DATA ANALYTICS FOR IMPROVING SHIPPING EFFICIENCY: MODELS, METHODS, AND APPLICATIONS

YAN RAN

PhD

The Hong Kong Polytechnic University

2022

The Hong Kong Polytechnic University

Department of Logistics and Maritime Studies

Data Analytics for Improving Shipping Efficiency: Models, Methods, and Applications

YAN Ran

A thesis submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

Jan 2022

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

_____(Signed)
YAN Ran (Name of student)

Abstract

This thesis aims to improve the effectiveness and efficiency of port state control (PSC) inspection, which is one of the most important international shipping policies, from the aspects of ship risk prediction and PSC inspector assignment and scheduling using data analytics and operations research models. In addition to a comprehensive summary and review of ship selection methods applied at ports over the world and proposed in existing literature, this thesis comprises three studies. In the first study, ship deficiency number, which is a ship risk indicator in the PSC inspection, is predicted using a state-of-the-art XGBoost model. The XGBoost model takes shipping domain knowledge regarding ship flag, recognized organization, and company performances into account to improve model accuracy and fairness. Based on the predictions, a PSC inspector scheduling model is proposed to help the ports optimally allocate inspection resources. According to the model structure, the concepts of *inspection template* and *un-dominated inspection template* are further proposed and incorporated in the optimization model to improve computation efficiency and model flexibility.

In the second study, three two-step approaches that match the inspection resources with the ships' deficiency conditions are proposed, aimed at identifying the most deficiencies of them. The three approaches combine prediction models with optimization models, and the optimization models are equivalent in all the approaches while the prediction models differ from each other regarding their prediction targets or structure. Specifically, the first approach predicts the number of deficiencies in each deficiency category for each ship and then develops an integer optimization model that assigns the inspectors to the ships to be inspected. The second approach predicts the number of deficiencies each inspector can identify for each ship and then applies an integer optimization model to assign the inspectors to the ships to be inspected. The third approach is a semi-"smart predict then optimize" (semi-SPO) method. It also predicts the number of deficiencies each inspector can identify for each ship and uses the same integer optimization model as the second approach. However, instead of minimizing the mean squared error as in the second approach, it adopts a loss function motivated by the structure of the optimization problem in the second approach. The prediction results are then input to PSC officer (PSCO) assignment models such that

the PSCOs' expertise and the ships' deficiency conditions can be matched, and the inspection efficiency can be improved.

In the third study, a data-driven ship risk prediction framework using features the same as the current ship selection scheme is developed for high-risk ship identification and selection based on gradient boosting regression tree (GBRT). Like existing ship risk prediction models, the proposed framework is of black-box nature whose decision process and working mechanism are opaque. To improve model explainability, the explanation of the prediction of individual ships by the Shapley additive explanations (SHAP) method with the properties of local accuracy and consistency is provided. Furthermore, the local SHAP method is innovatively extended to a fully explainable near linear-form global surrogate model of the original black-box data-driven model by deriving feature coefficients and fitting curves of feature values and SHAP values. This demonstrates that the behaviour of black-box data-driven models can be as interpretable as white-box models while retaining their prediction accuracy.

Key words: Maritime transportation; Marine policy; Port state control (PSC); Ship inspection; Data analytics; Ship risk prediction; Resource assignment and scheduling; Explainable artificial intelligence

Publications During PhD Study

Arising from the Thesis

1. Yan, R., Wang, S., Peng, C., 2021. Ship selection in port state control: Status and perspectives. Maritime Policy & Management, 1–16.
2. Yan, R., Wang, S., Cao, J., Sun, D., 2021. Shipping domain knowledge informed prediction and optimization in port state control. Transportation Research Part B: Methodological 149, 52–78.
3. Yan, R., Wang, S., Fagerholt, K., 2020. A semi-"smart predict then optimize" (semi-SPO) method for efficient ship inspection. Transportation Research Part B: Methodological 142, 100–125.
4. Yan, R., Wu, S., Jin, Y., Cao, J., Wang, S., 2022. Efficient and explainable ship selection planning in port state control. Transportation Research Part C: Emerging Technologies, under review.

Others

1. Yan, R., Wang, S., Psaraftis, H., 2021. Data analytics for fuel consumption management in maritime transportation: Status and perspectives. Transportation Research Part E: Logistics and Transportation Review 155, 102489.
2. Yan, R., Wang, S., Fagerholt, K., 2021. Coordinated approaches for ship inspection planning. Maritime Policy & Management, DOI: 10.1080/03088839.2021.1903599.
3. Yan, R., Mo, H., Wang, S., Yang, D., 2021. Analysis and prediction of ship energy efficiency based on the MRV system. Maritime Policy & Management, DOI: 10.1080/03088839.2021.1968059.
4. Chen, X., Yan, R.*, Wu, S., Liu, Z., Mo, H., Wang, S., 2021. A fleet deployment model to minimize the covering time of maritime rescue missions. Maritime Policy & Management, DOI: 10.1080/03088839.2021.2017042.
5. Yan, R., Wang, S., Peng, C., 2021. An artificial intelligence model considering data imbalance for ship selection in port state control based on detention probabilities. Journal of Computational Science 48, 101257.

6. Yan, R., Wang, S., Zhen, L., Laporte, G., 2021. Emerging approaches applied to maritime transport research: Past and future. Communications in Transportation Research, 100011.

7. Wang, S., Zhen, L., Psaraftis, H., Yan, R., 2021. Implications of the EU's inclusion of maritime transport in Emissions Trading System for shipping companies. Engineering 7(5), 554–557.

Acknowledgements

First and foremost, I am extremely grateful to my supervisor Professor Shuaian (Hans) Wang for his invaluable supervision, insightful comments, and continuous support from both an academic perspective and many other aspects during my MPhil and PhD studies. Doing a PhD under his supervision is a pleasant journey, and every time I think about my decision on pursuing my postgraduate degrees in LMS, PolyU with Hans as my supervisor, I feel extremely lucky. During this whole period, I am influenced by and learned a lot from him for his broad knowledge, dedication to high-quality and impactful research, innovative thinking, and sincere and pure personality. He is always there and willing to offer his help when I get in trouble in doing research and in daily life. This thesis would not have been possible without him. Hans is very concerned about the future development of his students. He offered me valuable chances to draft funding application proposals and gave me detailed comments on improving them, assist in supervising research assistants and MPhil students, gain internship experience in a shipping company, attend international conferences, and have meetings and discussions with maritime practitioners from governmental institutions and shipping companies. The limited space in this Acknowledge chapter is not enough to express my sincere gratitude for him, but the experience of my (MPhil and) PhD study with Hans will always motivate me to become a good supervisor, a good teacher, and a good person.

Second, I would like to express my gratitude to the Board of Examiners: Prof. LUO Meifeng, Prof. CHEN Nan, and Prof. Daniel LONG Zhuoyu, and Prof. Shuaian (Hans) Wang for their valuable comments to improve the quality of this paper. I would also like to thank the Department of Logistics and Maritime Studies and the HK PolyU for their financial support to my PhD study. Heartfelt thanks go to my collaborators of research papers: Prof. Cao Jiannong, Dr. Chen Xinyuan, Dr. Du Yuquan, Prof. Kjetil Fagerholt, Mr. Mo Haoyu, Prof. Harilaos N. Psaraftis, Prof. Gilbert Laporte, Dr. Liu Zhiyuan, Prof. Sun Defeng, Prof. Peng Chuansheng, Prof. Qu Xiaobo, Dr. Wu Lingxiao, Dr. Wu Shining, Dr. Yang Dong, Dr. Zhuge Dan, and Prof. Zhen Lu, etc. for their inspiring discussions and insightful comments. I am also very much appreciated the supports and comments given by our industrial collaborators: the

Marine Department of HKSAR, Wah Kwong, Pacific Basin, and China Merchants Energy Shipping.

Also, I would like to express my special thanks to my friends here in Hong Kong, who make my journey there happy and wonderful, especially the ones in LMS and EE departments. Particularly, I would like to thank the members of Prof. Wang's research team, who are excellent researchers and are constantly motivating me to keep improving and do better. I can always turn to them for help when I am confused. I would also sincerely thank my friends in the 'Menghunguoguan' group, the two badminton groups, and the lunch&dinner ('LMS Ganfan') group, for the joy and fun we had, the interesting things we shared, the delicious food we tasted, and the support and care you gave to me. My thanks also go to my good friends ever since my primary school, junior and senior high school, and college. Although we don't contact frequently, I know you are there whenever I need your help, care, and comfort. Especially, I would give my special thanks to my cat-brother, Mr. (Dr.) CHEN Xuliang, for his countless support, patience, and accompanying with love given to me through my MPhil and PhD studies. I appreciate the happy time we had together, which is the most important regulator in my busy and sometimes heavy stress postgraduate study. I also appreciate our deep-minded conversations and debate, which are sometimes quite interesting. I appreciate the most your patience when I told you bits and pieces, your insightful analysis and suggestions when I sought advice from you, and your time, efforts, and patience to accompany me for eating, playing, and going shopping.

Besides, I would like to thank Pinard Liu for his useful blogs introducing algorithms and codes for machine learning models and this patient and professional replies to my comments to his blogs. I am also grateful for the courses and the lecturers from the department of LMS, Coursera, and YouTube, which lay the foundation for my research. I would also like to thank the administrative staff in the General Office of the Department of LMS and the Graduate School for their assistant in my study and life at PolyU.

Finally, I would like to dedicate this thesis with love and gratitude to my family: my beloved parents and maternal grandparents, for their endless and unconditional love to me no matter where and how old I am. Thank you for supporting the decisions I made for my future ever since I chose to study at the Experiment School and Shishi High School, and then pursue my postgraduate degrees in Hong Kong. Your kindness

and diligence have always motivated me to pursue what I believe to realize my self-worth and contribute my efforts to make the world a better place. Especially, I spent a long time at home due to the COVID-19, and I am grateful for your patience, understanding, and care throughout my daily life during this period. Particularly, I really enjoyed the delicious meals you made every day (which is no doubt the biggest support for my research), hanging out and chatting, and playing mahjong with you every weekend :). This thesis would not have been possible without your support and encouragement. Meanwhile, this thesis is dedicated with deep love and gratitude to the memory of my beloved grandmother, who took the most care of me and had the greatest impact on me ever since I was born, and taught me literacy, singing, dancing, playing the piano, playing badminton, painting, reciting ancient poetry, and the truth of life. Unfortunately, she passed away just after I finished high school and had a very bad and unexpected performance in the college entrance examination. My regret has been with me for these years, for the reason that I didn't see her for the last time, and that I thought I didn't meet her expectations. Although she was gone during my PhD study, I believe she has always been with me and will be proud of me when I finish this thesis and my PhD study.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 BACKGROUND

Maritime transport is responsible for over 80% of global merchandise trade by volume and more than 70% by value (UNCTAD, 2021). Maritime safety is the backbone of running a smooth business, as the consequence of an accident can be very serious to the vessel and its crew, to the marine environment, and even to the whole society. More recently, emissions generated by vessels are receiving increasing attention as they may pollute the environment and exacerbate the greenhouse effect. To enhance maritime safety, protect the marine environment, and graduate decent living and working conditions of the crew, various international regulations and conventions are proposed and implemented which the vessels must comply with.

Generally, a ship is regarded as substandard if its condition is substantially below the standards or if the crew does not comply with the safe manning document (IMO, 2017). Distinguishing substandard ships from all ships in operation is essential. Ship flag state bears the main responsibility to inspect the ships under its registration or license, and it is regarded as the first line of defence against substandard shipping. Unfortunately, flag states cannot perform their duties efficiently due to internal and external reasons (Li and Zheng, 2008). As the second line of defence, port state control (PSC) inspection, which is the inspection conducted by port authorities targeted at foreign visiting ships, is proposed and implemented to ensure that these ships comply with various regulations and conventions (Cariou et al., 2007; Heij et al., 2011).

A typical PSC inspection starts from selecting high-risk foreign visiting ships to a port state, which is carried out by each port authority on the morning of a working day following the ship selection scheme adopted. Then, ship inspectors, i.e., PSC officers (PSCOs), are assigned for ship inspection by the decision makers at the port. During PSC inspections, a condition found not to comply with the relevant convention is denoted by a ship deficiency. Fatal deficiencies that may put too much danger to the sea can lead to detention, which is an intervention action carried out by the port state (IMO, 2017). Ship deficiency and detention are seen as the inspection target of the

PSC inspection, and they will be recorded in the corresponding database together with the ships' specifications.

To allow information and experience exchange, avoid multiple inspections of one ship within a short period, and apply standard inspection criteria and procedure, the regional Memorandum of Understandings on port state control (i.e., MoUs on PSC) are signed and established. The Hong Kong port belongs to the Tokyo MoU, which is in charge of the Asia-Pacific Region, and there are another eight regional MoUs on PSC around the world. Uniform inspection procedures and standards are required to be implemented within one MoU, including identifying and selecting high-risk ships, assigning inspection resource (mainly refers to PSCOs), deciding onboard inspection items and sequence, and recording ship deficiency and detention. Accurate identification of substandard visiting ships is the key to improve the effectiveness of PSC, as only a small proportion can be inspected among a large number of foreign visiting ships due to the limited inspection resources at a port. In addition, effective assignment and scheduling of the inspection resources, i.e., the available PSCOs, considering their working time and expertise is a foundation for effective PSC inspections, as such resources are scarce at a port while the background and experience of the PSCOs at the same port can be varied. To achieve both goals, this thesis proposes several prediction and optimization models to predict ship risk considering various factors and optimize the assignment and scheduling of inspection resources at a port. Specifically, predictions of ship overall condition regarding the total number of deficiencies and the number of deficiencies under each deficiency category are achieved by state-of-the-art machine learning (ML) models considering shipping domain knowledge and the downstream optimization model structure, which are followed by PSCO scheduling or assignment models. Furthermore, the black-box ML based ship risk prediction models are opened by using post-hoc explanation methods to achieve model explanation.

1.2 THESIS OUTLINE

The remainder of the thesis is organized as follows. Chapter 2 summaries and reviews ship selection methods applied at ports and the existing literature on improving the efficiency of PSC inspection. Chapter 3 develops a prediction model for the total number of deficiencies of a ship where the shipping domain knowledge regarding ship operation information is incorporated by modifying the structure and property of the

prediction model. The predictions are then input to a PSCO scheduling model to realize optimal inspection resource allocation. Chapter 4 proposes several multi-target regression models to predict the number of deficiencies under each deficiency category for each ship. The regression models are based on the classic random forest model, while they differ from each other regarding their structures and prediction targets. The predictions then serve as the input to the following PSCO assignment models. Chapter 5 aims to open up the black-box models for ship risk prediction by first developing an accurate prediction model for ship deficiency number, and then using a local post-hoc explanation method to explain the prediction results. The local explanation method is extended to a global one by developing a near linear-form surrogate model which is also fully explainable. Chapter 6 concludes the thesis.

Chapter 2: Current Approach and Literature Review[1]

This chapter first summaries the current ship selection methods used in different ports around the world. It then reviews the existing studies on improving PSC inspection efficiency, including models for high-risk ship selection and onboard inspection efficiency improvement.

2.1 SHIP SELECTION METHOD AT PORT

A uniform ship selection procedure is required to be adopted by all the ports in one MoU on PSC, and the ship selection models currently used in ports are easy to understand and implement. Take the example of the Tokyo MoU, the New Inspection Regime (NIR) is applied to determine the inspection priority and time interval between inspections of ships by calculating their ship risk profile (SRP) (Tokyo MoU, 2014). Ships are divided into three categories based on the SRP: high risk ships (HRS), standard risk ships (SRS), and low risk ships (LRS) according to the information sheet given in Table 2-1. Particularly, the flag Black-Grey-White list and the RO performance list are published by the Tokyo MoU in the annual report considering the inspection and detention history of the vessels under the corresponding flag and RO over the preceding three calendar years. Ship company performance is the performance of a ship's international safety management (ISM) company which is calculated daily on the basis of a running 36-month period considering the detention and deficiency history of the company's fleet. The time windows attached to HRS, SRS, and LRS are 2–4, 5–8, and 9–18 months, respectively, in the Tokyo MoU.

[1] Yan, R., Wang, S., Peng, C., 2021. Ship selection in port state control: Status and perspectives. Maritime Policy & Management, 1–16.

Table 2-1. Information sheet of SRP adopted by the Tokyo MoU

Parameters	Values	Weighting points	Criteria for LRS
Ship type	Chemical tanker, gas carrier, oil tanker, bulk carrier, passenger ship, container ship	2	\
Ship age (calculated based on the keel laid date)	All types with age > 12y	1	\
Flag performance in Black-Grey-White list of Tokyo MoU	Black	1	White, and should be IMO Audit
RO performance evaluated by Tokyo MoU	Low/very low	1	High, and should be an RO recognized by the Tokyo MoU
Company performance evaluated by Tokyo MoU	Low/very low/no inspection within previous 36 months [unknown]	2	High
Deficiencies within previous 36 months	Inspections which recorded over 5 deficiencies	The number of inspections which recorded over 5 deficiencies	All inspections have 5 or less deficiencies and has at least one inspection within previous 36 months
Detentions within previous 36 months	3 or more detentions	1	No detention

Ship risk profile	Criteria	Inspection time window
HRS	When the sum of weighting points $\geqslant 4$	2 to 4 months
SRS	Neither HRS nor LRS	5 to 8 months
LRS	All the criteria for LRS are met	9 to 18 months

For a foreign visiting ship attached with a specific risk profile, its inspection priority is determined by the relationship between its last inspection time and the inspection time window attached to its SRP. Especially, there are two levels of inspection priority: ships with the last inspection time beyond the upper bound of the inspection time window are of Priority I and must be inspected; ships with the last inspection time within the inspection time window are of Priority II and may be inspected. Meanwhile, ships with the last inspection time less than the lower bound of the inspection time window have no priority (Tokyo MoU, 2013).

Other MoUs on PSC have their own methods of ship selection. For example, the Paris MoU, the Abuja MoU, and the Black Sea MoU also adopt the NIR for SRP calculation and ship selection. However, the NIR used in these MoUs is slightly different from that used in the Tokyo MoU. For instance, their information sheet for SRP calculation is different from the sheet used in the Tokyo MoU (Abuja MoU 2012; Black Sea MoU 2016; Paris MoU 2014). Moreover, the time windows attached to HRS, SRS, and LRS are 5–6, 10–12, and 24–36 months, respectively, in the Abuja MoU and

the Paris MoU. Some ports adopt simpler ship selection methods. For example, ships can be exempted from further inspection if they have been inspected within the last six months and found to comply with the regulations of the Mediterranean MoU (Mediterranean MoU 2020).

The ship selection models currently used in ports are easy to understand and implement. In particular, risk factors related to ship characteristics and historical inspection records are considered in the NIR. However, several drawbacks of the NIR would adversely reduce its efficiency in identifying substandard ships. First, the parameters considered to calculate ship risk are limited. Only basic ship characteristics and rough historical inspection results are considered; other parameters, such as ships involved in accidents and incidents, are neglected. Second, the weights attached to the parameters are highly dependent on expert judgement, which may lead to inaccuracies and inconsistencies. Third, the total ship risk score is calculated by a simple weighted sum method and the correlations between the parameters are not taken into account, further compromising its effectiveness. Fourth, although the SRP divides ships into three risk categories, no specific risk score is attached to an individual ship. This further weakens its effectiveness as an indicator of ship risk level. Consequently, the European Commission (2017) pointed out that "there is room for improvement in the design of the ship selection method" for PSC inspections.

2.2 STUDIES ON IMPROVING PSC INSPECTION EFFICIENCY

A recent literature review classified the large body of literature on PSC into four main categories: factors influencing PSC inspection results, ship selection schemes in PSC, PSC inspection effects, and suggestions for MoU management (Yan and Wang 2019). In this chapter, we focus on the studies on improving PSC efficiency, which develop models for ship selection and onboard inspection efficiency improvement.

Li (1999) is the pioneer who proposed an innovative risk score system to evaluate ship quality in PSC inspection. The author considered several ship generic factors: ship age, flag, insurers, classification, and operators. Degré (2007) also adopted the risk score concept to select high-risk ships for PSC inspection. The developed model considered ship physical factors, i.e., type, size, and age, and the selection criteria used in Paris MoU, namely ship flag, recognized organization, and company. Xu et al. (2007a) developed a vessel risk assessment system for PSC based

on support vector machine (SVM). They further improved the system performance by combining web mining technique for extracting new features (Xu et al, 2007b). Gao et al. (2008) proposed another ship risk assessment system for PSC. The system combined k-nearest neighbor with support vector machine (KNN-SVM) to remove noisy training samples and adopted bag of words (BW) to extract new features. All the above three papers used ship detention as the prediction target: if a ship was predicted to be high-risk and was detained, the prediction was considered accurate. The highest accuracy of the three models is about 22% due to the highly imbalanced distribution of ship detention in the dataset: the number of detained records is much smaller than records without detention. The imbalanced data makes the prediction a complex task. Zhou and Sun (2010) implemented a self-evolutional ship targeting system for ship detention prediction using generalized additive modeling (GAM). The system was designed to relieve the negative Matthew Effect, which was caused by the ship target system at Ningbo port as it would unavoidably set ships with bad history into a vicious circle by increasing their inspection frequency.

