}^{q_j}$	The average number of deficiencies in two subcategories divided by feature q_j.

Algorithm 1: Data analysis procedures

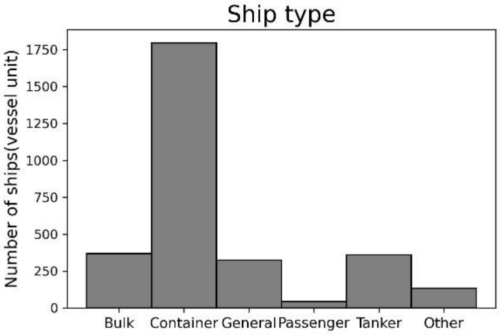
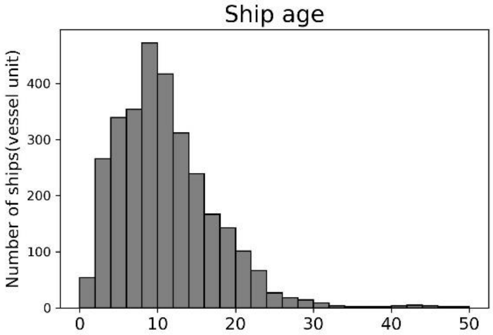
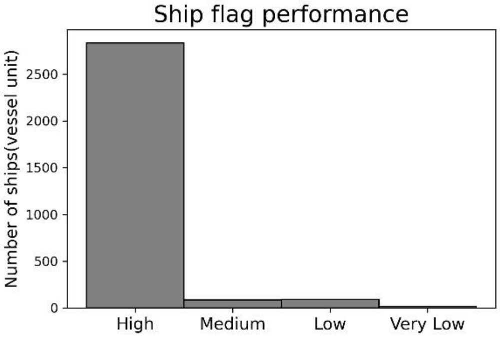
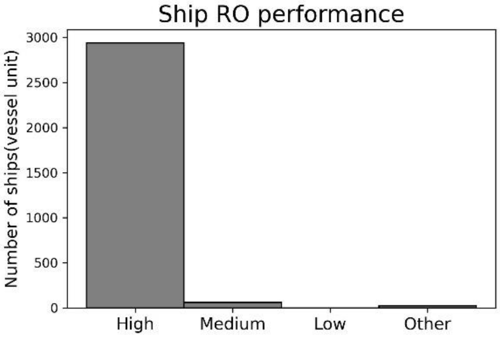
```
Input: $D=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$, $Q=\{q_1,...,q_j,...,q_J\}$, $Q_j'=Q\setminus\{q_j\}$
Output: Six decision trees
for $q_j \in Q$ do
    Denote ships with the same $x^{q_j'}$ by set $I_j^m$, where $q_j' \in Q_j'$
    for $m \in \{1,2,...,M\}$ do
        $Y_{m1}^{q_j}=0, Y_{m2}^{q_j}=0, n_{m1}^{q_j}=0, n_{m2}^{q_j}=0$    //initialize the total number of deficiencies and ships
        for $i \in I_j^m$ do
            if $q_j$ is a numerical feature then
                if $x_i^{q_j} > a_{q_j}$ then    //select ships in high-risk values of the numerical feature
                    $Y_{m1}^{q_j}=Y_{m1}^{q_j}+y_{m1}^{q_j}$, $n_{m1}^{q_j}=n_{m1}^{q_j}+1$
                else    //select ships in low-risk values of the numerical feature
                    $Y_{m2}^{q_j}=Y_{m2}^{q_j}+y_{m2}^{q_j}$, $n_{m2}^{q_j}=n_{m2}^{q_j}+1$
                end if
            end if
            if $q_j$ is a categorical feature then
                if $x_i^{q_j} = a_{q_j}$ then    //select ships in high-risk values of the categorical feature
                    $Y_{m1}^{q_j}=Y_{m1}^{q_j}+y_{m1}^{q_j}$, $n_{m1}^{q_j}=n_{m1}^{q_j}+1$
                else    //select ships in low-risk values of the categorical feature
                    $Y_{m2}^{q_j}=Y_{m2}^{q_j}+y_{m2}^{q_j}$, $n_{m2}^{q_j}=n_{m2}^{q_j}+1$
                end if
            end if
            return $avg_{m1}^{q_j}=\frac{Y_{m1}^{q_j}}{n_{m1}^{q_j}}, avg_{m2}^{q_j}=\frac{Y_{m2}^{q_j}}{n_{m2}^{q_j}}$    //one branch of the decision tree
        end for
    end for
    return the decision tree with the feature $q_j$ as the basic categorical feature
end for
Return six decision trees
```