In 2011, the NIR (and the SRP) first entered into force in the Paris MoU and replaced the existing ship target factor system at the time. In 2014, the Tokyo MoU also implemented the NIR (and the SRP). The SRP is easy to understand and implement. Moreover, it enhances PSC efficiency to improve maritime safety, security, pollution prevention, and working conditions to some extent (European Commission, 2017). It is also recognized that the implementation of the NIR has modernized the PSC inspection system. For example, Yang et al. (2020) analyzed the influence of the implementation of the NIR on the PSC inspection system and ship quality from macroscopic and microscopic perspectives. The authors concluded that the influence of the NIR was generally positive, as it prompted ship owners to maintain their vessels at a high-quality level.

In recent years, more advanced and accurate ship selection models have been proposed to improve SRP efficiency. Based on the inspection data of bulk carriers in the Paris MoU from 2005 to 2008, Yang et al. (2018a) implemented a Bayesian network (BN) approach to predict ship detention. The main factors influencing ship detention, namely the number of deficiencies, type of inspection, RO, and ship age, were also analyzed. The authors proposed a strategic game model incorporating the outcomes of the BN model to determine the optimal inspection rate for port states.

Recommendations for port authorities were generated based on the results: when port authorities have sufficient resources, they should choose the optimal inspection rate; otherwise, they should increase the severity of punishment to tackle the poor efforts and illegal actions of ship owners (Yang et al. 2018b). Wang et al. (2019) developed a BN model to predict the number of ship deficiencies identified during a PSC inspection. In addition, they compared the proposed BN model and the current SRP ship selection scheme in the Tokyo MoU, demonstrating the superiority of the BN model. Based on the static risk factors adopted by the NIR and the SRP, Dinis et al. (2020) developed a BN-based ship risk assessment and maritime traffic monitoring model. They conducted a quantitative assessment of the predictive validity of the model using historical PSC inspection records. The results were consistent with the SRP criteria and models developed in other studies.

In addition to the popular BNs, researchers have proposed various other types of models for ship selection for PSC inspection. For instance, Yan et al. (2021b) developed a balanced random forest (BRF) model for ship detention prediction which can address the problems brought by the highly imbalanced dataset due to low detention rate (about 3.55%). An SVM model was proposed by Wu et al. (2021) for ship detention prediction. Particularly, input features were selected by analytic hierarchy process and grey relational analysis to improve prediction accuracy. Apart from generic ship factors and historical inspection factors, some ship selection studies also consider ships involved in casualties and incidents, as they could indicate high ship risk and possible future accidents (Heij and Knapp, 2019; Knapp and Heij, 2020).

To improve the efficiency of onboard inspection, association rule mining technologies are widely proposed to figure out the relationship between various factors in existing literature. The generated rules can offer meaningful insights to onboard deficiency and detention identification. Tsou (2019) explored the detention database of Tokyo MoU using association rule mining techniques. The author identified the correlations between detention deficiencies and the correlations between deficiencies and ship-/inspection-related factors. Chung et al. (2020) analyzed the historical PSC inspection records in Taiwan Provence of China using Apriori algorithm. The correlations between ship characteristics and PSC deficiencies were identified. Yan et al. (2021c) also adopted Apriori algorithm to identify the relationship between ship deficiencies based on the inspection records at the Hong Kong port. Onboard

inspection schemes were then proposed according to the rules identified. Fu et al. (2020) analyzed the correlations between ship generic properties and ship deficiency and detention conditions using Apriori algorithm based on the inspection records in Tokyo MoU.

Chapter 3: Shipping Domain Knowledge Informed Prediction and Optimization in Port State Control[2]

This chapter addresses one critical issue faced by the port states about how to optimally allocate the limited inspection resources for inspecting the visiting ships. It first develops a state-of-the-art XGBoost model to accurately predict ship deficiency number. Particularly, the XGBoost model takes shipping domain knowledge regarding ship flag, recognized organization, and company performance into account to improve model performance and prediction fairness. Based on the predictions, a PSCO scheduling model is proposed to help the maritime authorities optimally allocate inspection resources. Considering that a PSCO can inspect at most four ships in a day, we further propose and incorporate the concepts of *inspection template* and *un-dominated inspection template* in the optimization models to reduce problem size as well as improve computation efficiency and model flexibility. Numerical experiments and sensitivity analysis using practical data and settings at the Hong Kong port are conducted to validate model performance and robustness.

3.1 INTRODUCTION

In PSC inspection, it is generally believed that accurate identification of high-risk visiting ships is a pre-requirement while effective assignment and scheduling of available PSCOs is a foundation for effective PSC inspections. The reasons are as follows. First, it is impossible to inspect all visiting ships as port inspection resources, especially the number of available PSCOs, are quite limited. Second, among all visiting ships, only a small portion of ships need to be inspected. The annual report of Tokyo MoU in Asia-Pacific region shows that only 60% of the inspections conducted between 2009 and 2019 identified deficiencies, and no more than 6% inspections were with detention (Tokyo MoU, 2020). Third, the proportion of ships inspected is crucial in port management. If too few substandard ships are inspected at a port, ship owners

2 Yan, R., Wang, S., Cao, J., Sun, D., 2021. Shipping domain knowledge informed prediction and optimization in port state control. Transportation Research Part B: Methodological 149, 52–78.

may lack the motivation to intensively maintain ship conditions, which in return attracts more substandard ships to the port. On the contrary, if too many qualified ships are inspected, the competitiveness of the port may be reduced and consequently leads ship owners to turn to other destinations with relaxed inspection policy (Yang et al. 2018b). Therefore, accurate identification of high-risk ships and rational allocation of inspection resources guarantee effective PSC inspections by picking out which ships are most worthy of inspection and finishing the inspection tasks efficiently without putting too much delay in shipment. They also help the port states to find a balance between stringent inspections of substandard ships and reducing un-necessary inspections of qualified ships and thus to better fulfill their responsibilities and enhance their competitiveness.

One of the widely adopted and the most advanced ship selection method applied by port states is the NIR. It calculates SRP based on ship generic parameters including ship type, ship age, flag performance, recognized organization (RO) performance, and company performance, and inspection historical factors including deficiency and detention conditions (Paris MoU, 2010; Tokyo MoU, 2014). It is noted that all the parameters are objective except for flag, RO, and company performance, which is calculated by the MoUs. More specifically, ship flag performance is established annually by taking its ships' inspection and detention conditions over the preceding three calendar years into account. Black-grey-white ship flag lists are published in an MoU's annual report, where flag performance gets worse from white to grey and to black. RO is a qualified organization which has been assessed and authorized by the flag state to provide necessary statutory services and certification of ships entitled to fly its flag (IMO 2017). The performance of all ROs is established annually considering their ships' inspection and detention history over the preceding three calendar years. The RO performance list is published in an MoU's annual report, where the performance of ROs gets worse from high, to medium, to low, and to very low. Ship company is the International Safety Management (ISM) company of a ship, and its performance is determined by their ships' detention and deficiency history calculated daily on the basis of a running 36-month period. Similar to ship RO performance, company performance gets worse from high, to medium, to low, and to very low.

As ship flag, RO, and company play an important role in ship management, operation, and maintenance, they are taken into account in the popular SRP ship selection scheme applied at ports. In return, a ship's performance in PSC inspection can influence the reputation of its flag, RO, and company and their performance evaluated by PSC MoUs. Under this condition, it is justifiable to conclude that given all other conditions being equal, a ship should be estimated to have worse performance in PSC inspection (e.g., more deficiencies and higher probability of detention) if the performance of its flag/RO/company gets worse. However, such domain knowledge is seldom considered in current literature of high-risk ship selection mainly because combining domain knowledge with ML models is not a trivial task as it requires modifications of the prediction models or finding good properties of them. Besides, PSCO assignment and scheduling models, which require allocating the available and scarce inspection resources as well as arranging the starting and ending time of the required activities, are also rarely proposed in current research. This chapter aims to bridge this gap with the contributions summarized as follows.

First, from a theoretical point of view, we first develop an ML prediction model considering proper and adequate domain knowledge to solve problems in maritime transportation. Specifically, a state-of-the-art tree-based model called XGBoost is developed to predict ship deficiency number in PSC inspection. In the XGBoost model, we combine the shipping domain knowledge regarding ship flag, RO, and company performance in a natural way. Based on the predictions, a PSCO scheduling model for ship inspection is then proposed considering a PSCO's work and rest time to guarantee inspection effectiveness. By taking the properties of the optimization model for PSCO scheduling into account, we propose the concepts of *inspection template*, *un-dominated inspection template*, and strengthened constraints to reduce problem size as well as improve model flexibility and solving efficiency.

Second, from a practical point of view, a practical problem in PSC inspection, which is one of the most important shipping policies, is addressed in this study. Numerical experiments show that the proposed combined model for ship deficiency number prediction and PSCO scheduling is more than 20% better than the current PSCO scheduling strategy at ports regarding the number of deficiencies identified. Meanwhile, the gap between the proposed model and the perfect-forecast policy is only about 8% regarding the number of deficiencies identified. The proposed model

can help port state authorities to identify higher risk ships and schedule inspection resources more efficiently. Especially, it contributes to assisting the port states to achieve a balance between effectively identifying and inspecting substandard ships and reducing un-necessary inspections of qualified ships and consequently frightening them from choosing this port in future shipment. Therefore, the main objectives of PSC to eliminate substandard shipping, to promote maritime safety and security, to protect the marine environment, and to safeguard seafarers' working and living conditions on board ships can be enhanced.

3.2 RESEARCH GAP

Based on the literature reviewed in Chapter 2, several limitations regarding high-risk ship selection and PSCO scheduling in PSC are summarized as follows. First, current studies of high-risk ship selection have failed to consider shipping domain knowledge in ship risk prediction, including the monotonicity regarding ship flag/RO/company performance in ship risk prediction. It is likely that the prediction models ignoring such domain knowledge give opposite prediction results due to model inaccuracy and noises in training data (Sill, 1997; Duivesteijn and Feelders, 2008; Daniels and Velikova, 2010; Pei et al., 2016). This indicates that only by taking such shipping domain knowledge into account in ship risk prediction models can fair and reasonable prediction results be generated. Here regarding such prediction results to be "fair" is because for ship flag/RO/company which adopt more effective management measures on their ships, it can be expected that their ships' performance in PSC inspection should be better than other flags/ROs/companies adopting worse management strategies. In return, reducing the inspection frequency of their ships can promote them to better fulfill their maintenance and operational duties and attract more shippers to choose their services. The reason to regard such prediction results to be "reasonable" is that considering monotonicity into an ML model "can be an important model requirement with a view toward explaining and justifying decisions" (Duivesteijn and Feelders, 2008) to the decision makers. It is also reported by Pazzani et al. (2001) that the learned rules with monotonicity constraints were significantly more acceptable to experts than rules learned without the monotonicity restrictions when experts expect certain monotonicity based on their experience.

Second, there is little literature aiming to design tailored PSCO assignment or scheduling schemes for ship inspection, and thus to validate the superiority of the

proposed ship risk prediction models over the current schemes at ports. Indeed, prediction model accuracy is figured out in many current studies, and their superiority over current ship selection scheme is also presented. However, port inspection resources (e.g., the number of available PSCOs) are scarce and the arrival and departure time of ships are not fixed. This indicates that not all ships can be inspected in practice, and the gap between the proposed and current schemes in practice remains to be validated. Formulation and solution techniques for assignment and scheduling models are proposed in current literature, and typical modeling approaches include column generation (Van Den Akker et al., 2005; Huisman, 2007; Janacek et al., 2017; Kulkarni et al., 2018) and Dantzig-Wolfe decomposition (Janacek et al., 2017; Kulkarni et al., 2018; Muñoz et al., 2018). Nevertheless, there is no tailored modelling approach considering the problem structure and the corresponding properties of PSC inspection as well as proposing intuitive solving strategies that are comprehensible to the decision makers at port authorities. Therefore, it is of vital importance to develop tailored and easy-to-understand PSCO assignment and scheduling modeling approach based on ship risk prediction models to figure out their superiority in practice and improve the efficiency of PSC inspection.

To address these issues, this chapter develops a highly accurate XGBoost model for ship deficiency number prediction considering shipping domain knowledge. It then proposes PSCO scheduling models based on the predictions which are consistent with the actual situation at port. Extensive computational experiments and sensitivity analysis are conducted to validate the model performance.

3.3 DATA AND MODEL VALIDATION METRICS

3.3.1 Data description

The case dataset of this study contains 1,974 PSC initial inspection records and the corresponding ship related factors at the Hong Kong port from January 2016 to December 2018. Especially, PSC inspection records are downloaded from the public database provided by Tokyo MoU[3], and ship related factors are searched from World Shipping Register database. The prediction target is the number of deficiencies detected in the current PSC inspection. We consider 14 features that are regarded to be highly related to ship deficiency number in the current literature and by domain

[3] http://www.tokyo-mou.org/inspections_detentions/psc_database.php

knowledge, namely ship age, gross tonnage (GT), length, depth, beam, type, flag performance, RO performance, and company performance in Tokyo MoU, last PSC inspection date in Tokyo MoU, the number of deficiencies in last inspection in Tokyo MoU, the number of total detentions in all historical PSC inspections, the number of flag changes, and whether the ship has a casualty in last 5 years. Moreover, as required by the Tokyo MoU, from the best to the worst, the states of ship flag performance are white, grey, and black, the states of ship RO and company performance are high, medium, low, and very low, respectively. After data preprocessing, the whole dataset contains 1,926 samples. The explanation, method of feature encoding, and the descriptive statistics of the prediction target and the 14 features in the whole dataset are shown in Table 3-1.

Table 3-1. Variable explanation, encoding method, and descriptive statistics

Variable name	Explanation	Encoding	Mean value	Min value	Max value
deficiency number	The number of deficiencies identified in the current PSC initial inspection.	No encoding	4.31	0	51
age	The time interval (in years) between the keel laid date and the current PSC inspection date.	No encoding	10.8	0	47
GT	A nonlinear measure of a ship's internal volume, with 100 cubic feet as the unit.	No encoding	44,908	497	266,681
length	The overall maximum length of a ship (in meters).	No encoding	214.88	32.29	400
depth	The vertical distance (in meters) measured from the top of the keel to the upper deck at side measured inside the plating.	No encoding	17.79	4.28	36.02
beam	The width of ship hull (in meters).	No encoding	31.93	7.38	60.05
type	Ships in the dataset are classified into the following types: bulk carrier, container ship, general cargo/multipurpose, passenger ship, tanker, and other.	One-hot encoding: is_bulk_carrier: 1 for bulk carrier and 0, otherwise; is_container_ship: 1 for container ship and 0, otherwise; is_ general cargo/multipurpose: 1 for general cargo/multipurpose and 0, otherwise; is_ passenger_ship: 1 for passenger ship and 0, otherwise; is_tanker: 1 for tanker and 0, otherwise; is_other: 1 for other ship types and 0, otherwise.	\	\	\
flag performance	Ship flag performance is calculated based on the flag Black-Grey-White list provided by Tokyo MoU (Tokyo MoU, 2018a).	Label encoding: white->1*; grey->2; black->3.	\	\	\
RO performance	Ship RO performance is calculated based on RO performance list provided by Tokyo MoU (Tokyo MoU, 2018a).	Label encoding: high->1; medium->2; low->3.	\	\	\

company performance	Ship company performance is calculated based on company performance matrix provided by Tokyo MoU (Tokyo MoU, 2018a)	Label encoding: high->1; medium->2; low->3; very low->4.	\	\	\
last inspection date	The time interval between the last and current PSC initial inspections within Tokyo MoU (in months). For ships that are inspected for the first time (i.e., with no previous inspection records), the state of this variable is set to be "−1".	No encoding.	10.2	0	180.7
last deficiency number	The number of deficiencies identified in last PSC initial inspection within Tokyo MoU. For ships that are inspected for the first time, the state of this variable is set to be "−1".	No encoding.	2.46	0	38
total detentions	The total number of detentions of a ship in all previous PSC inspections since the keel laid date.	No encoding.	0.59	0	18
the number of flag changes	The total number of times of ship flag change from keel laid date to the current PSC inspection date.	No encoding.	0.66	0	8
casualty in last 5 years	A binary variable indicating whether a ship was involved in casualties in the last five years.	One-hot encoding: casualty-in-5-years: 1 for any casualty occurs in the last 5 years and 0, otherwise.	\	\	\

Note *: this indicates that the state of "white" is encoded to be "1".

3.3.2 Model validation metrics

The deficiency number prediction models are validated using two common metrics for regression problems in ML: mean squared error (MSE) and mean absolute error (MAE). Given a total of n samples in the dataset, the real output y_i and the predicted output $\hat{y}_i$ for sample i, $i = 1,...,n$, the definitions of MSE and MAE are as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 , \tag{3.1}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| . \tag{3.2}$$

3.4 INTRODUCTION AND CONSTRUCTION OF XGBOOST MODEL

3.4.1 The structure of XGBoost model

Ensemble models in ML combine the predictions of multiple simpler base models to improving the overall model prediction performance (Friedman et al., 2001). Two main ensemble models are bagging (bootstrap aggregating) and boosting. Bagging builds several base models independently and then average their predictions. Boosting builds sequential and dependent base models in the way that one base model is built considering the errors of the base models built so far and then produces a powerful ensemble (Friedman et al., 2001). In boosting models, a base model is also called a weak learner which may be only slightly better than random guessing. Meanwhile, the main idea of boosting is to add new weak learners to the ensemble sequentially, and in each iteration, the weak learner is trained with respect to the error of the whole ensemble learned so far (Natekin and Knoll, 2013). As boosting is purely algorithm-driven, a gradient-descent based formulation of boosting methods is derived which is called gradient boosting machine (GBM) (Freund and Schapire, 1997; Friedman et al., 2000). The principal idea of GBM is to construct new weak learners to be maximally correlated with the negative gradient of the loss function associated with the whole ensemble.

XGBoost (short for eXtreme Gradient Boosting) is an implementation of GBM that uses tree-structured weak learners (Chen and Guestrin, 2016). It is highly effective (which allows parallel and distributed computing) and scalable (which is able to handle datasets containing billions of examples in distributed or memory-limited settings). The detailed procedure of constructing a XGBoost model is as follows (Chen, 2014;

Chen and Guestrin, 2016). Given a dataset with n samples and m features, denoted by $D=\{(\mathbf{x}_i, y_i), i=1,...,n\}$, $\mathbf{x}_i \in R^m$, $y_i \in R$, a tree ensemble model uses K additive functions to predict the target y_i (the predicted value is denoted by $\hat{y}_i$) is

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i),\ f_k \in F, \tag{3.3}$$

where F is a space of functions that contains all regression and classification tree (CART) based regression trees. In XGBoost, the learning objective function to be minimized, which aims to draw a balance between model accuracy and complexity, is as follows:

$$\overline{obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k). \tag{3.4}$$

The first term in Eq. (3.4) is the training loss regarding all training samples, and the second term is the tree complexity. In regression problems, a common choice for the training loss function is half of the MSE, which is given by

$$\sum_{i=1}^{n} l(y_i, \hat{y}_i) = \frac{1}{2}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{3.5}$$

where the multiplication of $1/2$ is for the ease of calculation. Eq. (3.5) is also the loss function used in this study. As XGBoost is developed based on additive training, the prediction value after finishing 0 to $t=1,...,K$ iterations can be written as

$$\begin{aligned}
&\hat{y}_i^{(0)} = 0,\\
&\hat{y}_i^{(1)} = f_1(\mathbf{x}_i) = \hat{y}_i^{(0)} + f_1(\mathbf{x}_i),\\
&\hat{y}_i^{(2)} = f_1(\mathbf{x}_i) + f_2(\mathbf{x}_i) = \hat{y}_i^{(1)} + f_2(\mathbf{x}_i),\\
&\quad ...\\
&\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(\mathbf{x}_i) = \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i).
\end{aligned} \tag{3.6}$$

By combining Eqs. (3.4) to (3.6), the objective function in the tth iteration can be given by

$$\begin{aligned}
\overline{obj}^{(t)} &= \frac{1}{2}\sum_{i=1}^{n} (y_i - \hat{y}_i^{(t)})^2 + \sum_{k=1}^{t} \Omega(f_k)\\
&= \frac{1}{2}\sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)))^2 + \Omega(f_t) + \sum_{k=1}^{t-1} \Omega(f_k).
\end{aligned} \tag{3.7}$$

Eq. (3.7) contains three terms. The first term is the loss function of the tth iteration. The second term is the penalty for tree complexity in the tth iteration. The last term is the sum of penalties for tree complexity of all the first $t-1$ iterations. Define $g_i^t = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ and $h_i^t = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ as the first and second order gradients of the

loss function in Eq. (3.7). The concrete expressions for g_i^t and h_i^t can be given if the loss function is explicitly defined. As we choose Eq. (3.5) as the loss function in this study, we can have $g_i^t = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) = \hat{y}_i^{(t-1)} - y_i$ and $h_i^t = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) = 1$. It should be mentioned that the values for g_i^t and h_i^t for sample i of the tth iteration are fixed as they are only related to the output generated in the $(t-1)$th iteration. The second order Taylor expansion at $\hat{y}_i^{(t-1)}$ of Eq. (3.7) should be

$$\overline{obj}^{(t)} \simeq \sum_{i=1}^{n} [\frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i^{(t-1)})^2 + g_i^t f_t(\mathbf{x}_i) + \frac{1}{2} h_i^t f_t(\mathbf{x}_i)^2] + \Omega(f_t) + \sum_{k=1}^{t-1} \Omega(f_k) . \tag{3.8}$$

In the first term of Eq. (3.8), $\frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i^{(t-1)})^2$ is the loss of the $(t-1)$th iteration and thus it is a constant. The last term of Eq. (3.8), i.e. $\sum_{k=1}^{t-1} \Omega(f_k)$, is the penalty of tree complexity of all the first $t-1$ iterations and thus is also a constant. All the constants can be removed. Therefore, we represent the objective function in the tth iteration as

$$obj^{(t)} = \sum_{i=1}^{n} [g_i^t f_t(\mathbf{x}_i) + \frac{1}{2} h_i^t f_t(\mathbf{x}_i)^2] + \Omega(f_t) . \tag{3.9}$$

The goal of the tth iteration is to construct a tree to minimize Eq. (3.9), which requires to decide the outputs of the leaf nodes and the structure of the tree. We first assume that the tree structure is fixed and discuss the way to determine the outputs of the leaf nodes. Define a tree by a vector of outputs (which are also called weights) in leaves, and a leaf index mapping function that maps a sample to a leaf as

$$f_t(\mathbf{x}) = w^t_{q^t(\mathbf{x})}, \mathbf{w}^t \in R^{T_t}, \ q^t : R^m \to \{1, 2, ..., T_t\} , \tag{3.10}$$

where T_t is the number of leaves in the tree, $\mathbf{w}^t$ is the vector of outputs in all the leaves, and q^t is the function assigning each sample to the corresponding leaf in the tth iteration. We use the following toy example to exemplify the notations used in Eq. (3.10).

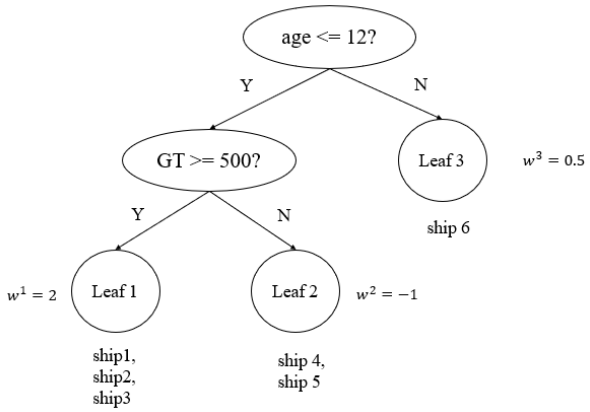


Figure 3-1. A toy regression tree in the t th iteration of a XGBoost model

Suppose that we have a total of six samples in a toy training set, and the developed regression tree in the t th iteration is shown in Figure 3-1. The notations in Eq. (3.10) can be exemplified as follows: $T_t = 3$, $\mathbf{w}^t = \{2, -1, 0.5\}$, $q^t(\text{ship}1) = 1$, $q^t(\text{ship}2) = 1$, $q^t(\text{ship}3) = 1$, $q^t(\text{ship}4) = 2$, $q^t(\text{ship}5) = 2$, and $q^t(\text{ship}6) = 3$. It should be noted that the leaf output in XGBoost is different from the leaf output in traditional CART regression tree: the leaf output in XGBoost is calculated by optimization models whereas the leaf output in CART regression tree is simply the mean of the output of the samples in that leaf node in regression problems. The tree complexity in the objective function of XGBoost is defined as

$$\Omega(f_t) = \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} w_j^{t\,2}, \tag{3.11}$$

where the first term is the penalty on the total number of leaves and the second term is the penalty on the sum of squares of the weights in the leaves in the tth iteration. γ and λ are two hyperparameters that need to be tuned and are used to balance model accuracy and complexity. Define the sample set in leaf j on the tree of the tth iteration as $I_j^t = \{i \mid q^t(\mathbf{x}_i) = j\}$, $i = 1, ..., n$, we can regroup the objective function in Eq. (3.9) by leaf and combine with Eq. (3.11) to be

$$\begin{aligned} obj^{(t)} &= \sum_{i=1}^{n} [g_i^t f_t(\mathbf{x}_i) + \frac{1}{2} h_i^t f_t(\mathbf{x}_i)^2] + \Omega(f_t) \\ &= \sum_{i=1}^{n} [g_i^t w_{q^t(\mathbf{x}_i)}^t + \frac{1}{2} h_i^t w_{q^t(\mathbf{x}_i)}^{t\,2}] + \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} w_j^{t\,2} \\ &= \sum_{j=1}^{T_t} [(\sum_{i \in I_j^t} g_i^t) w_j^t + \frac{1}{2} (\sum_{i \in I_j^t} h_i^t + \lambda) w_j^{t\,2}] + \gamma T_t. \end{aligned} \tag{3.12}$$

For simplicity, we define $G_j^t = \sum_{i \in I_j^t} g_i^t$ and $H_j^t = \sum_{i \in I_j^t} h_i^t$, Eq. (3.12) can be written as

$$obj^{(t)} = \sum_{j=1}^{T_t} [G_j^t w_j^t + \frac{1}{2}(H_j^t + \lambda) w_j^{t\,2}] + \gamma T_t \,. \tag{3.13}$$

As we have assumed that the tree structure (i.e. q^t) is fixed, and thus G_j^t, H_j^t, and T_t are all fixed. The optimal output w_j^t (denoted by w_j^{t*}) can be found by letting the first derivative of $obj^{(t)}$ with respect to w_j^t be 0, which is

$$w_j^{t*} = -\frac{G_j^t}{H_j^t + \lambda} \,. \tag{3.14}$$

The optimal value of the objective function is

$$obj^{(t)*} = -\frac{1}{2} \sum_{j=1}^{T_t} \frac{G_j^{t\,2}}{H_j^t + \lambda} + \gamma T_t \,. \tag{3.15}$$

After the outputs in the tree leaves are determined by assuming the tree structure is given, the last question is how to decide the tree structure (i.e., split a node into two child nodes) in an XGBoost tree. In practice, we grow the tree in a greedy manner by splitting nodes from the tree root by enumerating all values (or quantiles of values) of all features (or a subset of features) and calculating the reduction in objective function after adding a candidate split by

$$\begin{aligned} gain &= obj_{L+R}^{(t)} - (obj_L^{(t)} + obj_R^{(t)}) \\ &= \frac{1}{2}\left[\frac{G_L^{t\,2}}{H_L^t + \lambda} + \frac{G_R^{t\,2}}{H_R^t + \lambda} - \frac{(G_L^t + G_R^t)^2}{H_L^t + H_R^t + \lambda} \right] - \gamma, \end{aligned} \tag{3.16}$$

where $obj_{L+R}^{(t)}$, $obj_L^{(t)}$, and $obj_R^{(t)}$ are the objective functions of the node for splitting, the objective function of the left child node if adding this candidate split, and the objective function of the right child node if adding this candidate split, respectively. $gain$ is calculated for each candidate split of the current node. As G_L^t and G_R^t (H_L^t and H_R^t) are the sum of first (second) derivative of the samples contained in left and right child leaf respectively, different splits would lead to different values for G_L^t and G_R^t (H_L^t and H_R^t). If $gain < 0$, the candidate split is not considered. For all positive values for $gain$, we choose the feature and value corresponding to the maximum value of $gain$ to split the node as it could reach the maximum reduction of the objective function after the splitting.

3.4.2 Feature monotonic constraints in XGBoost

Apart from the state-of-the-art prediction performance, XGBoost also has the nice property to enforce monotonic constraint on the feature(s) regarding the prediction target (Chen, 2016). Suppose we have a total of m features and the feature vector is denoted by $\mathbf{x} = (x^1, ..., x^{m'}, ..., x^m)$. We put a monotonically increasing constraint on feature $\bar{m}$, which means that for two samples i_1 and i_2 that have the same feature

values except for $x^{\bar{m}}$, i.e. $x_{i_1}^{m'} = x_{i_2}^{m'}$, $m' = 1,...,\bar{m}-1$, $\bar{m}+1,...,m$ and $x_{i_1}^{\bar{m}} < x_{i_2}^{\bar{m}}$, the predicted target for i_1 should be no more than that for i_2, i.e. $\hat{y}_{i_1} \le \hat{y}_{i_2}$. As the monotonic constraint works in the context that all the features are equal in the samples except for the feature which is enforced to be monotonic (denote the set of samples by I'), the prediction process of the samples in I' in a tree can be simplified to only contain the splits on the monotonic feature (as using all other features and values will always lead the samples to the same tree nodes and thus have the same output). In this context, the work process to impose monotonic constraint on a feature can be illustrated as follows.

We still use feature $\bar{m}$ which we put a monotonically increasing constraint on as an example. An illustration of the tree structure is shown in Figure 3-2. The output of all samples in the root node is W_0. From splitting the root node, we would expect the weight assigned to the right child not to be lower than the weight assigned to the left child while using the monotonic feature for splitting. When feature $\bar{m}$ is picked to split the root node, if a candidate value of $\bar{m}$ leads to a higher weight in the left child than that in the right child, this candidate value will be abandoned for the current node splitting. That is, when enumerating all possible values of feature $\bar{m}$ to split the root node, only the values leading to no lower weights in right child than in left child will be retained for further comparison. If all possible splits lead to higher output in the left child than in the right child, the node would not be split any more. If feasible splits exist and the optimal splitting point is found, we could have $W_L \le W_R$, where W_L is the output of the left child node while W_R is the output of the right child node. When splitting node L to left child LL and right child LR, only the splits that lead to $W_{LL} \le W_{LR}$ will be considered, where W_{LL} is the output of LL and W_{LR} is the output of LR. As the weight of LR should be no more than the weight of node R, we further impose an upper bound for W_{LR} as $W_{LR} \le \text{mean}(W_L, W_R) = \frac{W_L + W_R}{2} \le W_R$. Similarly, apart from ensuring $W_{RL} \le W_{RR}$, where W_{RL} is the output of RL and W_{RR} is the output of RR, we also impose a lower bound for W_{RL} as $W_{RL} \ge \text{mean}(W_L, W_R) = \frac{W_L + W_R}{2} \ge W_L$. Consequently, in this tree level we can guarantee $W_{LL} \le W_{LR} \le \text{mean}(W_L, W_R) \le W_{RL} \le W_{RR}$. As a tree is split in a recursive manner, the monotonicity of the whole tree can be guaranteed.

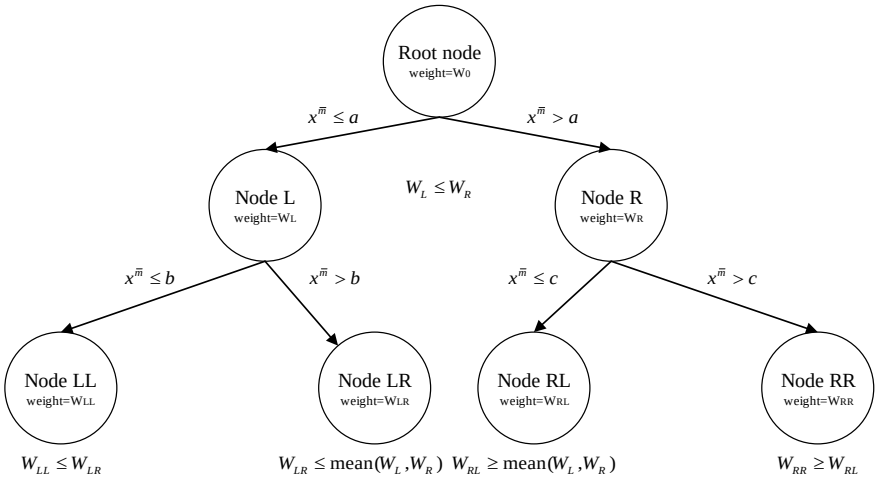


Figure 3-2. Illustration of feature monotonicity in XGBoost

It should also be noted that as XGBoost allows for feature subsampling when constructing each tree, the monotonic feature may not be included in some trees. For those trees, as the samples in I' have the same feature values except for the feature with monotonic constraint, all the samples will be assigned to the same leaf node and thus have the same output. As XGBoost is an additive model, the predicted output of each sample is the sum of the outputs in all the trees where the monotonicity constraints are preserved. Therefore, the feature monotonicity of the final output in the whole model can be preserved.

3.4.3 Construction of monotonic XGBoost

The whole dataset is randomly divided into training set (80% samples) and test set (20% samples, denoted by test set i), which contain 1,524 samples and 384 samples, respectively. The XGBoost model with monotonic constraints enforced on three features, i.e., ship flag, RO and company performance, is constructed using the training set (which we call monotonic XGBoost). Hyperparameters contained in XGBoost are in three categories: a. general parameter, which guides the overall functioning; b. booster parameters, which guide the individual booster at each iteration; and c. learning task parameter, which guides the optimization performed. We use regression decision tree as the weak learner in XGBoost model. The hyperparameters tuned in this study are summarized in Table 3-2.

Table 3-2. Hyperparameters in XGBoost model

Hyperparameter*	Meaning
learning_rate (c**)	Step size shrinkage used to update the predicted values after each boosting step to prevent overfitting which can be applied to Eq. (3.6).
n_estimators (a)	The number of weak learners (decision trees) in the XGBoost model (i.e. K in Eq. (3.3)).
max_depth (b)	The maximum depth of each tree.
min_child_weight (b)	The minimum sum of sample weight (Hessian) (i.e., H_j^t) needed in a child node. In a regression tree with loss function as MSE, the sum of sample weight in a node equals the number of samples contained in the node.
delta (b)	The minimum loss reduction required to make a split for a node.
sub_sample (b)	The fraction of samples to be randomly sampled for each tree.
colsample_bytree (b)	The fraction of columns (features) to be randomly sampled for each tree.
reg_gamma (b)	L_1 regularization term on tree complexity (i.e., γ in Eq. (3.11)).
reg_lambda (b)	L_2 regularization term on tree complexity (i.e., λ in Eq. (3.11)).

Note*: to avoid ambiguity, we have renamed some hyperparameters. For example, in the XGBoost Module for Python , '*delta*' is called '*lambda*', and '*reg_gamma*' is called '*reg_alpha*'.
Note**: this indicates the hyperparameter category.

Table 3-2 shows that there are totally nine hyperparameters that need to be tuned in an XGBoost model, which can be a huge burden if we apply cross validation with grid search to tune the hyperparameter tuple directly. To address this issue, we propose a three-step hyperparameter tuning method after giving the initial values of the hyperparameters based on experience. In the first step, the hyperparameters are tuned in turns according to their categories using grid search based on 5-fold cross validation with MSE as the metric, and their initial tuned values can be found. In the second step, an extended searching space for all the hyperparameters consisting of the initial tuned value and two more candidate values near the tuned value for each hyperparameter is formed. Then, grid search based on 5-fold cross validation with MSE as the metric is conducted on all hyperparameters simultaneously. In the third step, '*learning_rate*' is further reduced and '*n_estimators*' is further increased to improve model generalization ability. The finally adopted values for the hyperparameters are shown in Table 3-3.

Table 3-3. Finally adopted hyparameter values in monotonic XGBoost

Hyperparameter	*n_estimators*	*learning_rate*	*max_depth*	*min_child_weight*	*delta*
value	200	0.02	5	4	0.15
Hyperparameter	*sub_sample*	*colsample_bytree*	*reg_gamma*	*reg_lambda*	
value	0.75	0.4	0.1	0.1	

After hyperparameter tuning using cross validation on the training set, the final monotonic XGBoost model is constructed using the whole training set with the optimal hyperparameter values presented in Table 3-3. Its performance is validated by test set i. The MAE of the monotonic XGBoost model is 2.372 and the MSE is 12.470.

3.4.4 Analysis of monotonic XGBoost

We form another test set (denote by test set ii) as an extension of test set i to validate the monotonicity in the output of the monotonic XGBoost model regarding the three monotonic features: flag performance, RO performance, and company performance. For each sample in test set i, we form 10 variant samples by setting the values for flag performance from 1 to 3 (i.e. from white to black), RO performance from 1 to 3 (i.e. from high to low), and company performance from 1 to 4 (i.e. from high to very low) respectively while keeping the other features and their values unchanged. Totally we can have 3,840 samples ($3{,}840 = 384 \times 10$) in test set ii. We use a random sample in test set i as an example to show the construction process and the predicted results using the normal XGBoost model and the monotonic XGBoost model. Sample features except for flag, RO, and company performance are shown in Table 3-4. The flag, RO and company performance together with the prediction results are shown in Table 3-5.

Table 3-4. Features of an example in test set i except for flag, RO, and company performance

Feature	Value
age	12
GT	6813
length	132.6
depth	9.2
beam	19.2
type	container ship
last inspection date	4.3
last deficiency number	6
total detentions	0
the number of flag changes	0
casualty in the last 5 years	0

Table 3-5. An example of construction variant samples and the prediction results

Sample	flag	RO	company	Output of monotonic XGBoost	Increase between consecutive values	Output of normal XGBoost	Increase between consecutive values
Original sample	1	1	3	5.3443 (true: 5)	\	5.6563 (true: 5)	\
variant sample 1	**1**	1	3	5.3443	\	5.6563	\
variant sample 2	**2**	1	3	5.9879	0.6437	5.9409	0.2846
variant sample 3	**3**	1	3	6.3320	0.3441	5.8450	**-0.0959**
variant sample 4	1	**1**	3	5.3443	\	5.6563	\
variant sample 5	1	**2**	3	5.5915	0.2473	5.6383	**-0.0180**
variant sample 6	1	**3**	3	5.5915	0	5.6383	0
variant sample 7	1	1	**1**	3.9397	\	3.9384	\
variant sample 8	1	1	**2**	4.6101	0.6703	4.4111	0.4728
variant sample 8	1	1	**3**	5.3443	0.7342	5.6563	1.2452
variant sample 10	1	1	**4**	7.2423	1.8981	7.5755	1.9192

Table 3-5 indicates that in the monotonic XGBoost model, the predicted deficiency number increases as the performance of flag, RO and company gets worse, respectively. Moreover, increase between consecutive states of company performance is most significant in this example: on average, 1.1009 more deficiencies can be detected if it gets worse by one state. Meanwhile, change in RO performance is the least obvious: when its RO performance change from 1 (high) to 2 (medium), only 0.2473 more deficiencies will be detected; the number of detected deficiencies remains unchanged while the RO performance changes from 2 (medium) to 3 (low). Meanwhile, it can also be seen in Table 3-5 that in a normal XGBoost model, the monotonicity of the three features cannot be fully guaranteed: when flag performance changes from medium to low, and when RO performance changes from high to medium, the predicted deficiency number decreases instead, which is against domain knowledge.

We further calculate the average increase between consecutive states of each feature over the whole test set as shown in Table 3-6.

Table 3-6. Increase in predicted deficiency number of consecutive states in test set ii

State change	Flag performance	RO performance	Company performance
1->2	0.8030	0.2530	0.5312
2->3	0.2236	0	0.7787
3->4	\	\	1.4919

Table 3-6 indicates that when the states of flag performance change from high to medium and from medium to low, the increase of deficiency number gets smaller. While the state values increase by 1 in company performance, the increase of deficiency number gets larger. On the contrary, when RO performance gets from 2 (medium) to 3 (low), 0 more deficiency number will be detected as suggested by the monotonic XGBoost model. This is because there is only one sample in the training set with RO performance as low, which makes it hard for the model to capture the change in deficiency number when RO performance gets from medium to low. It should also be noted that although in Tokyo MoU the worst performance for RO is “very low”, as there are no such inspection records between 2016 and 2018, we only form variant samples with RO performance to be high, medium, and low.