3.2. Materials

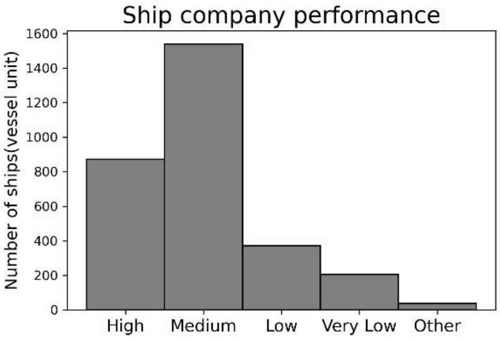
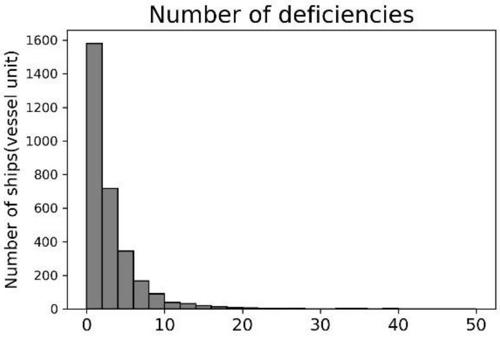
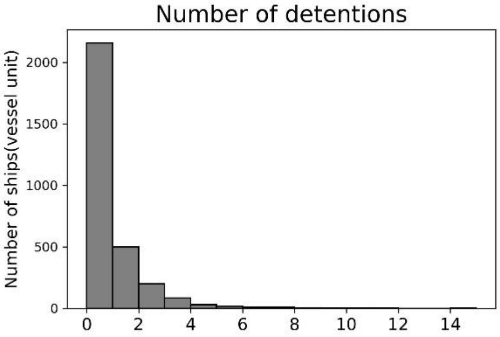
We collect PSC inspection records during January 2015 to December 2019 period at the Hong Kong port from the database of Tokyo MoU[1]. Data records with incomplete information are omitted, and we finally obtain 3026 PSC inspection records to be analyzed in this paper. The information we need comes from each PSC inspection record, including seven features that the NIR focuses on (i.e., ship type, ship age, ship flag performance, ship RO performance, ship company performance, the number of deficiencies within previous 36 months, and the number of detentions within previous 36 months) and the number of deficiencies identified. The distributions of the seven features over the 3026 cases are shown in Table 6. It is noticeable that the 3026 records do not have low and very low ship RO performance states. Because all ships in the dataset do not have to add points with respect to their ship RO performance, we ignore this feature in the following analysis.

[1] https://www.tokyo-mou.org/inspections_detentions/psc_database.php

Table 6. The distributions of the seven features.

Feature	Explanation	Distribution figure
Ship type	The main types of ships that have been inspected are chemical tanker, gas carrier, oil tanker, bulk carrier, passenger ship, and container ship.	
Ship age	The age of a ship is the time difference (in years) between the keel laid date and the PSC inspection date.	
Ship flag performance	The values of this variable are white, grey, black and not listed. Only flags that have been involved in more than 30 PSC inspections during the previous three years are listed in the black-grey-white lists; otherwise, the performance of the flag will not be listed.	
Ship RO performance	Ship RO is the classification society that carries out surveys and issues or endorses statutory certificates on behalf of a flag state. The states of performance of the ship RO are high, medium, low, and not listed.	

Continued on next page

Feature	Explanation	Distribution figure
Ship company performance	The ship company refers to the International Safety Management (ISM) company for the ship. The states of ship company performance are high, medium, low, and very low.	
Number of deficiencies within previous 36 months	The sum of deficiencies identified in the PSC inspections within previous 36 months.	
Number of detentions within previous 36 months	The sum of the detentions in the PSC inspections within previous 36 months.	

4. Results

Based on the method described in Section 3.1, we classify ships according to the values of a certain basic categorical feature by restricting that the values of remaining confounding features are identical. Because we select six features of the NIR's ship selection scheme, for each feature, the whole dataset could be divided into 32 (2^5) subcategories. In each subcategory, the values of all other features except for the chosen basic categorical feature are the same, so we can analyze whether there exists Simpson's paradox for this basic categorical feature in this subcategory. For each subcategory of one certain basic categorical feature, we can divide the subcategory into two groups according to the risk value of the basic categorical feature. We compute the average number of deficiencies in each group and display it in a histogram. The left red side of the histogram is the average number of deficiencies of ships with the high-risk value of the basic categorical value, while the right green side of the histogram is that of ships with the low-risk value. If the left value of the histogram is smaller than the right value, Simpson's paradox exists because this finding violates the assumption that ships with high-risk values should have more deficiencies than those with low-risk values. Then we present each subcategory as a branch. The 32 branches construct a decision tree, and the leaf node of each branch shows the histogram mentioned above. By doing this for each feature, we obtain six decision trees in total, and further select the branch with Simpson's paradox in each decision tree by blue boxes, as shown in Figures 1–6. In addition, we select the branches with Simpson's paradox and show them in the table, as shown in Tables 7–12. In each table, we analyze which categorical confounding features cause Simpson's paradox.

In each decision tree, we classify ships according to the values of one basic categorical feature, the values of the remaining confounding features of the ships are identical in the same branch. According to Figures 1–6, we find 23, 11, 16, 8, 12 and 10 cases of Simpson's paradox in six trees, respectively. Tables 7–12 show the average numbers of deficiencies of the branches showing Simpson's paradox.

Table 7. Branches showing Simpson's paradox of decision tree 1 based on the basic categorical feature ship type.

Branch no.	Categorical confounding features					Average deficiencies of ships in high-risk types	Average deficiencies of ships in other types
	Ship age	Ship flag performance	Ship company performance	Number of deficiencies within previous 36 months	Number of detentions within previous 36 months		
1	>12	Black	Low/very low	>5	⩾3	/*	23
3				⩽5	⩾3	/	21
4					<3	7.33	11.11
6			High/medium	>5	<3	1	7
7				⩽5	⩾3	5.33	23
8					<3	2.5	9
9		Grey/white	Low/very low	>5	⩾3	6.25	10
10					<3	7.41	8.06
11				⩽5	⩾3	6.92	8.61
12					<3	4.69	6.49
13			High/medium	>5	⩾3	7	9.75
15				⩽5	⩾3	5.72	6.65
16					<3	3.52	3.91
17	⩽12	Black	Low/very low	>5	⩾3	6	9
18					<3	/	12.83
19				⩽5	⩾3	/	11.5
24			High/medium	⩽5	<3	3.38	7.6
25		Grey/white	Low/very low	>5	⩾3	6	14.08
27				⩽5	⩾3	8.4	11.53
28					<3	4.23	5.24
29			High/medium	>5	⩾3	6.5	7.6
30				⩽5	<3	4.31	5.24
32					<3	2.46	2.53

Note: "/" means that there are no data samples in this subcategory.