3.4.5 Comparison with other popular ML models

We compare the performance of the other popular ML models with the monotonic XGBoost model using test set i and the same training set. Especially, we compare the performance of normal XGBoost, CART based regression decision tree (DT) (Breiman et al., 1984), random forest (RF) (Breiman, 2001), gradient boosting decision tree (GBDT) (Friedman, 2001), monotonic light gradient boosting machine (LightGBM) (Ke et al., 2017), least absolute shrinkage and selection operator (LASSO) regression (Santosa and Symes, 1986), ridge regression (Hoerl and Kennard, 1970), and support vector machine (SVM) (Drucker et al., 1996) with the monotonic XGBoost model. It should be noted that apart from LightGBM, none of the other ML models can guarantee the monotonic constraints of the three features. For SVM, DT, RF, LASSO regression and ridge regression, grid search with 5-fold cross validation is applied directly for hyperparameter tuning as they have fewer hyperparameters. For normal XGBoost, GBDT and monotonic LightGBM, the hyperparameter tuning method is similar to that used in the monotonic XGBoost model. The MSE and MAE in test set i are shown in Table 3-7.

Table 3-7. MSE and MAE in test set i of the ML models

Model	monotonic XGBoost*	normal XGBoost	DT	RF	GBDT	monotonic LightGBM*	LASSO regression	ridge regression	SVM
MSE	**12.470**	12.779	15.625	13.612	13.322	12.747	15.089	15.765	13.421
Rank	**1**	3	8	6	4	2	7	9	5
MAE	**2.372**	2.422	2.672	2.459	2.461	2.475	2.806	2.909	2.411
Rank	**1**	3	7	4	5	6	8	9	2

Note*: monotonicity of the three features can be preserved.

Table 3-7 shows that the prediction performance of monotonic XGBoost ranks first regarding both MSE and MAE among all the ML models considered. Regarding MSE, monotonic LightGBM ranks second, followed by normal XGBoost. Regarding MAE, SVM is slightly worse than monotonic XGBoost, followed by normal XGBoost. Ridge regression has the worst performance regarding both metrics. Especially, the monotonic XGBoost performs better than the normal XGBoost whose hyperparameter values are tuned by the same hyperparameter tuning method regarding both MSE and MAE, which is in line with the comment that if reasonable monotonic constraints on certain features are enforced, model prediction performance should be improved, meaning that the constrained models may generalize better (Sill, 1997; Duivesteijn and Feelders, 2008; Daniels and Velikova, 2010; Pei et al., 2016).

To conclude, a tree-based gradient boosting machine called XGBoost, where shipping domain knowledge regarding ship flag/RO/company performance for ship risk prediction in PSC inspection can be incorporated in a natural and rational way, is developed and validated in this section. The structure of XGBoost and detailed steps to develop an XGBoost model, especially how to incorporate monotonic constraints in the model are first introduced. The performance of the developed XGBoost model is then validated and compared with other popular ML models. It is shown that the XGBoost model considering domain knowledge has the best performance among all the ML models concerned.

3.5 PSCO SCHEDULING PROBLEM

The PSCO scheduling model aims to assign the available PSCOs to inspect the foreign visiting ships that need to be inspected as required (i.e., ships with no previous inspection records and ships out of/within the inspection time window). Human and time inspection resources, the predicted deficiency condition of the ships, and the berthing time of the ships at port should be considered in the model. As there are many foreign ships visiting a port for each day while the inspection resources are scarce, the PSCO scheduling model aims to decide the set of ships to be inspected and assign the selected ships to the PSCOs so as to maximize the inspection benefit, which is represented by the total number of deficiencies that can be identified.

Denote the set of foreign ships that need to be inspected on one day as S and one ship as $s \in S$. Denote the set of PSCOs on duty for this day as P and one PSCO as $p \in P$. The work time for the PSCOs is stable for each day: they work from 8:00 to 11:00 in the morning, and 14:00 to 17:00 in the afternoon. They spend one hour for lunch break during 11:00 to 14:00, and the other two hours for working. For example, if PSCO p has lunch break during 12:00 to 13:00, his/her work time should be from 8:00 to 12:00 and from 13:00 to 17:00. A typical PSC inspection takes about 2 hours, and thus we assume the duration of a PSC inspection to be two hours for all ships. For ship $s \in S$, its deficiency number d_s is predicted by the monotonic XGBoost model which should be treated as a parameter. Each ship berths at the port for a period in each day, and the available time for ship s during 8:00 to 17:00 (i.e., the daily work time for PSCOs) for PSC inspection is reported to the port state in advance. We divide the work hours from 8:00 to 17:00 for PSCOs into $T = 18$ time units with each time

unit as 0.5 hour, indexed by τ. The relationship between the time periods and the time units is illustrated in Table 3-8.

Table 3-8. Relationship between time periods and units

Time period	Time unit	Time period	Time unit	Time period	Time unit
8:00 to 8:30	1	11:00 to 11:30	7[a]	14:00 to 14:30	13
8:30 to 9:00	2	11:30 to 12:00	8	14:30 to 15:00	14
9:00 to 9:30	3	12:00 to 12:30	9[b]	15:00 to 15:30	15
9:30 to 10:00	4	12:30 to 13:00	10	15:30 to 16:00	16
10:00 to 10:30	5	13:00 to 13:30	11[c]	16:00 to 16:30	17
10:30 to 11:00	6	13:30 to 14:00	12	16:30 to 17:00	18

a: The latest time unit to start inspection before lunch break.
b: The earliest time unit to start inspection after lunch break.
c: The latest time unit to start lunch break.

Based on the ship berthing information reported in advance, we further introduce a parameter e_τ^s which is set to 1 if ship s stays at the port in the whole period of time unit τ. For example, if ship s stays at the port from 01:00 to 12:00, we should set $e_\tau^s = 1, \tau = 1,2,...,8$ and $e_\tau^s = 0, \tau = 9,10,...,18$. We assume that the inspection starting time of a ship and the lunch break starting time of a PSCO are at the beginning of one time unit. The PSCO scheduling problem aims to select the ships for inspection, to decide the inspection starting time of the selected ships, to assign the selected ships to the PSCOs, and to decide the lunch break starting time of the PSCOs to maximize the inspection benefits. The notation used in the PSCO scheduling problem is listed in Table 3-9.

Table 3-9. Notation used in the problem

Sets	
S	The set of foreign ships that need to be inspected for one day.
P	The set of PSCOs on duty for that day.
H	The set of *inspection templates.* An *inspection template* is a set of ships which is feasible to be inspected by one PSCO while guaranteeing his/her lunch break within the daily work time.
$\tilde{\mathrm{H}}$	The set of *un-dominated inspection templates*.
Indices	
s	The index for a ship in S.
p	The index for a PSCO in P.
τ	The index for a time unit.
η	The index of an *inspection template* in H.
Parameters	
T	The total number of time units for a working day.
d_s	The predicted deficiency number of ship s using the XGBoost model.
D_η	The number of deficiencies that can be detected if *inspection template* η is adopted.
e_τ^s	Binary parameter indicating whether ship s is available for inspection in time unit τ.
δ_s^η	Binary parameter indicating whether ship s is contained in *inspection template* η.
$B = [b_{s's''}]_{\lvert S\rvert\times\lvert S\rvert}$	Binary matrix indicating the relationship between each of the two ships that need to be inspected.

3.5.1 PSCO scheduling model M1

To formulate the PSCO scheduling problem, we define two types of main binary decision variables: $x_{sp\tau}$, which is set to 1 if ship s is inspected by PSCO p in time unit τ and 0, otherwise; and r_p^τ, which is set to 1 if PSCO p has lunch break in time unit τ and 0, otherwise. Besides, we also introduce three types of auxiliary binary decision variables: y_{sp}, which is set to 1 if ship s is inspected by PSCO p and 0, otherwise; σ_{sp}^τ, which is set to 1 if ship s starts to be inspected by PSCO p from time unit τ and 0, otherwise; and θ_p^τ, which is set to 1 if PSCO p starts to have lunch break from time unit τ and 0, otherwise. To maximize the inspection benefit by maximizing the estimated total number of deficiencies that can be detected, an integer linear optimization model M1 is proposed as follows.

[M1]
$$\max \sum_{s\in S}\sum_{p\in P} d_s y_{sp} \tag{3.17}$$

s.t.

$$\sum_{p\in P} y_{sp} \le 1, \forall s \in S \tag{3.18}$$

$$x_{sp\tau} \le e_\tau^s, \forall s \in S, \forall p \in P,\ \tau = 1,...,T \tag{3.19}$$

$$x_{sp\tau} \le 1 - r_p^\tau, \forall s \in S, \forall p \in P, \tau = 1,...,T \tag{3.20}$$

$$\sum_{s\in S} x_{sp\tau} \le 1,\ \forall p \in P,\ \tau = 1,...,T \tag{3.21}$$

$$\sum_{\tau=1}^{T} x_{sp\tau} = 4y_{sp},\ \forall s \in S, \forall p \in P \tag{3.22}$$

$$\sum_{\tau'=\tau}^{\tau+3} x_{sp\tau'} \ge 4\sigma_{sp}^\tau, \forall s \in S, \forall p \in P, 1 \le \tau \le 15 \tag{3.23}$$

$$\sum_{\tau=1}^{T} \sigma_{sp}^\tau = y_{sp}, \forall s \in S, \forall p \in P \tag{3.24}$$

$$\sigma_{sp}^\tau = 0, \forall s \in S, \forall p \in P, 16 \le \tau \le 18 \tag{3.25}$$

$$\sum_{\tau=7}^{12} r_p^\tau = 2,\ \forall p \in P \tag{3.26}$$

$$r_p^\tau = 0, \forall p \in P, \tau \in [1,6] \cup [13,18] \tag{3.27}$$

$$\sum_{\tau'=\tau}^{\tau+1} r_p^{\tau'} \ge 2\theta_p^\tau, \forall p \in P, 7 \le \tau \le 11 \tag{3.28}$$

$$\sum_{\tau=1}^{T} \theta_p^\tau = 1,\ \forall p \in P \tag{3.29}$$

$$\theta_p^\tau = 0, \forall p \in P, \tau \in [1,6] \cup [12,18] \tag{3.30}$$

$$x_{sp\tau} \in \{0,1\}, \forall s \in S, \forall p \in P,\ \tau = 1,...,T \tag{3.31}$$

$$y_{sp} \in \{0,1\}, \forall s \in S,\ \forall p \in P \tag{3.32}$$

$$\sigma_{sp}^\tau \in \{0,1\}, \forall s \in S,\ \forall p \in P, \tau = 1,...,T \tag{3.33}$$

$$r_p^\tau \in \{0,1\}, \forall p \in P,\ \tau = 1,...,T \tag{3.34}$$

$$\theta_p^\tau \in \{0,1\}, \forall p \in P,\ \tau = 1,...,T. \tag{3.35}$$

Objective function (3.17) maximizes the inspection benefits by maximizing the estimated total number of deficiencies that can be detected. Constraints (3.18) ensure that each ship can only be inspected by at most one PSCO. Constraints (3.19) and (3.20) guarantee that a ship can only be inspected when it is at port and when the corresponding PSCO does not have lunch break. Constraints (3.21) ensure that a PSCO can only inspect one ship in one time unit. Constrains (3.22) to (3.25) guarantee that if a ship is inspected, it should be inspected during 4 consecutive time units, and the start inspection time unit is between 1 and 15. Constraints (3.26) to (3.30) guarantee that each PSCO can have a one-hour consecutive lunch break between time units 7 and 12. Constraint (3.31) to (3.35) ensure the domain of the decision variables.

3.5.2 PSCO scheduling model M2

As the PSCOs are indifferent from each other, there will be an exponential number of optimal solutions to mathematical model M1, which will reduce the efficiency to solve M1. As the total work time of a PSCO for one day is 8 hours and an inspection would take 2 hours, a PSCO can inspect 0, 1, 2, 3, or 4 ships for one day. Therefore, the PSCO scheduling problem can be reformulated as identifying and assigning the sets of ships that can be inspected by one PSCO to the available PSCOs. Define L as the number of ships inspected by one PSCO, $L = 0,1,2,3,4$. Given the value for L, the total number of combinations of L ships from the total $|S|$ ships is $C_{|S|}^L$, $C_{|S|}^L = \binom{|S|}{L} = \frac{|S|!}{L!(|S|-L)!}$. Given a combination of L ships, denoted by set S', $S' \subset S$, $|S'| = L$. we examine whether it is feasible to inspect all the ships in S' by one PSCO. If it is feasible, then we call set S' an *inspection template* and our aim is to choose $|P|$ *inspection templates* (each template is assigned to one PSCO) that maximize the total number of deficiencies that can be detected while ensuring a ship is included in at most one chosen template (i.e., a ship is inspected at most once). Here

the concept of "template" is similar with the concept of berth template (Zhen, 2015) and yard template (Zhen 2016), which have been widely used in some pioneering work such as Zhen et al. (2011) in the field of port and shipping management.

To examine whether it is feasible to inspect all the ships in S' by one PSCO, $|S'|=L$, we note that a PSCO has to carry out $L+1$ activities to inspect all the ships in S' between time unit $\tau=1$ (8:00) and $\tau=18$ (17:00), that is, inspecting each of the L ships and having lunch break. We define α as an activity, and each activity has a duration t_α, an earliest start time ω_α, and a latest completion time ϖ_α. If an activity α is inspecting a ship, denoted by ship s, then $t_\alpha=4$, ω_α is the start time of the ship's berthing between $\tau=1$ and $\tau=18$ of the day, i.e., $\omega_\alpha=\min\{\tau=1,...,18\,|\,e_\tau^s=1\}$, ϖ_α is the ship's departure time if it departs before $\tau=18$ and otherwise $\varpi_\alpha=18$, i.e., $\varpi_\alpha=\max\{\tau=1,...,18\,|\,e_\tau^s=1\}$; if an activity α is having lunch break, then $t_\alpha=2$, $\omega_\alpha=7$, $\varpi_\alpha=12$. There are a total of $(L+1)!$ different sequences for the PSCO to conduct the activities (note that some, or even all of the sequences may be infeasible). For a particular sequence, we denote the activities carried out by $\alpha_1\to\alpha_2\to...\to\alpha_{L+1}$, that is, α_ℓ is the ℓth activity, t_{α_ℓ}, ω_{α_ℓ}, ϖ_{α_ℓ} are the duration, earliest start time, and latest completion time of activity α_ℓ, respectively. To check whether the $L+1$ activities can be carried out in the above sequence, we define $\mathcal{T}_\ell$ as decision variable representing the start time of carrying out activity α_ℓ, then the $L+1$ activities can be carried out in the above sequence by one PSCO if and only if there is a solution $\mathcal{T}_\ell$, $\ell=1,...,L+1$, that satisfies the following constraints:

$$\mathcal{T}_\ell\ge\omega_{\alpha_\ell},\ell=1,...,L+1 \tag{3.36}$$

$$\mathcal{T}_\ell+t_{\alpha_\ell}-1\le\varpi_{\alpha_\ell},\ell=1,...,L+1 \tag{3.37}$$

$$\mathcal{T}_{\ell+1}\ge\mathcal{T}_\ell+t_{\alpha_\ell}\ \ell=1,...,L\,. \tag{3.38}$$

Note that if an activity α_ℓ with duration t_{α_ℓ} starts at the beginning of time unit $\mathcal{T}_\ell$, its completion time should be at the end of time unit $\mathcal{T}_\ell+t_{\alpha_\ell}-1$.

Proposition 1: For an activity sequence, whether constraints (3.36)–(3.38) have a feasible solution can be checked below: for activity α_1, let its start time $\mathcal{T}_1^*=\omega_{\alpha_1}$; for activity α_l, $l=2,...,L+1$, let its start time $\mathcal{T}_\ell^*=\max\{\mathcal{T}_{\ell-1}^*+t_{\alpha_{\ell-1}},\omega_{\alpha_\ell}\}$, $\ell=2,...,L+1$; if

$\mathcal{T}_\ell^* \le \varpi_{\alpha_\ell} - t_{\alpha_\ell} + 1$, $l = 1,...,L+1$, then the *activity sequence* is feasible; otherwise it is infeasible.

Proof:

The "if" part of the proposition is straightforward because it is easy to check that $(\mathcal{T}_\ell^*, \ell = 1,...,L+1)$ is indeed feasible to constraints (3.36)–(3.38). To prove the "only if" part, suppose that constraints (3.36)–(3.38) have a feasible solution $(\mathcal{T}_\ell^\#, \ell = 1,...,L+1) \neq (\mathcal{T}_\ell^*, \ell = 1,...,L+1)$. Denote by $\hat{\ell}$ the index of the first different elements of vectors $(\mathcal{T}_\ell^\#, \ell = 1,...,L+1)$ and $(\mathcal{T}_\ell^*, \ell = 1,...,L+1)$, that is $\mathcal{T}_\ell^\# = \mathcal{T}_\ell^*$, $\ell = 1,...,\hat{\ell}-1$ and $\mathcal{T}_{\hat{\ell}}^\# \neq \mathcal{T}_{\hat{\ell}}^*$. If $\hat{\ell} = 1$, we define a new vector $(\mathcal{T}_\ell^\&, \ell = 1,...,L+1)$ such that $\mathcal{T}_1^\& = \omega_{\alpha_1}$ and $\mathcal{T}_\ell^\& = \mathcal{T}_\ell^\#$, $\ell = 2,...,L+1$. If $\hat{\ell} = 2,...,L+1$, we define a new vector $(\mathcal{T}_\ell^\&, \ell = 1,...,L+1)$ such that $\mathcal{T}_\ell^\& = \mathcal{T}_\ell^\#$, $\ell = 1,...,\hat{\ell}-1$, $\mathcal{T}_{\hat{\ell}}^\& = \max\{\mathcal{T}_{\hat{\ell}-1}^\& + t_{\alpha_{\hat{\ell}-1}}, \omega_{\alpha_{\hat{\ell}}}\}$ and $\mathcal{T}_\ell^\& = \mathcal{T}_\ell^\#$, $\ell = \hat{\ell}+1,...,L+1$. In both cases, it is easy to check that $(\mathcal{T}_\ell^\&, \ell = 1,...,L+1)$ is feasible to constraints (3.36)–(3.38). We can now set $(\mathcal{T}_\ell^*, \ell = 1,...,L+1) \leftarrow (\mathcal{T}_\ell^\&, \ell = 1,...,L+1)$ and repeat the above procedure. It can be seen that by repeating the above procedure at most $L+1$ times, we will generate a feasible solution $(\mathcal{T}_\ell^\&, \ell = 1,...,L+1)$ that is identical to $(\mathcal{T}_\ell^*, \ell = 1,...,L+1)$. In other words, constraints (3.36)–(3.38) have a feasible solution only if $(\mathcal{T}_\ell^*, \ell = 1,...,L+1)$ is feasible. This concludes the proof of the proposition. □

We use the following example to illustrate the steps to decide whether an *activity sequence* $\alpha_1 \to \alpha_2 \to ... \to \alpha_{L+1}$ is feasible.

Example 1. Given $L = 3$ and $S' = \{s_1, s_2, s_3\}$ for *activity sequence* $\alpha_1 \to \alpha_2 \to \alpha_3 \to \alpha_4$. Particularly, activities α_1, α_2, and α_4 are ship inspections for s_1, s_2, and s_3 respectively and activity α_3 is lunch break. The berthing periods of s_1, s_2, and s_3 are during 8:00 to 13:00, 9:00 to 18:30, and 13:00 to 17:30, respectively. Therefore, we have $\omega_{\alpha_1} = 1$ and $\varpi_{\alpha_1} = 10$ for activity α_1, $\omega_{\alpha_2} = 3$ and $\varpi_{\alpha_2} = 18$ for activity α_2, $\omega_{\alpha_3} = 7$ and $\varpi_{\alpha_3} = 12$ for activity α_3, and $\omega_{\alpha_4} = 11$ and $\varpi_{\alpha_4} = 18$ for activity α_4. The earliest start time of each activity should be $\mathcal{T}_1^* = \omega_{\alpha_1} = 1$, $\mathcal{T}_2^* = \max\{\mathcal{T}_1^* + 4, \omega_{\alpha_2}\} = 5$, $\mathcal{T}_3^* = \max\{\mathcal{T}_2^* + 4, \omega_{\alpha_3}\} = 9$, and $\mathcal{T}_4^* = \max\{\mathcal{T}_3^* + 2, \omega_{\alpha_4}\} = 11$. The earliest start time of each activity satisfies $\mathcal{T}_1^* \le \varpi_{\alpha_1} - 4 + 1 = 7$, $\mathcal{T}_2^* \le \varpi_{\alpha_2} - 4 + 1 = 15$, $\mathcal{T}_3^* \le \varpi_{\alpha_3} - 2 + 1 = 11$, and

$\mathcal{T}_4^* \le \varpi_{\alpha_4} - 4 + 1 = 15$, and thus the *activity sequence* is feasible and S' is an *inspection template*.□

Proposition 2: Given a combination of $L = 1,2,3$ ships denoted by $\bar{S}$, if it is not an *inspection template*, any set of $L+1$ ships (denoted by $\hat{S}$) containing all ships in $\bar{S}$, i.e. $\bar{S} \subset \hat{S}$ cannot be an *inspection template*.

Proposition 3: Given an *inspection template* containing $L = 2,3,4$ ships denoted by $\bar{S}$, all subsets of $\bar{S}$ containing $L' = L - 1$ ships are *inspection templates*.