Table 7 shows that, when we choose ship type as the basic categorical feature, there are 13 cases with other flag state performance and 12 cases with fewer than 5 deficiencies identified within previous 36 months where Simpson's paradox occurs. This result indicates that ship flag performance and the number of deficiencies within previous 36 months are two confounding features that are coupled with ship type. Table 8 shows that, when we choose ship age as the basic categorical feature, there are 6 cases under other flags and 6 cases with fewer than 3 detentions identified within previous 36 months where Simpson's paradox occurs. This result indicates that ship flag performance and the number of detentions within previous 36 months are two confounding features that are coupled with ship age. Table 9 shows that, when we choose ship flag performance as the basic categorical feature, there are 9 cases with ship age below 12 years old and 10 cases with fewer than 5 deficiencies identified within previous 36 months where Simpson's paradox occurs. This result indicates that ship age and ship company performance are two confounding features that are coupled with ship flag performance. Table 10 shows that, when we choose ship company performance as the basic categorical feature, there are 5 cases with ship age below 12 years old and 5 cases with high or medium ship company performance where Simpson's paradox occurs. This result indicates that ship age and the number of deficiencies within previous 36 months are two confounding features that are coupled with ship company performance. In Table 11, because all the numbers of cases with low-risk values of each confounding feature do not exceed half of the total 12 cases, it is hard to tell which features are coupled with the number of deficiencies within previous 36 months. Table 12 shows that, when we choose the number of detentions within previous 36 months as the basic categorical feature, there are 6 cases with ship age below 12 years old where Simpson's paradox occurs, which indicates that ship age is coupled with the number of detentions within previous 36 months. This result may indicate that because young ships are assumed to have less possibility of being detained, shipping companies may neglect these ships, resulting in more deficiencies detected in official PSC inspections. At last, Figure 7 displays the correlation of pairwise features among six features based on the above results.

Based on our results, we find that the selected features are nonlinear to the risk level of a ship, so the simple weighted-sum method could not identify ships' conditions accurately. The PSC authorities can consider improving the ship selection scheme of NIR by reformulating a scoring system that considers the correlations between pairwise features. Based on the decision rules revealed in this paper, port states should pay attention to certain ship features when scoring ships. For example, for features such as ship age, ship flag performance, and ship company performance, even if one of them of a ship is in high-risk values, it does not indicate that this ship is supposed to have a large number of deficiencies. The reason is that those features are coupled with three other features (see Figure 7), respectively. For example, although some ships are registered under black flags, their ages might be below 12 years old, or their ship company performance is high or medium, leading to a lower average number of deficiencies. However, for features such as the number of detentions within previous 36 months, they can reflect the actual condition of a ship because there is only one feature coupling to them. Therefore, these features should be given more attention during inspections. In addition, when developing advanced models for calculating ship risk level, e.g., statistical and ML models, the correlations between pairwise features can be further considered in the modeling procedures.

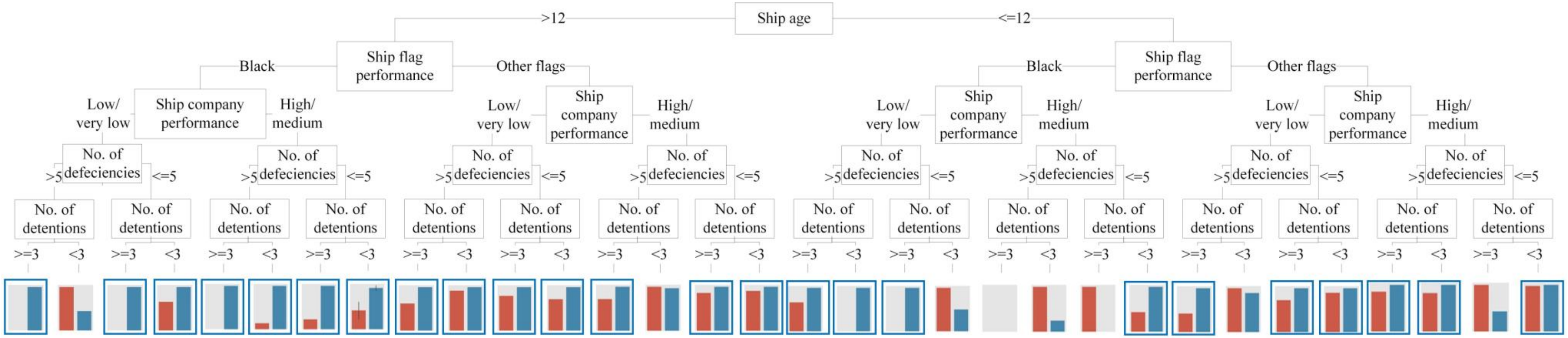


Figure 1. Decision tree 1 with ship type as the basic categorical feature.

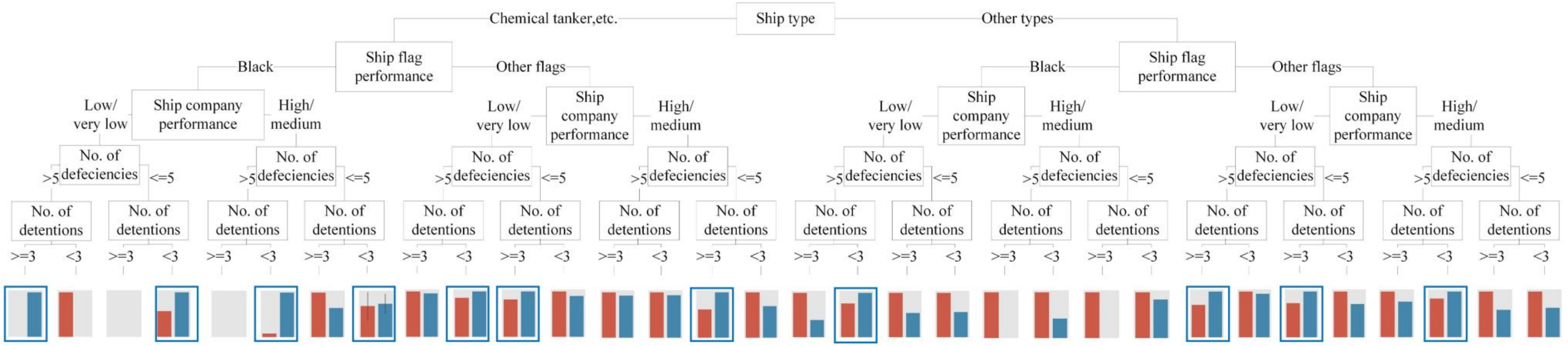


Figure 2. Decision tree 2 with ship age as the basic categorical feature.