Proposition 2 and **Proposition 3** are the basis of *inspection template* construction. They are intuitive and thus we omit their proof. Based on the two propositions, the following two properties of *inspection templates* can be derived to reduce the trials and the total number of generated *inspection templates*.

Property 1: If there is a combination with $L = 1$ ship associated with berthing period smaller than 4 time units or from time unit 8 to time unit 11, or with zero predicted deficiency number, we can simply ignore it as it cannot be inspected during its berthing period or inspecting the ship will not bring benefits.

Property 2: Candidate $L \ge 2$ *inspection templates* can be formulated by combining all pairs of $L-1$ *inspection templates* with the first $L-2$ items the same.

Procedure 1: generation of the set of *inspection templates* H

Input: the set of foreign ships that need to be inspected S, duration of a PSC inspection, ship berthing information e_τ^s, $s \in S$, $\tau = 1,...,\mathrm{T}$, the duration and period of PSCO lunch break, the total number of time units T.

Output: the set of all feasible *inspection templates* H, binary parameter δ_s^η, $s \in S$, $\eta \in \mathrm{H}$ indicating whether ship s is contained in *inspection template* η.

Initialize $\mathrm{H} = \varnothing$, δ_s^η, $s \in S$, $\eta \in \mathrm{H}$.

for $L = 0,1,2,3,4$ do

  if $L < 2$:

    Formulate all combinations containing L ships that can be inspected based on **Property 1** from S denoted by $\mathbb{Q}$.

  else:

    Formulate the combinations containing L ships as denoted by $\mathbb{Q}$ such that each combination contains two items of $L-1$ ships from H and they have the same $L-2$ ship based on **Property 2**.

  end if

  for each combination $\hat{Q} \in \mathbb{Q}$ do

    Initialize feasibility = False.

    Formulate set Q that contains all permutations (i.e. *activity sequences*) of the activities of inspecting the ships in $\hat{Q}$ and having lunch break.

    for each *activity sequence* $q \in Q$ do

      Test the feasibility of q using **Proposition 1**.

      if q is feasible:

        Add $\hat{Q}$ to H by updating $\mathrm{H} = \mathrm{H} \cup \hat{Q}$.

        Update parameter $\delta_s^\eta = 1$, $s \in \hat{Q}$, $\eta \in \mathrm{H}$.

        Update feasibility = True.

        break

      else:

        continue

      end if

    end for

    if feasibility = True:

      break

    else:

      continue

  end for

end for

Return H and δ_s^η.

Property 1 and **Property 2** can highly improve the efficiency of *inspection templates* generation. The overall procedure to generate the set of all *inspection templates* (denoted by H) is shown in Procedure 1.

After obtaining H and δ_s^η, $s \in S$, $\eta \in \mathrm{H}$ by executing Procedure 1, the estimated number of deficiencies that can be detected in *inspection template* η is $D_\eta = \sum_{s \in S} \delta_s^\eta d_s$, $\eta \in \mathrm{H}$. To assign the *inspection templates* to the PSCOs, we introduce binary decision

variable z_η which is set to 1 if *inspection template* $\eta \in \mathrm{H}$ is adopted and 0, otherwise. The PSCO scheduling problem aiming to maximize the total number of detected deficiencies based on *inspection templates* can be formulated by mathematical model M2.

[M2]

$$\max \sum_{\eta \in \mathrm{H}} D_\eta z_\eta \tag{3.39}$$

s.t.

$$\sum_{\eta \in \mathrm{H}} z_\eta \le |P| \tag{3.40}$$

$$\sum_{\eta \in \mathrm{H}} \delta_s^\eta z_\eta \le 1, \forall s \in S \tag{3.41}$$

$$z_\eta \in \{0,1\}, \forall \eta \in \mathrm{H}\,. \tag{3.42}$$

Objective function (3.39) maximizes the estimated total number of deficiencies that can be detected. Constraint (3.40) ensures that the total number of adopted *inspection templates* should be no more than the total number of PSCOs. Constraints (3.41) guarantee that each ship can only be inspected at most once.

3.5.3 PSCO scheduling model M3

Model M2 considers all the *inspection templates* in H indifferently, which is time-consuming when $|\mathrm{H}|$ is large. Meanwhile, it is noted that if we reformulate constraints (3.41) which require that a ship can only be inspected at most once, we can only consider the *inspection templates* that are not contained in any other *inspection template(s)*, which we denote by *un-dominated inspection templates*, as inspecting them can always detect more deficiencies than inspecting the *inspection templates* contained in them according to **Property 1**. In this way, the number of *inspection templates* considered in the PSCO scheduling optimization model can be reduced largely. However, one problem is that the number of deficiencies of one ship might be calculated several times in the objective function of M2 as it can be contained in several *inspection templates* selected by a solution. To overcome this issue, we further introduce binary decision variables ξ_s, $s \in S$ which is set to 1 if ship s is inspected and 0, otherwise. In addition, we form set $\tilde{\mathrm{H}}$ which contains all *un-dominated inspection templates* using **Procedure 1** by adding only the *inspection templates* with $L = 1,2,3$ that cannot be further combined with others to generate larger valid *inspection templates* and all the *inspection templates* with $L = 4$. Mathematical model M3 is developed based on *inspection template* set $\tilde{\mathrm{H}}$ and decision variables ξ_s,

$s \in S$ and z_η, $\eta \in \mathrm{H}$ to reduce the number of *inspection templates* considered in the master problem as follows.

[M3]

$$\max \sum_{s \in S} d_s \xi_s \tag{3.43}$$

s.t.

$$\xi_s \le \sum_{\tilde{\eta} \in \tilde{\mathrm{H}}} \delta_s^{\tilde{\eta}} z_{\tilde{\eta}}, \forall s \in S \tag{3.44}$$

$$\sum_{\tilde{\eta} \in \tilde{\mathrm{H}}} z_{\tilde{\eta}} \le |P| \tag{3.45}$$

$$z_{\tilde{\eta}} \in \{0,1\}, \forall \tilde{\eta} \in \tilde{\mathrm{H}} \tag{3.46}$$

$$\xi_s \in \{0,1\}, \forall s \in S\,. \tag{3.47}$$

Like Eq. (3.39), objective function (3.43) also maximizes the total estimated number of deficiencies that can be detected. Constraints (3.44) indicate the relationship between ξ_s and $z_{\tilde{\eta}}$. Constraints (3.45) require the maximum number of *un-dominated inspection templates* that can be selected. Constraints (3.46) and (3.47) guarantee the domain of the decision variables. It should be noted that although M3 does not require that each ship can only be inspected at most once, the objective function only calculates its estimated deficiency number once it is inspected and thus model M3 is equivalent to model M2.

To further improve the efficiency of model M3, we propose the following proposition:

Proposition 4: For two ships s_1 and s_2, if $d_{s_1} > d_{s_2}$ and $\{e_\tau^{s_2} \mid e_\tau^{s_2} = 1, \forall \tau \in \mathrm{T}\} \subseteq \{e_\tau^{s_1} \mid e_\tau^{s_1} = 1, \forall \tau \in \mathrm{T}\}$, i.e. ship s_1 has larger estimated number of deficiencies than ship s_2 and the set of berthing period of ship s_2 is a sub-set of that of ship s_1 (we denote the relationship between s_1 and s_2 by "s_1 dominates s_2"), we must have $\xi_{s_1} \ge \xi_{s_2}$ in an optimal solution.

Proof:

Consider two ships s_1 and s_2 with $d_{s_1} > d_{s_2}$ and $\{e_\tau^{s_2} \mid e_\tau^{s_2} = 1, \forall \tau \in \mathrm{T}\} \subseteq \{e_\tau^{s_1} \mid e_\tau^{s_1} = 1, \forall \tau \in \mathrm{T}\}$, i.e. s_1 dominates s_2. If an optimal solution chooses a set of ships S' for inspection, there can be several situations regarding ships s_1 and s_2:

Situation 1: if $s_1 \in S'$ and $s_2 \in S'$, $\xi_{s_1} \geq \xi_{s_2}$ is satisfied.

Situation 2: if $s_1 \notin S'$ and $s_2 \notin S'$, $\xi_{s_1} \geq \xi_{s_2}$ is satisfied.

Situation 3: if $s_1 \in S'$ and $s_2 \notin S'$, $\xi_{s_1} \geq \xi_{s_2}$ is satisfied.

Situation 4: if $s_1 \notin S'$ and $s_2 \in S'$, we can expect that another feasible set of ships $\tilde{S}'$ formulated by substituting ship s_2 by s_1 in S' can increase the value of the objective function by $d_{s_1} - d_{s_2}$ and thus S' should not be an optimal solution, which is contradictory to the given conditions. Therefore, Situation 4 cannot be a case in any optimal solution, and $\xi_{s_1} \geq \xi_{s_2}$ can always be satisfied in the optimal solution(s).□

To incorporate **Proposition 4** into model M3, we introduce a binary matrix $B = [b_{s's"}]_{|S|\times|S|}$ which can be derived directly from ship visiting information and deficiency condition to indicate whether ship $s' \in S$ dominates $s" \in S$. If s' dominates $s"$, we set $b_{s's"} = 1$; otherwise, $b_{s's"} = 0$. Especially, we require $b_{s's"} = 0$ if $s' = s"$. The following strengthened constraints based on B can be added to M3 to improve its efficiency:

$$\xi_{s'} - \xi_{s"} \geq b_{s's"} - 1, s' \in S, s" \in S \cdot \tag{3.48}$$

3.6 COMPUTATIONAL EXPERIMENTS

We take Hong Kong port as an example to validate the proposed PSCO scheduling models M1, M2, and M3. Particularly, we first compare the computing performance of the three models in section 3.6.1. Then, comparisons between the current and proposed PSCO scheduling models are conducted in section 3.6.2. In section 3.6.3, results of extensive sensitivity analysis are presented to further validate the proposed models.

3.6.1 Comparison of computing performance of M1, M2, and M3

To compare the computing performance of M1, M2, and M3 (including generation of all *inspection templates*, *un-dominated inspection templates*, and binary matrix $B = [b_{s's"}]_{|S|\times|S|}$), we set the number of PSCOs to 4, 6, 8 and 10, and the number of ships that need to be inspected to 30, 40, 50, and 60 and combine them one by one in several scenarios. Ships for inspection are selected from test set i and the number of deficiencies of them is predicted by the XGBoost model developed in Section 3.4. We

assume that a ship can arrive at a port at any time during a day, and their staying period ranges from 0 to 18 consecutive time units from 8:00 to 17:00. As all the PSCO scheduling models M1, M2, and M3 are integer linear programming (ILP) models, they are solved by the off-the-shelf optimization solver CPLEX. In addition, we compare the performance of PSCO scheduling decisions generated by M1, M2, and M3 with the current greedy PSCO scheduling strategy applied at the Hong Kong port, whose detailed description is presented in Appendix A. We call the current scheduling strategy "random scheduling case", and it aims to assign as many ships as possible to each available PSCO for inspection in a greedy manner. Besides, we present the performance of the proposed PSCO scheduling model utilizing the predicted deficiency number from a perfect-foresight prediction which knows the actual deficiency number of all ships in advance (denoted by "perfect-forecast policy"). The identified deficiency number based on the perfect-foresight policy is an upper bound in theory which cannot be achieved.

All experiments are conducted on a laptop (Intel Core i7, 3.40 GHz, 16GB RAM) using programming language Python. The *inspection templates* in M2 are generated using Procedure 1, and the *un-dominated inspection templates* in M3 are generated based on Procedure 1. Table 3-10 summarizes the computing performance of the three models, including the average computation time (in CPU seconds), the standard deviation of computation time, the number of *inspection templates* generated, the number of *un-dominated inspection templates* generated, the reduction in percentage of the number of *un-dominated inspection templates* compared to that of all the *inspection templates*, the average improvement of M1/M2/M3 over random scheduling case, and the average gap between M1/M2/M3 and the perfect-forecast policy in all cases of each scenario.

Table 3-10. Comparison of computing performance of M1, M2, and M3

No. of PSCOs	Scenarios	Model	Number of ships			
			30	40	50	60
4	Average total computation time*	M1	5.75	7.80	10.18	13.35
		M2	0.48	2.30	9.11	25.47
		M3	0.46	2.09	7.95	23.89
	Standard deviation of computation time*	M1	4.21	2.69	7.98	7.51
		M2	0.18	0.74	3.52	8.65
		M3	0.16	0.64	2.87	8.01
	The number of *inspection templates* in H	M2	1883.0	5465.8	13834.8	26972.7
	The number of *un-dominated inspection templates* in $\tilde{H}$	M3	1189.6	3871.7	10456.4	21226.0
	$(\lvert\tilde{H}\rvert - \lvert H\rvert) / \lvert H\rvert \times 100\%$	\	36.82%	29.16%	24.42%	21.31%
	Average improvement of M1/M2/M3 over random scheduling case	\	22.10%	34.57%	35.56%	43.72%
	Average gap between M1/M2/M3 and the perfect-forecast policy	\	9.64%	11.24%	16.28%	17.52%
6	Average total computation time	M1	19.16	30.85	36.26	47.98
		M2	0.47	2.32	8.58	25.06
		M3	0.46	2.07	7.83	23.25
	Standard deviation of computation time	M1	10.74	13.49	23.90	28.24
		M2	0.18	0.70	3.37	8.13
		M3	0.15	0.63	3.16	6.61
	The number of *inspection templates* in H	M2	1883.0	5465.8	13834.8	26972.7
	The number of *un-dominated inspection templates* in $\tilde{H}$	M3	1189.6	3871.7	10456.4	21226.0
	$(\lvert\tilde{H}\rvert - \lvert H\rvert) / \lvert H\rvert \times 100\%$	\	36.82%	29.16%	24.42%	21.31%
	Average improvement of M1/M2/M3 over random scheduling case	\	13.93%	20.17%	24.57%	29.81%
	Average gap between M1/M2/M3 and the perfect-forecast policy	\	5.88%	7.83%	10.90%	12.95%
8	Average total computation time	M1	31.97	64.18	102.95	195.14
		M2	0.52	2.21	8.22	24.78
		M3	0.49	2.00	7.66	24.04
	Standard deviation of computation time	M1	57.54	48.71	70.15	257.11
		M2	0.23	0.67	3.08	7.39
		M3	0.23	0.62	2.83	8.03
	The number of *inspection templates* in H	M2	1883.0	5465.8	13834.8	26972.7
	The number of *un-dominated inspection templates* in $\tilde{H}$	M3	1189.6	3871.7	10456.4	21226.0
	$(\lvert\tilde{H}\rvert - \lvert H\rvert) / \lvert H\rvert \times 100\%$	\	36.82%	29.16%	24.42%	21.31%
	Average improvement of M1/M2/M3 over random scheduling case	\	6.04%	13.74%	18.32%	24.73%
	Average gap between M1/M2/M3 and the perfect-forecast policy	\	4.19%	5.94%	7.52%	8.97%
10	Average total computation time	M1	89.53	614.45	295.13	377.39
		M2	0.49	2.24	8.53	25.22
		M3	0.43	2.05	7.46	22.36
	Standard deviation of computation time	M1	195.82	1492.04	340.93	366.52
		M2	0.19	0.67	3.39	7.80
		M3	0.17	0.64	2.73	6.44
	The number of *inspection templates* in H	M2	1883.0	5465.8	13834.8	26972.7
	The number of *un-dominated inspection templates* in $\tilde{H}$	M3	1189.6	3871.7	10456.4	21226.0
	$(\lvert\tilde{H}\rvert - \lvert H\rvert) / \lvert H\rvert \times 100\%$	\	36.82%	29.16%	24.42%	21.31%
	Average improvement of M1/M2/M3 over random scheduling case	\	1.21%	8.07%	13.90%	18.55%
	Average gap between M1/M2/M3 and the perfect-forecast policy	\	1.40%	3.94%	5.74%	6.47%

Note*: the computation time of M2 includes the time to generate all *inspection templates*, and the computation time of M3 includes the time to generate matrix B and all *un-dominated inspection templates*.

For the average computation time, it is indicated in Table 3-10 that in almost all the cases, much less time is required to solve M2 and M3 compared to the time used to solve M1, except when the number of PSCOs is 4 and the number of ships is 60. The difference of the computation time between M1 and M2/M3 becomes larger as the number of PSCOs increases. Meanwhile, the difference of the computation time between M2 and M3 shows an increasing trend when there are more visiting ships. To be more specific, when the number of PSCOs is fixed and the number of ships increases, i.e. from left to right in each row of the table, the computation time of all the three models shows an increasing trend as expected. When the number of ships is fixed and the number of PSCOs increases, i.e. from top to bottom in each column of the table, the model computation time increases faster and faster in M1. Meanwhile, there are only some minor fluctuations in the model computation time of M2 and M3. This is because the number of PSCOs has no influence on the generation of *inspection templates* and *un-dominated inspection templates*, which occupies most of the computation time of M2 and M3, respectively.

The standard deviation of model computation time of M1 is much larger than that of M2 and M3 in most of the cases listed in Table 3-10, and M2 is a little bit larger than M3 in most cases. Particularly, in M1, when the number of ships is fixed and the number of PSCOs increases, the standard deviation of computation time shows a rapid upward trend. There is also a general upward trend in the standard deviation of computation time when the number of ships increases while the number of PSCOs remains unchanged in M1. Meanwhile, in M2 and M3 with similar pattern, the standard deviation of computation time increases dramatically when the number of ships increases given a certain number of PSCOs. When the number of PSCOs increases with a fixed number of ships, there are many fluctuations in the standard deviation of computation time of both M2 and M3.

When the number of visiting foreign ships increases from 30 to 60, the numbers of *inspection templates* and *un-dominated inspection templates* grow, while the difference between them decreases. On average, the number of *un-dominated inspection templates* considered in M3 is about 72% of the *inspection templates* considered in M2. In addition, M1/M2/M3 perform better than the currently implemented random PSCO scheduling strategy at the ports in all cases. When the number of PSCOs increases given a certain number of visiting ships, the advantage of

M1/M2/M3 over random scheduling and the advantage of perfect-forecast policy over M1/M2/M3 are reduce. When there are more visiting ships while the number of PSCOs is fixed, both the gap between M1/M2/M3 and random scheduling and the gap between perfect-forecast policy and M1/M2/M3 increase.

To summarize, the average model computation time and its standard deviation of M2 are much smaller than those of M1 in most cases, and the average total model computation time and its standard deviation of M3 are smaller than those of M2 in most cases as shown in Table 3-10. Besides, model computation time of M2 and M3 is less sensitive to the increase of the number of PSCOs given a fixed number of ships, as the process to generate *inspection templates* and *un-dominated inspection templates* is not influenced by the number of PSCOs. In all scenarios, the proposed M1/M2/M3 perform better than the current random scheduling strategy, and the gap between M1/M2/M3 and the perfect-forecast policy decreases when there are more PSCOs or fewer visiting ships. We can therefore conclude that M3 is the most efficient, stable, and flexible model among M1, M2, and M3. Especially, M3 is more suitable to be applied to the ports where there are a larger number of available PSCOs or more visiting ships.

3.6.2 Comparison of current and the proposed PSCO scheduling strategies

We compare the performance of current PSCO scheduling strategy applied at port and the proposed models in this section. For each day, we randomly select 20 ships from test set i as the visiting ships that need to be inspected at the Hong Kong port. We further assume that the number of PSCOs on duty for that day is 3, and their daily work time is fixed as mentioned in section 3.5. As M1, M2 and M3 are equivalent and section 3.6.1 shows that M3 is more efficient than M1 and M2, the following experiments are only conducted on M3. We randomly generate 30 groups of ships from test set i in the experiment. The performance of random scheduling case (average of 100 runs), M3, and the perfect-forecast policy solved by M3 and their comparisons are presented in Table 3-11.

Table 3-11. Performance and comparison of PSCO scheduling models

Group	Actual identified deficiency number of random scheduling case	Actual identified deficiency number of M3	Identified deficiency number under perfect-forecast policy as solved by M3	Improvement of M3 over random scheduling case	Gap between M3 and the perfect-forecast policy
1	16	24	30	46.6%	20.0%
2	59	68	71	15.4%	4.2%
3	36	37	43	4.0%	14.0%
4	46	74	78	61.8%	5.1%
5	56	57	65	2.3%	12.3%
6	53	58	62	10.4%	6.5%
7	27	41	41	51.5%	0.0%
8	65	73	78	12.6%	6.4%
9	55	71	72	29.6%	1.4%
10	37	41	51	9.9%	19.6%
11	43	57	65	33.8%	12.3%
12	17	25	31	46.6%	19.4%
13	59	75	79	27.0%	5.1%
14	42	47	63	12.5%	25.4%
15	42	54	56	29.6%	3.6%
16	60	69	75	15.2%	8.0%
17	59	66	66	11.6%	0.0%
18	41	48	55	15.9%	12.7%
19	39	55	59	39.8%	6.8%
20	55	66	67	20.8%	1.5%
21	61	64	68	5.4%	5.9%
22	46	49	54	7.2%	9.3%
23	34	48	49	42.5%	2.0%
24	33	45	46	36.2%	2.2%
25	32	36	37	11.6%	2.7%
26	63	77	78	22.1%	1.3%
27	29	39	47	35.4%	17.0%
28	47	55	58	18.3%	5.2%
29	51	56	65	9.5%	13.8%
30	50	61	72	22.2%	15.3%
Average	**45.0067**	**54.5333**	**59.3667**	**21.2%**	**8.1%**

Table 3-11 shows that the average improvement of M3 with the prediction of XGBoost as the input over the random PSCO scheduling case is over 20%. This implies that the combination of XGBoost model for ship deficiency number prediction and the mathematical models M1/M2/M3 for PSCO scheduling can identify 20% more deficiencies than the current PSCO scheduling scheme with the same inspection resources. Besides, the gap between the proposed model and the perfect-forecast policy is about 8%, which indicates that the proposed combined model can identify about 92% of all existing deficiencies.