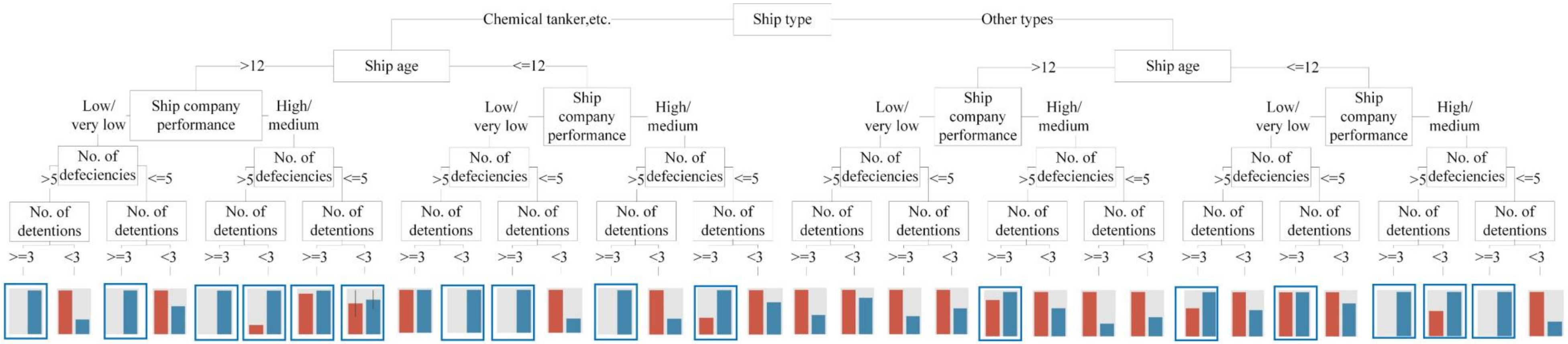


Figure 3. Decision tree 3 with ship flag performance as the basic categorical feature.

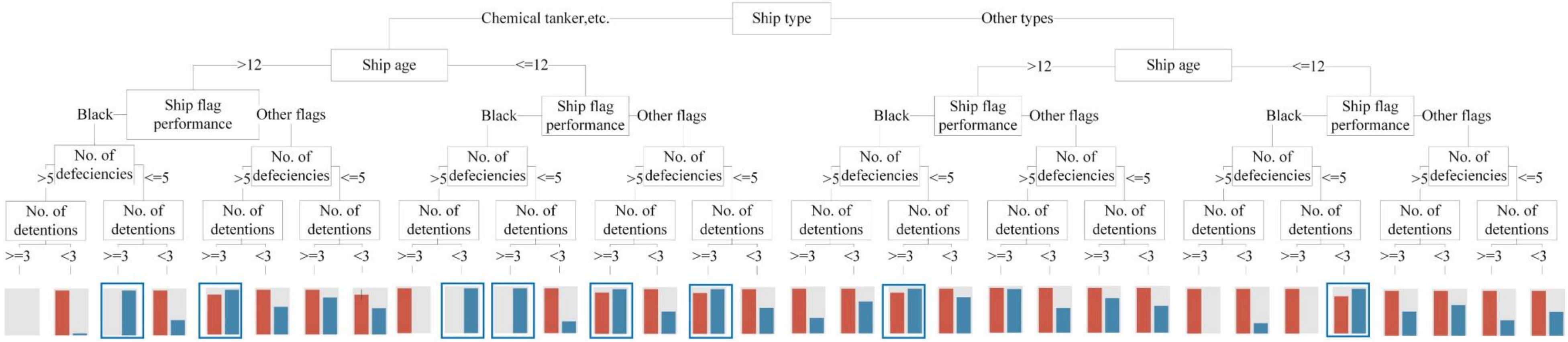


Figure 4. Decision tree 4 with ship company performance as the basic categorical feature.

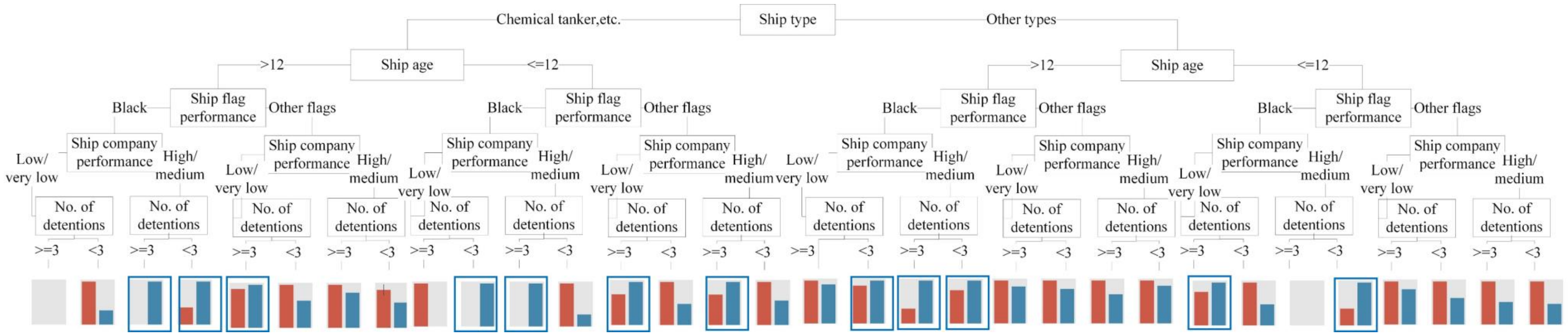


Figure 5. Decision tree 5 with the number of deficiencies within previous 36 months as the basic categorical feature.