3.6.3 Sensitivity analysis

In this section, we analyze how the number of ships to be inspected, the number of available PSCOs for conducting inspection, and ship berthing duration and period will influence the performance of M3 (and M1, M2). Four groups of sensitivity analysis (SA) are performed: SA1: different numbers of ships for inspection; SA2: different numbers of available PSCOs; SA3: different berthing durations of ships; SA4: different berthing periods of ships. In each group of SA, the number of deficiencies identified is calculated based on 10 runs.

SA1: different numbers of ships for inspection

First, we analyze how the number of ships that need to be inspected would influence the performance of M3. We set the number of ships to 15, 20, 25, 30, 35, 40, 45, and 50, respectively while fixing the number of PSCOs to 3 in SA1G1 to SA1G8. The performance of random scheduling case (based on 100 runs), M3, and the perfect-foresight policy and their comparison are presented in Table 3-12.

Table 3-12. Performance of the groups in SA1

Group	**SA1G1**	**SA1G2**	**SA1G3**	**SA1G4**	**SA1G5**	**SA1G6**	**SA1G7**	**SA1G8**
Number of ships	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**
Random scheduling case	34.1	44.4	48.3	52.9	54.5	55.7	55.9	57.4
M3	41.2	54.4	58.0	66.8	71.5	78.8	81.2	84.0
Perfect-foresight policy	44.3	59.1	66.1	76.6	82.9	91.4	97.7	104.8
Superiority of M3 over random scheduling case	20.9%	22.5%	20.0%	26.2%	31.2%	41.4%	45.3%	46.3%
Gap between M3 and the perfect-foresight policy	7.5%	8.6%	14.0%	14.7%	15.9%	16.0%	20.3%	24.8%

Table 3-12 shows that when the number of ships increases from 15 to 50 while the number of PSCOs remains unchanged, the numbers of deficiencies identified in random scheduling case, M3, and the perfect-foresight policy increase. This can be explained as follows. In random scheduling case which aims to assign as many ships to each PSCO as possible, a larger number of ships can be inspected by one PSCO as the total number of visiting ships increases. For M3 and the perfect-foresight policy, although the inspection resources are fixed, more ships with larger number of deficiencies can be selected for inspection when the total number of visiting ships grows. Meanwhile, Table 3-12 also indicates that both the superiority of M3 over random scheduling case and the gap between M3 and the perfect-foresight policy show an increasing trend. This is because as the perfect-foresight policy can capture the ships with more deficiencies more efficiently than M3, the gap between them became larger as the number of visiting ships increases. This explanation can also be applied for the changes in the gap between M3 and random scheduling case.

SA2: different numbers of available PSCOs

Second, we analyze how the number of available PSCOs to carry out PSC inspection would influence the performance of M3. We set the number of ships for inspection to 30, and the number of PSCOs to 2, 3, 4, and 5 in SA2G1 to SA2G4, respectively. The performance of random scheduling case (based on 100 runs), M3, and the perfect-foresight policy and their comparison are presented in Table 3-13.

Table 3-13. Performance of the groups in SA2

Group	SA2G1	SA2G2	SA2G3	SA2G4
Number of PSCOs	2	3	4	5
Random scheduling case	37.4	53.1	63.9	70.5
M3	51.8	66.8	77.8	86.4
Perfect-foresight policy	61.6	76.6	86.1	92.9
Superiority of M3 over random scheduling case	38.5%	25.8%	21.8%	22.6%
Gap between M3 and the perfect-foresight policy	18.9%	14.7%	10.7%	7.5%

Table 3-13 indicates that when the number of ships that need to be inspected remains to be 30 while the number of PSCOs increases from 2 to 5, the total number of deficiencies that can be detected grows as expected. In addition, both the superiority of M3 over random scheduling case and the gap between M3 and the perfect-foresight policy show a decreasing trend. Particularly, such decreasing trend is more obvious in the gap between M3 and the perfect-foresight policy. This can be explained by the fact that as the number of available PSCOs increases, more ships can be assigned for inspection and thus to reduce the superiority of models with better performance as more ships with large number of deficiencies can be captured. Especially, for M3 which is based on the prediction given by XGBoost, more ships with larger real deficiency number can be captured for inspection although the XGBoost model is not perfect. As a consequence, the gap between M3 and the perfect-foresight policy gets closer more quickly than the superiority of M3 over random scheduling case as the number of inspected ships grows.

SA3: different berthing durations of ships

Third, we analyze model performance when the berthing duration of ships varies. We assume that the number of ships for inspection is 30 and the number of available PSCOs is 3. As only when a ship berths at a port for no less than two hours can the ship be inspected, we consider eight groups where the berthing duration of all ships is 2, 3, …., 8, 9 hours respectively denoted by SA3G1 to SA3G8. The consecutive berthing time units are randomly generated for all ships in each group. The performance of random scheduling case (based on 100 runs), M3, and the perfect-foresight policy and their comparison are presented in Table 3-14.

Table 3-14. Performance of the groups in SA3

Group	SA3G1	SA3G2	SA3G3	SA3G4	SA3G5	SA3G6	SA3G7	SA3G8
Berthing duration of all ships	2 hours	3 hours	4 hours	5 hours	6 hours	7 hours	8 hours	9 hours
Random scheduling case	38.3	42.8	48.4	50.7	52.2	45.9	48.8	51.3
M3	58.9	62.6	66.4	72.4	74.4	72.0	74.2	75.0
Perfect-foresight policy	71.8	82.9	89.3	94.5	96.0	90.7	95.1	95.0
Superiority of M3 over random scheduling case	53.6%	46.1%	37.2%	42.9%	42.4%	56.8%	52.0%	46.1%
Gap between M3 and the perfect-foresight policy	21.9%	32.4%	34.5%	30.5%	29.0%	26.0%	28.2%	26.7%

Table 3-14 shows that as the berthing duration of all ships increases, the total number of deficiencies detected also shows an upward trend although there are fluctuations due to the randomness in ship conditions. Meanwhile, there is no obvious pattern in the change of the gap between random scheduling case and M3 and the gap between M3 and the perfect-foresight policy when ship berthing duration increases. The superiority of M3 over random scheduling case is maximized at 56.8% when the berthing duration of all ships is 7 hours. The gap between M3 and the perfect-foresight policy is maximized at 34.5% when the berthing duration of all ships is 4 hours.

SA4: different berthing periods of ships

Fourth, we analyze how ship berthing period (during the work time of PSCOs) can influence the model performance. We set the number of ships for inspection to be 30 and the number of available PSCOs to be 3. We consider four groups of berthing periods as denoted by SA4G1 to SA4G4, respectively. In SA4G1, the berthing period of all ships is only in the morning (from 8:00 to 12:30). In SA4G2, the berthing period of all ships is only in the afternoon (from 12:30 to 17:00). In SA4G3, the berthing period of one-third of the ships is in the morning, in the afternoon, and both in the morning and in the afternoon, respectively. In SA4G4, the berthing period of half of the ships is in the morning and the other half is in the afternoon. The berthing duration is randomly generated for all ships. The performance and comparison of random scheduling case (based on 100 runs), M3, and the perfect-foresight policy are presented in Table 3-15.

Table 3-15. Performance of the groups in SA4

Group	SA4G1	SA4G2	SA4G3	SA4G4
Distribution of berthing period	All ships in the morning	All ships in the afternoon	1/3 ships in the morning, 1/3 ships in the afternoon, and 1/3 ships in the morning and afternoon	1/2 ships in the morning and 1/2 ships in the afternoon
Random scheduling case	23.1	22.2	47.1	50.5
M3	41.4	40.8	69.5	66.2
Perfect-foresight policy	61.1	59.1	88.2	87.1
Superiority of M3 over random scheduling case	79.4%	83.5%	47.7%	31.0%
Gap between M3 and the perfect-foresight policy	47.6%	44.9%	26.9%	31.6%

Table 3-15 shows that the number of deficiencies identified is smaller when there is more overlap in ship berthing period (i.e. in SA4G1 and SA4G2) than less overlap (i.e. in SA4G3 and SA4G4). Meanwhile, the average gaps between random scheduling case and M3 as well as between M3 and the perfect-foresight policy in SA4G1 and SA4G2 are much larger than those in SA4G3 and SA4G4. This is also because more ships can be inspected when their berthing period is more scattered, which would reduce the superiority of prediction models with better performance.

3.7 CONCLUSION

PSC inspection is a safeguard of maritime safety, the marine environment, and the rights of seafarers. To improve ship selection efficiency, this chapter first proposes an accurate XGBoost model to predict ship deficiency number. Particularly, domain knowledge regarding ship flag, RO, and company performance is considered in the XGBoost model, which improves its accuracy and fairness. Based on the predictions, an initial PSCO scheduling model is proposed to assign the PSCOs to inspect the predicted high-risk ships which also considers the number of available PSCOs and their work and rest time. To reduce problem size and improve model computation efficiency and flexibility, concepts of *inspection template* and *un-dominated inspection template* are further proposed and incorporated in the PSCO scheduling models.

In numerical experiments, we use the real PSC inspection records at the Hong Kong port from January 2016 to December 2018 as the case dataset to construct and validate the proposed models. Numerical experiments show that the MSE and MAE of the XGBoost model is 12.5 and 2.4 in the test set, respectively, which are better than the other popular ML models compared in this study. Moreover, when ship flag performance gets worse from white to grey and from grey to black, 0.8 and 0.2 more deficiency will be detected on average, respectively. When RO performance gets

worse from high to medium, 0.3 more deficiency will be detected on average. When company performance gets worse from high to medium, from medium to low, and from low to very low, 0.5, 0.8, and 1.5 more deficiencies will be detected on average, respectively. When combining the predictions with PSCO scheduling models, it is shown that the superiority of the proposed PSCO scheduling models over the current inspection scheme regarding the number of deficiencies identified is more than 20%. The gap between the proposed model and the model under perfect-forecast policy is about 8% regarding the number of deficiencies identified. Meanwhile, computation efficiency and flexibility of the PSCO scheduling model with *inspection templates* are higher than the initial PSCO scheduling model. Problem size can be reduced and the computation efficiency can be further improved in the PSCO scheduling model which takes *un-dominated inspection templates* and the relationship between each of the two ships into consideration. Extensive sensitivity analysis shows that when changing the numbers of ships for inspection, the numbers of available PSCOs, the berthing durations of ships, and the berthing periods of ships, the performance of the proposed PSCO scheduling model is stable and it is always better than the current model used at ports.

This chapter addresses an important practical problem in maritime industry. Theoretically, it proposes the first ship risk prediction model for PSC inspection considering domain knowledge. It also develops the first PSCO scheduling models based on the predictions to efficiently allocate scarce inspection resources for ship inspection. Moreover, the concepts of *inspection template* and *un-dominated inspection template* are proposed and incorporated in the PSCO scheduling model to improve computation efficiency and model flexibility. Practically, it helps port states to identify high-risk ships and assign the PSCOs more efficiently. Therefore, the main objectives of PSC to eliminate substandard shipping and safeguard the sea can be enhanced.

Chapter 4: A Semi-"smart predict then optimize" (semi-SPO) Method for Efficient Ship Inspection[4]

This chapter improves ship inspection efficiency by proposing three two-step approaches that match inspection resources with ship conditions so as to identify the most deficiencies (non-compliances with regulations) of the ships. It contains three combined prediction and optimization approaches. The first approach predicts the number of deficiencies in each deficiency category for each ship and then develops an integer optimization model that assigns the inspectors to the ships to be inspected. The second approach predicts the number of deficiencies each inspector can identify for each ship and then applies an integer optimization model to assign the inspectors to the ships to be inspected. The third approach is a semi-"smart predict then optimize" (semi-SPO) method. It also predicts the number of deficiencies each inspector can identify for each ship and uses the same integer optimization model as the second approach, however, instead of minimizing the MSE as in the second approach, it adopts a loss function motivated by the structure of the optimization problem in the second approach. Numerical experiments show that the proposed approaches improve the current inspection efficiency by over 4% regarding the total number of detected deficiencies. Through comprehensive sensitivity analysis, several managerial insights are generated, and the robustness of the proposed approaches is validated.

4.1 INTRODUCTION

17 deficiency codes are required by Tokyo MoU as presented in Table 4-1. The deficiency items in accordance with the deficiency codes are the inspection targets during a PSC inspection. Except for D99, the remaining 16 deficiency codes can be grouped into four deficiency categories as follows: C1: ship safety (D4 Emergency system, D7 Fire safety, D11 Life saving appliances, and D12 Dangerous goods), C2: ship management (D1 Certificates and documentation, D9 Working and living conditions, D14 Pollution prevention, D15 International Safety Management (ISM),

[4] Yan, R., Wang, S., Fagerholt, K., 2020. A semi-"smart predict then optimize"(semi-SPO) method for efficient ship inspection. Transportation Research Part B: Methodological 142, 100-125.

and D18 Labour conditions), C3: ship condition and structure (D2 Structural condition, D3 Water/Weathertight condition, D6 Cargo operations including equipment, and D13 Propulsion and auxiliary machinery), and C4: communication and navigation (D5 Radio communication, D8 Alarms, and D10 Safety of navigation). It should be noted that the deficiencies and deficiency categories are of equal importance as they are all derived from major international maritime regulations and conventions.

Table 4-1. Description of deficiency codes

Code	Meaning	Code	Meaning	Code	Meaning
D1	Certificates and documentation	D7	Fire safety	D13	Propulsion and auxiliary machinery
D2	Structural condition	D8	Alarms	D14	Pollution prevention
D3	Water/Weathertight condition	D9	Working and living conditions	D15	International Safety Management (ISM)
D4	Emergency system	D10	Safety of navigation	D18	Labour conditions
D5	Radio communication	D11	Life saving appliances	D99	Other
D6	Cargo operations including equipment	D12	Dangerous goods		

The overall inspection process suggests that the PSCOs play the key role in PSC inspection as they are responsible for conducting the inspection and deciding the inspection results (Ravira and Piniella, 2016; Graziano et al., 2017, 2018a). A PSCO should be an experienced person with both theoretical knowledge and seagoing experience. Common backgrounds of PSCOs can be naval architects, merchant marine captains, chief engineers, and ratio officers (Ravira and Piniella, 2016). As required, during an inspection, a PSCO will use his/her professional judgment to decide whether and in what aspects the ship should be further inspected. The PSCO will also use his/her expertise to decide what deficiencies should be recorded and whether to detain a ship. However, it is indicated that due to discretion, subjectivity, individuality, professional judgement, different backgrounds and expertise, PSCOs at the same port may have different expertise in identifying different categories of deficiencies (Ravira and Piniella, 2016; Graziano et al., 2017; Graziano et al., 2018a). For instance, there are two PSCOs on duty for one day, and PSCO 1 used to be a captain who is good at dealing with deficiencies related to communication and navigation, while PSCO 2 has naval background and is good at addressing deficiencies on the ship condition and structure. Assume now that two ships visiting the port are selected to be inspected: ship 1 has main deficiencies in structure and ship 2 has many deficiencies in radio communication. Ideally, we should assign PSCO 1 to inspect ship 2 and assign PSCO

2 to inspect ship 1; otherwise, deficiencies might be missing due to the lack of professional backgrounds and knowledge. This example shows that the inspection efficiency and effectiveness can be improved if ship deficiency conditions and the expertise of PSCOs are matched. To achieve this objective, the deficiency conditions of the ships, which can be represented by the number of deficiencies in each category (the number can be zero) need to be first predicted. Then, the expertise of PSCOs should be considered when assigning them to the ships to be inspected.

Considering the PSCOs' different expertise, this chapter proposes three approaches for ship deficiency condition prediction and PSCO assignment to improve the inspection efficiency. Our key contributions from a theoretical and practical point of view are summarized as follows.

First, from a theoretical point of view, we develop three sequential prediction and optimization approaches for the PSCO assignment problem. The first approach predicts the number of deficiencies in each deficiency category for each ship in a way that minimizes the MSE. The numbers of deficiencies are a natural choice of target to predict. The predicted values are subsequently used in a PSCO assignment model (model M1 in Section 4.4.1). The second approach predicts, instead of the number of deficiencies in each category for each ship, the number of deficiencies each PSCO can identify for each ship (also in a way that minimizes the MSE). The predicted values are subsequently used in a slightly different PSCO assignment formulation (model M2 in Section 4.4.2). The prediction models in the first two approaches do not account for how the predictions will be used in the optimization models, and this may lead to sub-optimal decisions (Elmachtoub and Grigas, 2017). Methods that fully integrate prediction and optimization, called "smart predict then optimize" (SPO) by Elmachtoub and Grigas (2017), are often computationally challenging. Instead of ignoring optimization models in the prediction or fully integrating optimization models into the prediction, semi-SPO methods partially integrates optimization models into the prediction, improving the performance while incurring limited extra computational burden (Demirović et al., 2019). The third approach proposed in our study is a semi-SPO method. It also predicts the number of deficiencies each PSCO can identify for each ship. However, instead of minimizing the MSE as in the second approach, this approach adopts a loss function motivated by the structure of the optimization problem. It aims to minimize the mean squared difference regarding the

overestimates (i.e., predicted value minus actual value) in the numbers of deficiencies that can be detected among the PSCOs for each ship (denoted by MSO for short). The prediction results are then applied to a PSCO assignment formulation (model M2 in Section 4.4.2). We demonstrate, on the basis of the three approaches, that (i) there may be different choices of targets to predict in the prediction model and then feed the targets into an optimization model and (ii) the structure of the optimization model may provide useful information to guide the training of the prediction model, even if the overall prediction and optimization procedure is sequential. Therefore, prediction models that show worse performance regarding classical regression metric (e.g., MSE) would not necessarily generate worse decisions in the following optimization models. Besides, we have rigorously proved that the optimization models can be solved in polynomial time of the length of its input parameters.

Second, from a practical point of view, we address a meaningful problem in maritime transportation. Improving inspection efficiency and effectiveness is a critical measure for PSC MoUs to guarantee maritime safety and protect the marine environment. One key point is realizing accurate identification of the deficiencies of the coming ships, which benefits from accurate prediction. Based on the three prediction models and the optimization model proposed in this study, the expertise of PSCOs can be fully utilized in dealing with various deficiency conditions of the ships. Particularly, compared with random assignment of PSCOs, the proposed three models can help to detect 4.70%, 4.55%, and 4.86% more deficiencies, respectively, after inspecting the same groups of ships by using the same PSCO resources. Comprehensive robustness analysis shows that even if there may be some uncertainties in measuring the expertise of PSCOs, the PSCO assignment scheme generated by the third proposed model can still identify more than 90% of the real deficiencies and significantly outperforms random PSCO assignment. From the perspective of application, as reported by Tokyo MoU, there were totally 31,589 PSC inspections and the total number of deficiencies detected was 73,441 in 2017 (Tokyo MoU, 2018a). This indicates that the average number of deficiencies of one ship in one PSC inspection is about 2.32. If our models are applied, about 3,569 more deficiencies can be detected (as 4.86% more deficiencies can be identified compared with random PSCO assignment), which can be viewed as inspecting about 1,538 more ships with

the same inspection resources. Therefore, human, material and financial resources could be saved if the inspection efficiency is improved.

4.2 RESEARCH GAP

Although the effectiveness of PSC inspections in improving the safety level of maritime transport has been widely recognized by industry and academia, there are still critical challenges faced by port state authorities. One of the biggest challenges is the discrepancy in the inspection process and criteria among different PSC MoUs, port states of the same MoU, and even PSCOs at the same port. More specifically, variations in the treatment of vessels across the MoUs were identified and reported by Sampson and Bloor (2007), Knapp and Franses (2007), and Knapp and van de Velden (2009), and the differences in inspections within the same MoU were found by Bateman (2012), Graziano et al. (2018b), and Şanlıer (2020), while the different treatment caused by different backgrounds and expertise of the PSCOs was investigated by Ravira and Piniella (2016) and Graziano et al. (2017, 2018a). It is of vital importance to achieve harmonization in PSC inspections, or the ship operators will recognize that they no longer necessarily gain a great deal from efforts to comply with regulations and thus substandard ships will "port shop", i.e., choose to call the ports with looser PSC inspection criteria.