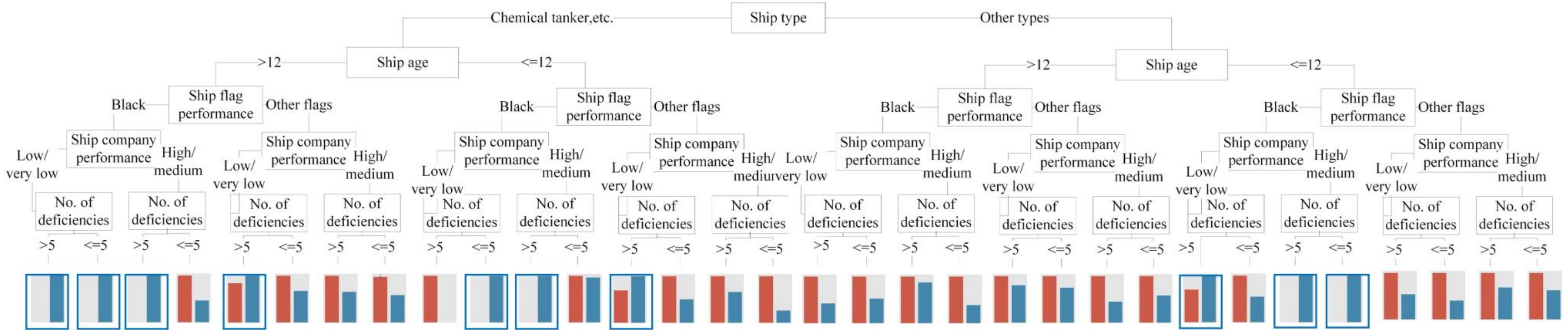


Figure 6. Decision tree 6 with the number of detentions within previous 36 months as the basic categorical feature.

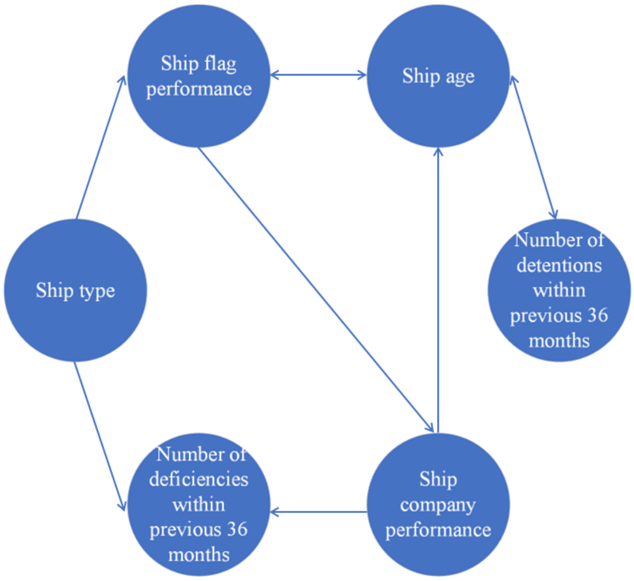

Ship flag performance
Ship age
Ship type
Number of detentions within previous 36 months
Number of deficiencies within previous 36 months
Ship company performance

Figure 7. Correlations of pairwise features among six features.

Table 8. Branches showing Simpson's paradox of decision tree 2 based on the basic categorical feature ship age.

Branch no.	Categorical confounding features					Average deficiencies of ships in age > 12	Average deficiencies of ships in age <= 12
	Ship type	Ship flag performance	Ship company performance	Number of deficiencies within previous 36 months	Number of detentions within previous 36 months		
1	Chemical tanker, Gas Carrier, Oil tanker, Bulk carrier, etc.	Black	Low/very low	>5	⩾3	/	6
4				⩽5	<3	7.33	12.67
6			High/medium	>5	<3	1	12
8				⩽5	<3	2.5	3.38
10		Grey/ white	Low/very low	>5	<3	7.41	8.62
11				⩽5	⩾3	6.92	8.4
15			High/medium	⩽5	⩾3	5.73	9.2
18	Other types	Black	Low/very low	>5	<3	9.8	12.83
25		Grey/ white	Low/very low	>5	⩾3	10	14.08
27				⩽5	⩾3	8.62	11.53
30			High/medium	>5	<3	4.44	5.24

Table 9. Branches showing Simpson's paradox of decision tree 3 based on the basic categorical feature ship flag performance.

Branch no.	Categorical confounding features					Average deficiencies of ships under black flags	Average deficiencies of ships under grey or white flags
	Ship type	Ship age	Ship company performance	Number of deficiencies within previous 36 months	Number of detentions within previous 36 months		
1	Chemical tanker, Gas Carrier, Oil tanker, Bulk carrier, etc.	>12	Low/very low	>5	≥3	/	6.25
3				≤5	≥3	/	6.92
5			High/medium	>5	≥3	/	7
6					<3	1	4.6
7				≤5	≥3	5.33	5.73
8					<3	2.5	3.53
10		≤12	Low/very low	>5	<3	/	8.62
11				≤5	≥3	/	8.4
13			High/medium	>5	≥3	/	6.5
15				≤5	≥3	3.5	9.2
21	Other types	>12	High/medium	>5	≥3	8	9.75
25		≤12	Low/very low	>5	≥3	9	14.08
27				≤5	≥3	11.5	11.53
29			High/medium	>5	≥3	/	7.6
30					<3	3	5.24
31				≤5	≥3	/	4

Table 10. Branches showing Simpson's paradox of decision tree 4 based on the basic categorical feature ship company performance.