The models proposed in this paper could help to address the problems brought about by the diverse backgrounds and expertise of PSCOs at the same port by matching the ship deficiency conditions with PSCOs' expertise. Besides, the phenomenon of "port shop" can also be alleviated by improving inspection efficiency.

Based on the literature review in this chapter and in chapter 2, it can be seen that although there are a large number of studies on improving PSC inspection efficiency, to the best of our knowledge, there is no literature on developing PSCO assignment schemes to improve inspection efficiency by considering the expertise and backgrounds of PSCOs and the deficiency conditions of the ships.

4.3 DATA DESCRIPTION AND THE PSCO ASSIGNMENT PROBLEM

The Asia Pacific Computerized Information System (APCIS) provided by the Tokyo MoU and World Register of Ships (WRS) database are used in this study. APCIS is a public website-based database of PSC inspections conducted by the

member authorities of the Tokyo MoU. It contains ship generic information and historical PSC inspection records within the Tokyo MoU (including the specific deficiencies detected for each inspected ship). WRS is a comprehensive database providing hundreds of features on ship construction, engine, dimension, registration, ownership, fixtures, and class, etc. We select the most relevant features of PSC inspection from WRS based on the literature. The features selected from APCIS and WRS are combined by ship IMO number, and there are 15 input features in total. The description of the features and their statistical information used in this study are provided in Table 4-2. For ships that have never had any inspection within Tokyo MoU, the values for "last inspection time", "last deficiency number" and "follow-up inspection rate" are set to be "none" (not included in Table 4-2).

Table 4-2. Description of input features

Feature name	Meaning	Min value	Max value	Average value
Age (year)	Difference between keel laid date and inspection date.	0	47	11.00
GT (100 cubic feet)	A measure of a ship's overall internal volume.	299.00	1,995,636.00	44,927.76
Length (meter)	The overall maximum length of a ship.	32.29	400.00	212.73
Depth (meter)	The vertical distance measured from the top of the keel to the underside of the upper deck at side.	3.70	36.02	17.60
Beam (meter)	The width of the hull.	7.38	60.05	31.64
Type	Bulk carrier (12.70%), container ship (57.05%), general cargo/multipurpose (10.95%), passenger ship (1.35%), tanker (11.50%), other (6.45%).	/	/	/
Number-of-times-of-changing-flag	The sum of the times the ship's flag has been changed after keel laid date.	0	8	0.69
Total-detention-times	The sum of the times the ship has been detained by all PSC authorities.	0	18	0.62
Casualties-in-last-five-years	1, if the ship is encountered with casualties in last five years; 0, otherwise.	0	1	0.09
Ship-flag-performance*	White (92.10%), grey (3.20%), black (4.05%), not listed (0.65%).	/	/	/
Ship-RO-performance*	High (95.85%), medium (2.30%), low (0.10%), very low (0), not listed (1.75%).	/	/	/
Ship-company-performance*	High (34.50%), medium (40.25%), low (15.70%), very low (9.10%), not listed (0.45%).	/	/	/
Last-inspection-time (month)	The time of last PSC inspection within Tokyo MoU.	0.03	180.70	10.12
Last-deficiency-number	The deficiency number of last PSC inspection within Tokyo MoU.	0	55	3.41
Follow-up-inspection-rate	The total number of follow-up inspections divided by total number of inspections within Tokyo MoU.	0.00	1.00	0.15

* Note: Ship flag performance, RO performance, and company performance are calculated based on flag Black-Grey-White list, RO performance list, and company performance list provided by Tokyo MoU, respectively. The performance of the flags on white-list is better than those on grey-list, and much better than those on black-list. For RO and company, the performance gets worse in the sequence of "high", "medium", "low", and "very low". If the performance of the ROs and companies is not shown on the lists, the performance state is recorded as "not listed".

We use a total of 2,000 inspection records at the Hong Kong port in 2016 (638 records), 2017 (641 records) and 2018 (721 records) in our study. We use the PSC inspection records at the Hong Kong port because we have visited the Marine Department of Hong Kong Special Administrative Region (HKSAR) and discussed with the PSCOs here for several times. We learned that the PSC Section of Hong Kong Marine Department has four PSCOs who are all experienced experts in all aspects of PSC. Besides, it is required that a PSCO should participate in strict trainings and assessments before becoming a qualified PSCO according to the requirements of the Hong Kong Marine Department, and the PSCOs also need to attend regular training programs and seminars. Therefore, we suppose that the PSCOs at the Hong Kong port can identify all the deficiencies in each category for each inspected ship. Nevertheless, it should be noted the PSCOs at some ports may not be that experienced, and thus the Tokyo MoU has developed several co-operation programs to enhance consulting, cooperating and exchanging information among the authorities (Tokyo MoU, 2018a). The models proposed in our study aiming to match the ship conditions with the PSCOs' expertise can also be viewed as a type of cooperation and thus are more suitable for those ports with PSCOs of divergent expertise. We randomize the whole dataset and divide it into training set, validation set and test set with each containing 70%, 15% and 15% of all data entries, i.e.,1400, 300 and 300 data entries, respectively.

According to the working process of the PSC authorities, in the morning of each day, a set of ships (denoted by S) to be inspected will be selected among all the ships coming to the port state on that day. A total of P PSCOs will then be assigned for ship inspection. It is not uncommon that some PSCOs have limited expertise in some aspects of PSC because of limited work experience and training. It is, therefore, valuable to leverage historical inspection data and predict the number of deficiencies in each category for each ship, and based on the predicted number, to assign PSCOs with the relevant expertise to inspect the ships. Let $C = 4$ be the number of categories of deficiencies (i.e., ship safety, ship management, ship condition and structure, and communication and navigation mentioned in Section 4.1). The expertise of PSCO p for inspecting deficiency category c is denoted by u_{pc}, $p = 1,...,P$, $c = 1,2,3,C$. u_{pc} is actually the percentage of deficiencies of category c that can be detected by PSCO p, and $0 \le u_{pc} \le 1$. The smaller u_{pc} is, the more deficiencies in c are likely to be ignored by PSCO p. The expertise (which is represented by percentage) can be

evaluated by tests, questionnaires, and interviews. Considering the workload for the PSCOs, we further require that the maximum number of ships that a PSCO can inspect for each day is Θ. We try to assign the available PSCOs to the selected ships in a way that maximizes the total number of deficiencies in all the C categories of all the ships that can identified.

The prediction and optimization models proposed in this study work in the following way: deficiency prediction models with three different targets/model structures are first developed. Based on the prediction results, optimization models for PSCO assignment to maximize the inspection efficiency are then proposed. Several comparisons are made and comprehensive sensitivity analyses is conducted to generate managerial insights and validate the robustness of the models.

4.4 PREDICTION AND OPTIMIZATION APPROACHES

In our prediction and optimization approaches, a prediction model is first developed to predict the key unknown parameters in the optimization model. Based on the predicted values, an optimization model is then constructed to generate decisions. The main difference between the prediction models proposed in this chapter from that proposed in Chapter 3 is that the output is of multi-dimension in this chapter, while that in Chapter 3 is of one-dimension. The reason for choosing random forest-based prediction model instead of XGBoost is that CART-based decision tree and random forest are widely used in developing SPO frameworks in existing literature as they are intuitive and easy to be modified. To split each node, all features and values are enumerated to find the best split that can reduce the MSE the most in a greedy manner, and the output of a leaf node is just the mean value of all the samples contained in that node. Therefore, it is not a difficult task to change the loss function in node splitting to satisfy tailored needs. In contrast, although the XGBoost is generally believed to be more accurate than random forest, its working mechanism is much more complex, especially the derivative calculation and regularization are required in the loss function. Therefore, it might be very hard to modify its loss function while considering the properties and structure of the following optimization model. Therefore, we choose random forest consisting of CART regression tree in the SPO framework. Particularly, we propose three prediction models denoted by MTR-RF1, MTR-RF2, and MTR-RF3 and two assignment models denoted by M1 and M2 with details provided in Table 4-3.

Table 4-3. Prediction and optimization models

Model	Prediction targets	Splitting criteria	Decision trees	Assignment model	Assignment decision
MTR-RF1	Number of deficiencies under each deficiency category	MSE	$f^{MTR}(\mathbf{x})$	M1	A1
MTR-RF2	Number of deficiencies identified by each PSCO	MSE	$f'^{MTR}(\mathbf{x})$	M2	A2
MTR-RF3	Number of deficiencies identified by each PSCO	MSO	$f''^{MTR}(\mathbf{x})$	M2	A3

4.4.1 Prediction of natural targets and optimization

It is natural to predict the number of deficiencies in each category for each ship based on historical records. Therefore, we first develop random forest regression model (denoted by MTR-RF1) to predict the number of deficiencies in each category for each ship based on the features in Table 4-2.

4.4.1.1 Prediction model

We use random forest (RF) as the prediction model. RF is a state-of-the-art ML model with high accuracy and is widely used (Friedman et al., 2001; Liaw and Wiener, 2002; Breiman, 2017). We first present the construction procedure of a decision tree, and then the RF.

Decision tree (denoted by DT for short) is a popular supervised ML model. At the beginning, all the training examples are stored in the root node. Then, the root node is recursively split into successive nodes which contains subsets of the training set until coming to the preset stopping criterion or the current node cannot be further split (i.e., all the examples are of the same output value). Each split of the nodes in the decision tree aims to reduce the variance among the records in the successive nodes. According to the target, decision trees that predict categorical target are called classification trees while decision trees that predict numerical target are called regression trees. The target is one-dimensional in traditional decision tree while the targets can be multi-dimensional in multi-target regression (MTR) tree (Blockeel and De Raedt, 1998). In this study, the outputs are four-dimensional (either the number of deficiencies under the four categories or the number of deficiencies detected by the four PSCOs), and thus the MTR trees are constructed by using CART algorithm (Friedman, 2001; Harrington, 2012; Breiman, 2017). The procedure is as follows (Blockeel, 1998; Friedman et al., 2001).

The input information for decision tree construction contains the training dataset and termination conditions. We denote the set of J input features as $(x_1, x_2, ..., x_J)$ and the set of K targets as $(y_1, y_2, ..., y_K)$. An input feature is denoted by x_j, $j = 1, ..., J$, and the value set of this feature is denoted by Ω_j. A specific value of this feature is denoted by w_j, $w_j \in \Omega_j$. For example, for the variable ship-flag-performance which has four states: white, grey, black, and not listed, the states are first changed to numbers, with 1 representing white, 2 representing grey, 3 representing black, and 4 representing not listed. Then, we can have $\Omega_j = \{1, 2, 3, 4\}$. A target is denoted by y_k, $k = 1, ..., K$ and $K = 4$. In addition, we denote the training dataset containing N data entries as $D = \{(\mathbf{x}^1, \mathbf{y}^1), (\mathbf{x}^2, \mathbf{y}^2), ..., (\mathbf{x}^N, \mathbf{y}^N)\}$. We use $e = 1, ..., N$ to refer to both an inspection record and the ship in the current record. Notably, if a ship is inspected several times, its inspection records are treated independently. A data entry is denoted by $(\mathbf{x}^e, \mathbf{y}^e)$ with $e = 1, ..., N$, where $\mathbf{x}^e = (x^{e1}, x^{e2}, ..., x^{ej}, ..., x^{eJ})$ contains J features and $\mathbf{y}^e = (y^{e1}, y^{e2}, ..., y^{ek}, ..., y^{eK})$ contains K targets. The construction process of a CART-based MTR tree requires finding the *best split* pair $(j^*, w_{j^*}^*)$, $w_{j^*}^* \in \Omega_{j^*}$ of the nodes that minimizes the total within-subset variation in the two successive nodes when splitting. Denote the set of I termination conditions as $\Gamma = (\gamma_1, \gamma_2, \gamma_3, ..., \gamma_I)$. The main steps to construct an MTR tree are presented as shown in Appendix B.1.

In our problem, we choose the 15 features in Table 4-2 as $\mathbf{x}$ ($J = 15$) and the number of deficiencies in each category as $\mathbf{y}$ ($K = 4$). For each ship in record $e = 1, ..., N$, the input features are $\mathbf{x}^e = (x^{e1}, x^{e2}, ..., x^{ej}, ..., x^{eJ})$. Because we have several ML models, in this model we represent the targets by $\boldsymbol{\alpha}^e = (\alpha^{e1}, \alpha^{e2}, \alpha^{e3}, \alpha^{eC})$ instead of using $\mathbf{y}$, where α^{ec} is the number of deficiencies in category c, $c = 1, ..., C$ (we use C to represent the number of deficiency categories instead of using K) of ship e, and then

$$(j^*, w_{j^*}^*) \in \underset{\substack{j \in (x_1, ..., x_J) \\ w_j \in \Omega_j}}{\arg\min} \left[\sum_{e \in R_1(j, w_j)} \sum_{c=1}^{C} \left(\alpha^{ec} - \frac{1}{|R_1(j, w_j)|} \sum_{e_1 \in R_1(j, w_j)} \alpha^{e_1 c} \right)^2 + \sum_{e \in R_2(j, w_j)} \sum_{c=1}^{C} \left(\alpha^{ec} - \frac{1}{|R_2(j, w_j)|} \sum_{e_2 \in R_2(j, w_j)} \alpha^{e_2 c} \right)^2 \right] \tag{4.1}$$

where $R_1(j, w_j) = \{e \in R_0 \mid x^{ej} \le w_j\}$ and $R_2(j, w_j) = \{e \in R_0 \mid x^{ej} > w_j\}$.

Like traditional DTs, the MTR trees can also be ensembled by using bagging (Breiman, 1996) and bootstrapping (Breiman, 2001) to reduce overfitting and increase prediction accuracy. In this study, we adopt random forest (which is based on bagging)

to ensemble MTR trees proposed in Section 4.4.1.1. Compared to a single decision tree, the decision trees contained in the RF have two layers of randomness: a new training set generated by bootstrapping (i.e. randomly selecting a certain number of samples from the whole dataset with replacement) in the original training set is used to construct each decision tree, and a subset (with a preset fixed size) of all features is used to split each node in each decision tree (Friedman et al., 2001). A detailed construction procedure of MTR tree based random forest (MTR-RF) model is provided in Appendix B.2 (Breiman, 2001; Kocev et al., 2007).

4.4.1.2 Optimization model

Among all the foreign ships visiting the port, the ships to be inspected are selected based on guidelines provided by the Tokyo MoU (2018). For each ship $s \in S$ selected to be inspected, we can only obtain its input features, while the number of deficiencies under deficiency category c is unknown. With a little abuse of notation, we denote the unknown number of deficiencies in category c of ship s by α^{sc}, $s \in S$, $c = 1,2,3,C$ (α^{ec} is the known number of deficiencies in category c of ship e in the training set, $e = 1,...,N$). The predicted values of α^{sc}, denoted by $\hat{\alpha}^{sc}$, $s \in S$, $c = 1,2,3,C$, can be obtained by using the RF model proposed in Section 4.4.1.1. To achieve the maximum inspection efficiency, the sum of the product of the estimated deficiency number of each deficiency category and the corresponding inspection expertise of that deficiency category of the assigned PSCO (denoted by "inspection expertise" for short) should be as large as possible. The justification for matching the deficiency categories with the expertise of PSCOs is as follows. The decision (outcome) of a PSC inspection contains ship deficiency (specific deficiency types and total deficiency number) and ship detention. During a PSC inspection, the PSCO gets onboard and inspect the condition of the ship. For any condition that is not in compliance with the related regulations and conventions, it will be recorded as a deficiency. On the contrary, ship detention is not directly observed; instead, it is determined by the detected deficiencies and the PSCOs' judgement. Therefore, if the deficiency condition of the ships can be matched with the expertise of the PSCOs, the most proper PSCO (who can identify the existing deficiencies as many as possible and make rational detention decision) can be assigned to inspect the ship for better ship deficiency identification and detention decision making. Following this idea, we define binary decision variable z_{ps} that is set to 1 if PSCO p is assigned to inspect

ship s and 0, otherwise, and the PSCO assignment model can be expressed by mathematical model M1.

[M1]

$$\max \sum_{p=1}^{P} \sum_{s \in S} \sum_{c=1}^{C} \hat{\alpha}^{sc} u_{pc} z_{ps} \tag{4.2}$$

subject to

$$\sum_{s \in S} z_{ps} \leq \Theta, \ p = 1, ..., P \tag{4.3}$$

$$\sum_{p=1}^{P} z_{ps} = 1, \ s \in S \tag{4.4}$$

$$z_{ps} \in \{0,1\}, \ p = 1, ...P, \ s \in S. \tag{4.5}$$

Objective (4.2) maximizes the inspection expertise of the PSCOs by maximizing the sum of the product of the estimated deficiency number under each deficiency category and the expertise of the selected PSCO for that corresponding deficiency category for all inspected ships. Constraints (4.3) limit the maximum number of ships that can be inspected by a PSCO for one day. Constraints (4.4) guarantee that each ship is inspected by one PSCO. Constraints (4.5) ensure the domain of the decision variable.

Although model M1 is an integer program, it has the following nice property, whose proof is in Appendix B.3.

Proposition 1: Model [M1] can be solved in polynomial time of the length of the input parameters.

Proposition 1 implies that the PSCO assignment model [M1] is an easy problem: even if there are hundreds of ships and tens of PSCOs, [M1] can be solved efficiently (e.g., in less than 1 second).

4.4.2 Prediction of coefficients in the objective function of optimization model

4.4.2.1 Prediction model

In model M1, the coefficients of the decision variables in the objective function are $\sum_{c=1}^{C} \hat{\alpha}^{sc} u_{pc}$, $s \in S$ and $p = 1, ..., P$. Therefore, instead of predicting α^{sc} (i.e. the number of deficiencies in category c for ship s), we can directly predict $\sum_{c=1}^{C} \alpha^{sc} u_{pc}$ (i.e. the total number of deficiencies of ship s that can be detected by PSCO p).

Define $\beta^{sp} = \sum_{c=1}^{C} \alpha^{sc} u_{pc}$, $s \in S$ and $p = 1,...,P$. For ship s , $\boldsymbol{\beta}^s = (\beta^{s1},...,\beta^{sp},...,\beta^{sP})$ denotes the number of deficiencies that can be detected by assigning PSCO $p = 1,...,P$. The values for β^{sp} (and thus $\boldsymbol{\beta}^s$) can be predicted by using the RF models developed in Section 4.1.1.2, and the prediction model is denoted by MTR-RF2. The predicted values generated by MTR-RF2 are denoted by $\hat{\beta}^{sp}$. The procedure of constructing the MTR trees $f'^{MTR}(\mathbf{x})$ in MTR-RF2 is slightly different from the $f^{MTR}(\mathbf{x})$ in MTR-RF1: a data entry $(\mathbf{x}^e, \boldsymbol{\beta}^e)$ represents ship s , where the input features are $\mathbf{x}^e = (x^{e1}, x^{e2},..., x^{ej},..., x^{eJ})$, $J = 15$, and the targets are $\boldsymbol{\beta}^e = (\beta^{e1},...,\beta^{ep},...,\beta^{eP})$. Both MTR-RF1 and MTR-RF2 generate multi-dimensional targets: MTR-RF1 has C targets for the deficiency numbers under C deficiency categories while MTR-RF2 has P targets for the deficiency numbers identified by the P PSCOs. In particular, the choice of the best split in Step 1 in Procedure 1 for constructing an MTR tree should be revised as

$$(j^*, w_{j*}^*) \in \underset{\substack{j \in (x_1,...,x_J) \\ w_j \in \Omega_j}}{\arg\min} \left[\sum_{e \in R_1(j,w_j)} \sum_{p=1}^{P} (\beta^{ep} - \frac{1}{|R_1(j,w_j)|} \sum_{e_1 \in R_1(j,w_j)} \beta^{e_1 p})^2 + \sum_{e \in R_2(j,w_j)} \sum_{p=1}^{P} (\beta^{ep} - \frac{1}{|R_2(j,w_j)|} \sum_{e_2 \in R_2(j,w_j)} \beta^{e_2 p})^2 \right] \quad (4.6)$$

where $R_1(j, w_j) = \{e \in R_0 \mid x^{ej} \le w_j\}$ and $R_2(j, w_j) = \{e \in R_0 \mid x^{ej} > w_j\}$.

4.4.2.2 Optimization model

Based on the predicted values $\hat{\beta}^{sp}$, optimization model M1 can be reformulated as

[M2]

$$\max \sum_{p=1}^{P} \sum_{s \in S} \hat{\beta}^{sp} z_{ps} \quad (4.7)$$

subject to constraints (4.3) to (4.5). The structure of [M2] is the same as that of [M1] and hence [M2] can also be solved as a linear program.