Branch no.	Categorical confounding features					Average deficiencies of ships with low or very low company performance	Average deficiencies of ships with high or medium company performance
	Ship type	Ship age	Ship flag performance	Number of deficiencies within previous 36 months	Number of detentions within previous 36 months		
3	Chemical tanker, Gas Carrier, etc.	>12	Black	≤5	≥3	/	5.33
5			Grey/white	> 5	≥3	6.25	7
10		≤12	Black	>5	<3	/	12
11				≤5	≥3	/	3.5
13			Grey/white	>5	≥3	6	6.5
15				≤5	≥3	8.4	9.2
19	Other types	>12	Black	≤5	≥3	21	23
28		≤12	Black	≤5	<3	6.33	7.6

Table 11. Branches showing Simpson's paradox of decision tree 5 based on the basic categorical feature the number of deficiencies within previous 36 months.

Branch no.	Categorical confounding features					Average deficiencies of ships with more than 5 deficiencies	Average deficiencies of ships with at most 5 deficiencies
	Ship type	Ship age	Ship flag performance	Ship company performance	Number of detentions within previous 36 months		
3	Chemical tanker, Gas Carrier, Oil tanker, Bulk carrier, etc.	>12	Black	High/medium	⩾3	/	5.33
4					<3	1	2.5
5			Grey/ white	Low/very low	⩾3	6.25	6.92
10		⩽12	Black	Low/very low	<3	/	12.67
11				High/medium	⩾3	/	3.5
13			Grey/ white	Low/very low	⩾3	6	8.4
15				High/medium	⩾3	6.5	9.2
18	Other types	>12	Black	Low/very low	<3	9.8	11.11
19				High/medium	⩾3	8	23
20					<3	7	9
25		⩽12	Grey/ white	Low/very low	⩾3	9	11.5

Table 12. Branches showing Simpson's paradox of decision tree 6, based on the basic categorical feature the number of detentions within previous 36 months.

Branch no.	Categorical confounding features					Average deficiencies of ships with at least 3 detentions	Average deficiencies of ships with fewer than 3 detentions
	Ship type	Ship age	Ship flag performance	Ship company performance	Number of deficiencies within previous 36 months		
1	Chemical tanker, Gas Carrier, Oil tanker, Bulk carrier, etc.	>12	Black	Low/very low	>5	/	22
2					⩽5	/	7.33
3				High/medium	>5	/	1
5			Grey/ white	Low/very low	>5	6.25	7.41
10		⩽12	Black	Low/very low	>5	/	12.67
11				High/medium	>5	/	12
13			Grey/ white	Low/very low	>5	6	8.62
25	Other types	⩽12	Black	Low/very low	>5	9	12.83
27				Low/very low	>5	/	3
28					⩽5	/	7.6

5. Conclusions

Certain selection features, such as the ship's flag, age, and type, are believed to directly influence how well a ship is likely to be operated. Currently, the ship selection method widely adopted by PSC authorities regulated by Tokyo MoU is the NIR's simple weighted-sum scheme. This paper aims to investigate the plausibility of the NIR's weighted-sum method; that is, investigating whether there are paradoxes with respect to it. If Simpson's paradox exists, we could further explore which feature flips the effect. By observing the results, we find that many features selected by the NIR are coupled. Ship age, ship flag performance, and ship company performance are all coupled with three other features, respectively. Ship type and the number of deficiencies within previous 36 months are coupled with two other features, respectively. The number of detentions within previous 36 months is coupled with only one feature.

The results of this study indicate that selected features of the NIR are nonlinear to the risk level of ships, so the weighted-sum method should be improved according to their nonlinear relationships. The finding suggests that PSC authorities should pay attention to certain features like ship age, ship flag performance and ship company performance, since none of them could reflect the condition of the ship directly and there exist at least three pairwise correlations between each of them and other features. If we apply the nonlinear models which consider the correlations between the features (e.g., ML models) to evaluate a ship's risk level, the models can achieve better effectiveness in ship selection in PSC than the linear model (i.e., the weighted-sum method). Although this paper is the first study to examine the plausibility of the NIR, we do not quantitatively analyze the impact of compounding features on the final ship selection results. In the future, more advanced analytics techniques should be investigated for ship selection in PSC inspection based on the findings discovered in this paper, such as machine learning [39,40], online learning [41], prediction and optimization[42–44], evolutionary algorithms [45,46], multi-objective optimization [47], parameter control [48], and scheduling and routing [49], which have been applied as powerful solution approaches in many other domains.

Acknowledgments

The authors thank the four reviewers for their valuable comments.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. O. F. Abioye, M. A. Dulebenets, M. Kavoosi, J. Pasha, O. Theophilus, Vessel schedule recovery in liner shipping: Modeling alternative recovery options, *IEEE Trans. Intell. Transp. Syst.*, **22** (2021), 6420–6434. https://doi.org/10.1109/TITS.2020.2992120
2. S. Baştuğ, H. Haralambides, S. Esmer, E. Eminoğlu, Port competitiveness: Do container terminal operators and liner shipping companies see eye to eye?, *Mar. Policy.*, **135** (2022), 104866. https://doi.org/10.1016/j.marpol.2021.104866

3. M. A. Dulebenets, Multi-objective collaborative agreements amongst shipping lines and marine terminal operators for sustainable and environmental-friendly ship schedule design, *J. Clean. Prod.*, **342** (2022), 130897. https://doi.org/10.1016/j.jclepro.2022.130897
4. Z. Elmi, P. Singh, V. K. Meriga, K. Goniewicz, M. Borowska-Stefańska, S. Wiśniewski, M. A. Dulebenets, Uncertainties in liner shipping and ship schedule recovery: A state-of-the-art review, *J. Mar. Sci. Eng.*, **10** (2022), 563. https://doi.org/10.3390/jmse10050563
5. K. Wang, S. Wang, L. Zhen, X. Qu, Cruise service planning considering berth availability and decreasing marginal profit, *Transp. Res. Part B Methodol.*, **95** (2017), 1–18. https://doi.org/10.1016/j.trb.2016.10.020
6. L. Zhen, Y. Hu, S. Wang, G. Laporte, Y. Wu, Fleet deployment and demand fulfillment for container shipping liners, *Transp. Res. Part B Methodol.*, **120** (2019), 15–32. https://doi.org/10.1016/j.trb.2018.11.011
7. L. Zhen, Q. Sun, W. Zhang, K. Wang, W. Yi, Column generation for low carbon berth allocation under uncertainty, *J. Oper. Res. Soc.*, **72** (2021), 2225–2240. https://doi.org/10.1080/01605682.2020.1776168
8. L. Wu, Y. Adulyasak, J. F. Cordeau, S. Wang, Vessel service planning in seaports, *Oper. Res.*, **70** (2022), 2032–2053. https://doi.org/10.1287/opre.2021.2228
9. S. Wang, L. Zhen, D. Zhuge, Dynamic programming algorithms for selection of waste disposal ports in cruise shipping, *Transp. Res. Part B Methodol.*, **108** (2018), 235–248. https://doi.org/10.1016/j.trb.2017.12.016
10. L. Zhen, Y. Wu, S. Wang, G. Laporte, Green technology adoption for fleet deployment in a shipping network, *Transp. Res. Part B Methodol.*, **139** (2020), 388–410. https://doi.org/10.1016/j.trb.2020.06.004
11. W. Yi, L. Zhen, Y. Jin, Stackelberg game analysis of government subsidy on sustainable off-site construction and low-carbon logistics, *Clean. Logist. Supply Chain.*, **2** (2021), 100013. https://doi.org/10.1016/j.clscn.2021.100013
12. W. Yi, S. Wu, L. Zhen, G. Chawynski, Bi-level programming subsidy design for promoting sustainable prefabricated product logistics, *Clean. Logist. Supply Chain.*, **1** (2021), 100005. https://doi.org/10.1016/j.clscn.2021.100005
13. S. Wang, D. Zhuge, L. Zhen, C. Y. Lee, Liner shipping service planning under sulfur emission Regulations, *Transp. Sci.*, **55** (2021), 491–509. https://doi.org/10.1287/trsc.2020.1010
14. Paris MoU, *Organization of Paris MoU*, 2019. Available form: https://www.parismou.org/about-us/organisation
15. Tokyo MoU, *Information Sheet of the New Inspection Regime (NIR)*, 2014. Available from: http://www.tokyo-mou.org/doc/NIR-information%20sheet-r.pdf
16. European Commission, *Ex-post evaluation of Directive 2009/16/EC on Port State Control: Final Report*, 2018. Available from: https://data.europa.eu/doi/10.2832/154686
17. R. Yan, S. Wang, Ship inspection by port state control—review of current research, *Smart Transp. Syst.*, (2019), 233–241. https://doi.org/10.1007/978-981-13-8683-1_24
18. P. Cariou, M. Q. Mejia, F. C. Wolff, An econometric analysis of deficiencies noted in port state control inspections, *Marit. Policy Manag.*, **34** (2007), 243–258. https://doi.org/10.1080/03088830701343047
19. P. Cariou, M. Q. Mejia, F. C. Wolff, Evidence on target factors used for port state control inspections, *Mar. Policy.*, **33** (2009), 847–859. https://doi.org/10.1016/j.marpol.2009.03.004

20. M. C. Tsou, Big data analysis of port state control ship detention database, *J. Mar. Eng. Technol.*, **18** (2019), 113–121. https://doi.org/10.1080/20464177.2018.1505029
21. S. Knapp, P. H. Franses, A global view on port state control: Econometric analysis of the differences across port state control regimes, *Marit. Policy Manag.*, **34** (2007), 453–482. https://doi.org/10.1080/03088830701585217
22. F. J. Ravira, F. Piniella, Evaluating the impact of PSC inspectors' professional profile: A case study of the Spanish Maritime Administration, *WMU J. Marit. Aff.*, **15** (2016), 221–236. https://doi.org/10.1007/s13437-015-0096-y
23. A. Graziano, P. Cariou, F. C. Wolff, M. Q. Mejia, J. U. Schröder-Hinrichs, Port state control inspections in the European Union: Do inspector's number and background matter?, *Mar. Policy.*, **88** (2018), 230–241. https://doi.org/10.1016/j.marpol.2017.11.031
24. R. F. Xu, Q. Lu, W. J. Li, K. X. Li, H. S. Zheng, A risk assessment system for improving port state control inspection, in: *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics*, (2007), 818–823. https://doi.org/10.1109/ICMLC.2007.4370255
25. Z. Yang, Z. Yang, J. Yin, Z. Qu, A risk-based game model for rational inspections in port state control, *Transp. Res. Part E Logist. Transp. Rev.*, **118** (2018), 477–495. https://doi.org/10.1016/j.tre.2018.08.001
26. S. Wang, R. Yan, X. Qu, Development of a non-parametric classifier: Effective identification, algorithm, and applications in port state control for maritime transportation, *Transp. Res. Part B Methodol.*, **128** (2019), 129–157. https://doi.org/10.1016/j.trb.2019.07.017
27. D. Dinis, A. P. Teixeira, C. Guedes Soares, Probabilistic approach for characterising the static risk of ships using Bayesian networks, *Reliab. Eng. Syst. Saf.*, **203** (2020), 107073. https://doi.org/10.1016/j.ress.2020.107073
28. R. Yan, S. Wang, C. Peng, An artificial intelligence model considering data imbalance for ship selection in port state control based on detention probabilities, *J. Comput. Sci.*, **48** (2021), 101257. https://doi.org/10.1016/j.jocs.2020.101257
29. R. Yan, S. Wang, Ship detention prediction using anomaly detection in port state control: model and explanation, *Electron. Res. Arch.*, **30** (2022), 3679–3691. https://doi.org/10.3934/era.2022188
30. E. H. Simpson, The interpretation of interaction in contingency tables, *J. R. Stat. Soc. Ser. B Methodol.*, **13** (1951), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x
31. C. R. Blyth, On Simpson's paradox and the sure-thing principle, *J. Am. Stat. Assoc.*, **67** (1972), 364–366. https://doi.org/10.1080/01621459.1972.10482387
32. J. Zidek, Maximal Simpson-disaggregations of 2 × 2 tables, *Biometrika.*, **71** (1984), 187–190. https://doi.org/10.2307/2336411
33. Y. Bishop, S. Fienberg, P. Holland, R. Light, F. Mosteller, Discrete multivariate analysis: Theory and practice, *Appl. Psychol. Meas.*, **1** (1977). https://doi.org/10.1177/014662167700100218
34. M. G. Pavlides, M. D. Perlman, How likely is Simpson's paradox?, *Am. Stat.*, **63** (2009), 226–233. https://www.jstor.org/stable/25652271
35. S. Sunder, Simpson's reversal paradox and cost allocation, *J. Account. Res.*, **21** (1983), 222–233. https://doi.org/10.2307/2490944
36. A. Mehrez, J. R. Brown, M. Khouja, Aggregate efficiency measures and Simpson's Paradox, *Contemp. Account. Res.*, **9** (1992), 329–342. https://doi.org/10.1111/j.1911-3846.1992.tb00884.x

37. S. P. Curley, G. J. Browne, Normative and descriptive analyses of Simpson's paradox in decision making, *Organ. Behav. Hum. Decis. Process.*, **84** (2001), 308–333. https://doi.org/10.1006/obhd.2000.2928
38. N. D. Melumad, A. Ziv, Reduced quality and an unlevel playing field could make consumers happier, *Manag. Sci.*, **50** (2004), 1646–1659. https://doi.org/10.1287/mnsc.1040.0277
39. W. Zhu, J. Wu, T. Fu, J. Wang, J. Zhang, Q. Shangguan, Dynamic prediction of traffic incident duration on urban expressways: a deep learning approach based on LSTM and MLP, *J. Intell. Connect. Veh.*, **4** (2021), 80–91. https://doi.org/10.1108/JICV-03-2021-0004
40. N. Lyu, Y. Wang, C. Wu, L. Peng, A. F. Thomas, Using naturalistic driving data to identify driving style based on longitudinal driving operation conditions, *J. Intell. Connect. Veh.*, **5** (2022), 17–35. https://doi.org/10.1108/JICV-07-2021-0008
41. H. Zhao, C. Zhang, An online-learning-based evolutionary many-objective algorithm, *Inf. Sci.*, **509** (2020), 1–21. https://doi.org/10.1016/j.ins.2019.08.069
42. S. Wang, R. Yan, A global method from predictive to prescriptive analytics considering prediction error for "Predict, then optimize" with an example of low-carbon logistics, *Clean. Logist. Supply Chain.*, **4** (2022), 100062. https://doi.org/10.1016/j.clscn.2022.100062
43. R. Yan, S. Wang, Integrating prediction with optimization: Models and applications in transportation management, *Multimodal Transp.*, **1** (2022), 100018. https://doi.org/10.1016/j.multra.2022.100018
44. S. Wang, X. Tian, R. Yan, Y. Liu, A deficiency of prescriptive analytics—No perfect predicted value or predicted distribution exists, *Electron. Res. Arch.*, **30** (2022), 3586–3594. https://doi.org/10.3934/era.2022183
45. M. A. Dulebenets, R. Moses, E. E. Ozguven, A. Vanli, Minimizing carbon dioxide emissions due to container handling at marine container terminals via hybrid evolutionary algorithms, *IEEE Access.*, **5** (2017), 8131–8147. https://doi.org/10.1109/ACCESS.2017.2693030
46. M. Dulebenets, A diploid evolutionary algorithm for sustainable truck scheduling at a cross-docking facility, *Sustainability.*, **10** (2018), 1333. https://doi.org/10.3390/su10051333
47. J. Pasha, A. L. Nwodu, A. M. Fathollahi-Fard, G. Tian, Z. Li, H. Wang, et al., Exact and metaheuristic algorithms for the vehicle routing problem with a factory-in-a-box in multi-objective settings, *Adv. Eng. Inform.*, **52** (2022), 101623. https://doi.org/10.1016/j.aei.2022.101623
48. M. Kavoosi, M. A. Dulebenets, O. F. Abioye, J. Pasha, H. Wang, H. Chi, An augmented self-adaptive parameter control in evolutionary computation: A case study for the berth scheduling problem, *Adv. Eng. Inf.*, **42** (2019), 100972. https://doi.org/10.1016/j.aei.2019.100972
49. M. Rabbani, N. Oladzad-Abbasabady, N. Akbarian-Saravi, Ambulance routing in disaster response considering variable patient condition: NSGA-II and MOPSO algorithms, *J. Ind. Manag. Optim.*, **18** (2022), 1035. https://doi.org/10.3934/jimo.2021007