4.4.3 Prediction of fundamental parameters that are fed into the optimization model

It is common that to predict the values $\hat{\beta}^{sp}$ in [M2], we try to minimize the sum of squared errors between the predicted value and the actual value, as shown in Eq. (4.6). However, a closer examination into the structure of the optimization model [M2] reveals that if the predicted number of deficiencies the P PSCOs can identify for a

ship are overestimated or underestimated by the same value, the final optimal assignment decision will not be influenced. We use the following example to illustrate this finding:

Example: For any ship that is selected to be inspected, if the actual numbers of deficiencies four PSCOs can identify are 6, 7, 8, and 9, but the predicted numbers are 8, 9, 10, and 11 (i.e., all four outputs are overestimated by "2"), then the optimal assignment is not changed and we should assign PSCO 4 to inspect the ship. If the predicted number are 5, 6, 7, 8 (i.e., all four outputs are underestimated by "1"), then the optimal assignment is also not changed and we should assign PSCO 4 to inspect the ship.

Generally, if the actual numbers of deficiencies the P PSCOs can identify for a ship are n_1, n_2, …, n_P, but the predicted numbers are $n_1+\varepsilon$, $n_2+\varepsilon$,…, $n_P+\varepsilon$ ($\varepsilon \in R$; if $\varepsilon<0$, $|\varepsilon| \le \min(n_1, n_2, ..., n_P)$), then the resulting prediction errors do *not* adversely affect the PSCO assignment decision, because it is the *difference* in the predicted numbers among the PSCOs, rather than the absolute prediction values, that affects the assignment decision. Based on this observation, the third approach (denoted by MTR-RF3) minimizes the squared difference regarding the overestimates (i.e., predicted value minus actual value) in the predicted numbers of deficiencies among the PSCOs and then uses the prediction in a PSCO assignment formulation (model M2 in Section 4.4.2.2). The prediction model is revised as follows.

Decision trees contained in MTR-RF3 is denoted by $f''^{MTR}(\mathbf{x})$. Splitting criterion of $f''^{MTR}(\mathbf{x})$ is changed to minimize the sum of variance of the predicted deficiencies that can be detected by each PSCO for each ship. More specifically, in Procedure 1, the best split pair $(j^*, w_{j^*}^*)$ of the current splitting node is calculated by

$$(j^*, w_{j^*}^*) \in \underset{\substack{j\in(x_1,\ldots,x_J)\\ w_j\in\Omega_j}}{\arg\min} \left[\begin{array}{l} \sum_{e\in R_1(j,s_j)} \sum_{p=1}^{P-1} \sum_{p'=p+1}^{P} ((\beta^{ep} - \frac{1}{|R_1(j,w_j)|} \sum_{e_1\in R_1(j,w_j)} \beta^{e_1 p}) - (\beta^{ep'} - \frac{1}{|R_1(j,w_j)|} \sum_{e_1\in R_1(j,w_j)} \beta^{e_1 p'}))^2 + \\ \sum_{e\in R_2(j,s_j)} \sum_{p=1}^{P-1} \sum_{p'=p+1}^{P} ((\beta^{ep} - \frac{1}{|R_2(j,w_j)|} \sum_{e_2\in R_2(j,w_j)} \beta^{e_2 p}) - (\beta^{ep'} - \frac{1}{|R_2(j,w_j)|} \sum_{e_2\in R_2(j,w_j)} \beta^{e_2 p'}))^2 \end{array} \right]. \quad (4.8)$$

The predicted numbers of deficiencies that can be detected by each PSCO given by MTR-RF3 based on Eq. (4.8) are then input to optimization model M2 to generate PSCO assignment decisions.

4.5 COMPUTATIONAL EXPERIMENTS

4.5.1 Construction of MTR-RF

4.5.1.1 Introduction of hyperparameters in RF

A hyperparameter in ML is a parameter used to control the learning process and whose value is set before the learning process begins. As RF is an ensembled ML model which contains DTs as weak learners, an RF model has hyperparameters to control the overall structure and properties of the RF as well as those for its DTs. Hyperparameters for DTs are mainly used to control the complexity and serve as the regularization of the model. The hyperparameters for RF are summarized below.

(a) n_estimators: the total number of DTs contained in an RF model. As the main principle underlying bagging is that more trees are better while too few trees can lead to unstable performance, this hyperparameter should be set to the largest computationally manageable value and do not need to be tuned (Breiman, 2001; Probst and Boulesteix, 2017).

(b) max_features: the number of features considered for each split. The value range of this hyperparameter is from 1 to the total number of features in the dataset and it is an integer. Too small value will negatively affect the average performance of the trees, while too large value will reduce the randomness of each tree and thus badly influence the overall performance. Denote the total number of features as n_features, $\text{n\_features} = J = 15$. It is suggested setting $\text{max\_features} = \lfloor \text{n\_features} / 3 \rfloor$ for regression trees (Friedman et al., 2001; Probst et al., 2019).

(c) max_depth: the maximum depth of each DT in the RF model. The depth of a leaf is the number of splits taken from the root node to that leaf node (Elmachtoub et al., 2020). The value range of this hyperparameter can be set from one to unlimited and it is an integer. Larger value of max_depth leads to more complex single trees.

(d) min_samples_leaf: the minimum number of examples required to be at a leaf node. The minimum value for this hyperparameter is 1 and it is an integer. Smaller value of min_samples_leaf leads to more complex single trees. It is recommended to set the value of min_samples_leaf to be 5 for regression models by default (Friedman et al., 2001).

Hyperparameters (a) and (b) control the overall structure and the property of randomness of RF, while hyperparameters (c) and (d) are related to each DT. It should also be noted that in practice the best values for these parameters will depend on the problem, and should be treated as tuning parameter (Friedman et al., 2001).

4.5.1.2 Hyperparameter tuning in RF

Hyperparameters can have a large impact on model performance and generalization ability. Although it has been proved that RF models will not overfit, several studies have shown that tuning the hyperparameters in RF would yield slightly better performance and generalization ability (Biau and Scornet, 2016; Probst et al., 2019). In this study, we aim to tune three hyperparameters: max_features, max_depth, and min_samples_leaf which can only take integer values by using a training set and a validation set. We choose MSE as the performance evaluation measure for MTR-RF1 and MTR-RF2 and MSO in the predicted ship deficiency number that can be identified among the PSCOs as the performance evaluation measure for MTR-RF3. To tune the three hyperparameters, we propose a revised grid search method. Denote the pre-defined set containing all the possible values for a hyperparameter as its constrained value space. Unlike the classical grid search which exhaustively considers all hyperparameter combinations in the constrained value spaces to form the grid, the revised grid search method could gradually reduce the search space by iteration. The procedure to tune the hyperparameters by the revised grid search is presented in Appendix B.4. In this study, the default value for max_features should be 5 (recall that we have 15 input features) and min_samples_leaf should be 5. To form the constrained value space, we extend the value spaces of the two hyperparameters by increasing/decreasing the default value to the same extent, i.e. we set the constrained value space for max_features as $\{3,4,5,6,7\}$ and for min_samples_leaf as $\{2,3,4,5,6,7,8\}$. For the constrained value space of max_depth, as there is no recommended default value, we set it to be a moderate range as $\{4,5,6,7,8\}$.

4.5.2 Performance of the MTR-RF models and PSCO assignment schemes

4.5.2.1 Experiment settings and hyperparameters in MTR-RF

The settings in the numerical experiments are in accordance with the real situation at the Hong Kong port: there are 4 available PSCOs, and about 10 ships are selected for inspection every day with 2 to 3 ships assigned to one PSCO. We further

assume that PSCO 1 is good at dealing with deficiency category C1, PSCO 2 is good at dealing with deficiency category C2, PSCO 3 is good at dealing with deficiency categories C3 and C4, and PSCO 4 is good at dealing with deficiency categories C4. The assumed expertise of each PSCO to inspect each deficiency category is presented in Table 4-4. After applying the revised grid search method to the three hyperparameters under the given constrained value spaces in MTR-RF1, MTR-RF2, and MTR-RF3, the best hyperparameter tuples for the three models are shown in Table 4-5.

Table 4-4. Expertise of each PSCO in each deficiency category

PSCO/deficiency category	C1	C2	C3	C4
PSCO 1	0.8	0.5	0.7	0.6
PSCO 2	0.7	0.9	0.4	0.5
PSCO 3	0.7	0.6	0.8	0.7
PSCO 4	0.4	0.7	0.6	0.7

Table 4-5. Best hyperparameter tuples for MTR-RF1, MTR-RF2, and MTR-RF3

Model	max_features	max_depth	min_samples_leaf
MTR-RF1	4	8	5
MTR-RF2	4	7	3
MTR-RF3	6	8	4

After finding the optimal hyperparameter tuple for each model by using the training set and the validation set, in the following experiments we form a new training set by combining the current training and validation sets, and thus it contains 1,700 inspection records at the Hong Kong port. The test set contains another 300 inspection records at the Hong Kong port. We randomly and evenly divide them into 30 groups where each group contains 10 ships. We assume that the 10 ships in a group come to the port on one day and the totally 300 ships come to the port on 30 days. We also require that the maximum number of ships that can be inspected by one PSCO is three.

4.5.2.2 Performance of the three MTR-RF models

We set n_estimators = 200 for the proposed three MTR-RF models. Each MTR-RF model is trained by using the new training set and the hyperparameter tuple tuned by the revised grid search. Run each of the MTR-RF model 10 times, and the min, max, mean, and variance of MSE/MSO on the test set in the 10 runs for the three models are shown in Table 4-6. It can be seen that the min, mean, and max values of MSE of MTR-RF1 are all much smaller than those in MTR-RF2. The differences are caused by the values of the prediction targets in the MTR-RF models: in MTR-RF1, the prediction targets are the deficiency number under each deficiency category; while in MTR-RF2, the prediction targets are the *total* number of deficiencies a PSCO can detect if she/he is assigned to inspect the ship. Besides, it is shown that the min, mean, and max values of MSE of MTR-RF2 are all smaller than those of MTR-RF3, which indicates that MTR-RF2 performs better than MTR-RF3 as a regression model evaluated by MSE. The differences in MSE between MTR-RF2 and MTR-RF3 are caused by the property of the MTR-RF models: the splitting criteria in MTR-RF2 is to reduce the MSE in successive nodes while those in MTR-RF3 is to reduce MSO. In

addition, the variance in each model is small, which implies that the performance of MTR-RF containing 200 MTR trees is stable.

Table 4-6. Prediction performance of MTR-RF1, MTR-RF2, and MTR-RF3

Model	Metric	Min	Mean	Max	Variance
MTR-RF1	MSE	3.9756	4.0173	4.0762	0.0009
MTR-RF2	MSE	15.4953	15.8342	16.1237	0.0437
MTR-RF3	MSE	16.7775	17.1684	17.5571	0.0443
MTR-RF3	MSO	3.0242	3.0513	3.0863	0.0002

Table 4-6 shows that compared to the prediction outputs of MTR-RF3, the outputs of MTR-RF2 have less variability. Meanwhile, even if the differences in the prediction targets of MTR-RF1 and MTR-RF2 are considered, the variability of MTR-RF1 is less than MTR-RF2. The reasons are as follows. For the difference between the variance of MTR-RF2 and MTR-RF3, the splitting criterion of the DTs is to minimize the MSE of ship deficiency number detected by each PSCO in MTR-RF2, whereas the splitting criterion of the DTs in MTR-RF3 is to minimize the MSO of ship deficiency number detected by each PSCO. Therefore, the target of the prediction generated by MTR-RF2 is to make the outputs as close as to their real values, while the target of the prediction generated by MTR-RF3 is to make the differences of the overestimates of each two of the outputs as small as possible. As a result, MTR-RF3 generates more flexible prediction results and when evaluating the variance of the outputs, the variance of the outputs of MTR-RF3 is larger than MTR-RF2. For the difference between the variance of MTR-RF1 and MTR-RF2, recall that the prediction targets of MTR-RF1 only consider the deficiency number under each deficiency category while both the deficiency number and the PSCOs' expertise in each deficiency category are considered in the prediction targets of MTR-RF2. Due to the nonlinearity of MTR-RF models, the impacts of the PSCOs' expertise on the deficiency number in the outputs variability can be magnified. As a result, the uncertainties are propagated to the output predictions, which leads to higher variance in MTR-RF2 compared to MTR-RF1.

4.5.2.3 Performance of the combined prediction and optimization model

We assign PSCOs based on the prediction results (10 runs) in Section 4.5.2.1 to the 30 groups of ships in accordance with the settings. The assignment decisions generated by assignment models based on MTR-RF1, MTR-RF2, and MTR-RF3 are denoted by A1, A2, and A3, respectively. Apart from making comparisons among the three models themselves, we also compare them with random assignment scheme and

best assignment scheme in theory. The performance of random assignment scheme is the mean inspection expertise of 10,000 times of random PSCO assignment. The best assignment scheme in theory is making PSCO assignment decisions under the assumption that there is a perfect ML model that could predict the parameters for the optimization model totally accurate. However, this is an ideal situation that never exists because the generalization error cannot be zero. The comparison results are shown in Table 4-7. We further analyze the randomness of A1, A2, and A3 by calculating the min and max values of the inspection expertise and the variance of inspection expertise among the 30 groups of PSCO assignment. The results are presented in Table 4-8.

Table 4-7. Mean inspection expertise of the three models

Group	Inspection expertise of random PSCOs assignment	Mean inspection expertise of A1	Mean inspection expertise of A2	Mean inspection expertise of A3	Best inspection expertise in theory
1	80.56	84.45	84.00	86.55	89.40
2	44.03	45.67	46.42	46.18	48.30
3	49.10	51.20	51.17	51.08	52.40
4	39.24	39.27	39.65	39.00	43.00
5	34.11	37.10	37.14	36.72	38.20
6	25.36	26.86	26.72	27.13	28.30
7	48.15	50.83	50.92	50.81	51.90
8	61.70	64.10	63.97	64.21	67.10
9	32.41	34.35	34.12	34.42	35.40
10	20.22	20.80	20.74	21.23	23.10
11	60.19	64.00	64.20	64.19	65.30
12	33.69	35.11	35.11	34.84	36.80
13	33.84	34.48	34.43	34.88	37.90
14	37.98	38.60	38.14	38.20	41.60
15	22.63	24.43	24.71	24.52	25.70
16	63.36	66.45	66.22	66.88	69.50
17	27.67	29.08	28.19	29.14	30.40
18	38.16	38.82	38.56	38.60	40.80
19	31.75	34.76	34.95	33.69	36.00
20	44.36	47.58	47.85	47.67	49.50
21	32.82	34.26	34.33	34.55	36.20
22	31.22	34.19	33.80	34.27	35.10
23	29.67	31.37	31.40	31.49	34.30
24	33.42	34.09	33.99	34.36	37.00
25	51.16	52.93	53.00	53.24	55.70
26	23.23	25.18	24.99	24.07	26.60
27	25.88	26.50	26.65	26.62	28.70
28	60.52	61.99	62.02	62.28	65.60
29	22.75	24.97	24.55	25.22	26.00
30	27.88	28.50	28.15	27.57	31.10
Average	**38.90**	**40.73**	**40.67**	**40.79**	**42.90**
Ratio*	**(90.68%)**	**(94.94%)**	**(94.80%)**	**(95.08%)**	**(100%)**

Note*: calculated by $\frac{Average\ of\ mean\ inspection\ expertise}{The\ best\ inspection\ expertise\ in\ theory} \times 100\%$.

Table 4-8. Randomness of model performance

	Min inspection expertise			Max inspection expertise			Variance of inspection expertise		
Group/ inspection scheme	A1	A2	A3	A1	A2	A3	A1	A2	A3
1	83.4	83.7	83.4	86.3	85.7	87.1	1.4745	0.3600	1.1745
2	44.3	45.2	45.6	46.5	47.2	46.5	0.6261	0.2456	0.0876
3	51.2	50.9	50.3	51.2	51.2	51.2	0.0000	0.0081	0.0756
4	38.7	39.0	38.9	40.2	40.7	39.3	0.2361	0.1925	0.0120
5	37.1	37.1	35.1	37.1	37.5	37.1	0.0000	0.0144	0.3636
6	26.3	25.7	26.7	27.7	27.7	28.0	0.2844	0.4556	0.1641
7	50.7	50.8	50.8	50.9	51.2	50.9	0.0081	0.0096	0.0009
8	63.1	63.1	63.1	65.0	65.0	65.0	0.4400	0.2181	0.4089
9	33.6	33.8	34.0	34.8	34.8	34.8	0.2005	0.1216	0.0456
10	20.3	20.0	20.7	20.9	21.3	21.4	0.0300	0.1544	0.0621
11	64.0	64.0	64.0	64.0	64.7	64.7	0.0000	0.0940	0.0849
12	35.1	33.9	33.8	35.2	35.7	35.8	0.0009	0.1989	0.5864
13	33.2	33.2	34.2	35.7	35.5	35.3	0.5896	0.8121	0.0876
14	38.0	37.6	37.6	39.6	38.3	39.6	0.4440	0.0504	0.5740
15	23.9	23.9	24.2	24.8	25.2	25.4	0.0561	0.1049	0.1196
16	65.4	65.5	66.7	67.2	66.6	66.9	0.4145	0.1416	0.0036
17	28.1	27.8	28.1	29.4	28.8	29.4	0.2716	0.1029	0.2704
18	38.7	38.3	37.9	38.9	38.9	38.9	0.0096	0.0404	0.2100
19	33.6	33.6	32.6	35.6	35.6	34.3	0.5244	0.6585	0.2909
20	47.2	47.5	47.2	48.0	48.0	48.0	0.0996	0.0525	0.0801
21	33.8	33.8	33.8	34.7	34.8	35.1	0.1304	0.1161	0.1885
22	33.4	33.4	33.4	34.6	34.6	35.1	0.2189	0.1800	0.2261
23	30.5	30.3	31.4	31.6	31.6	31.6	0.0921	0.1420	0.0029
24	33.8	33.8	33.9	35.4	34.3	35.7	0.2069	0.0289	0.3444
25	52.7	53.0	52.7	53.2	53.0	53.5	0.0261	0.0000	0.0444
26	24.6	24.3	23.6	25.3	25.3	24.8	0.0596	0.1509	0.1821
27	25.7	26.2	25.9	26.7	26.7	26.8	0.1100	0.0225	0.0636
28	61.3	61.3	61.3	62.7	63.0	63.5	0.2109	0.2316	0.6176
29	24.5	24.1	25.0	25.2	25.0	25.3	0.0421	0.0565	0.0136
30	27.1	26.8	27.1	30.3	30.3	29.6	1.9640	1.3205	0.8461
Average	**40.11**	**40.05**	**40.10**	**41.29**	**41.27**	**41.35**	**0.2924**	**0.2095**	**0.2411**

Table 4-7 shows that on average, all the three models can realize about 95% of the best inspection expertise in theory on average while A3 has the best performance regarding mean inspection expertise. Table 4-8 indicates that the performance of A1, A2, and A3 are stable. We can draw the following conclusions:

(a) The performance of all the three newly proposed PSCO assignment schemes is stable and is much better than the performance of random PSCO assignment. This

shows that the PSCO assignment schemes generated by combining MTR-RF models with PSCO assignment models are efficient compared with the currently used random PSCO assignment at the port states.

(b) The performance of A1 is better than A2, although they both use MSE as the splitting criterion for constructing the MTR-RFs. The difference between A1 and A2 is that they have different prediction targets. The prediction targets in A1 are the deficiency numbers under each deficiency category which are natural choices. Meanwhile, the prediction targets in A2 are the deficiency numbers that can be identified by each PSCO which also considers the expertise of PSCs and is determined by the parameters of the following optimization model. The difference in performance of A1 and A2 indicates that although different targets can be chosen for a combined prediction and optimization model, their performance can be divergent.

(c) The performance of A3 is better than A2, although MTR-RF3 performs much worse as a regression model than MTR-RF2 if evaluated by MSE. This indicates that when combining ML model (e.g. decision tree and random forest) with optimization model, the choices for prediction targets, the properties of ML model (e.g. splitting criteria in decision trees), and model evaluation metrics can be varying. High-quality decisions are based on either precise prediction generated by the ML model or combination of the structure and property of the optimization model with the ML model.

(d) Table 4-8 indicates that A2 has the least variance while A1 has the largest variance in the total inspection expertise generated by the optimal assignment among A1, A2, and A3. The possible reasons are as follows. Although the splitting criterion of MTR-RF2 and that of MTF-RF3 is different, the outputs of MTR-RF2 and MTR-RF3 can serve as the parameters of the decision variables in the optimization model of A2 and A3 directly. On the other hand, the outputs of MTR-RF1 need to be further combined with the expertise of the PSCOs to serve as the parameters of the decision variables in the optimization model of A1. The further processing might magnify the variability of the total inspection expertise in the final optimal assignment decision, which leads to highest variance of A1 compared to A2 and A3. Although MTR-RF3 predicts the deficiency number detected by each PSCO like MTR-RF2, the splitting criteria in A3 is not relevant to the values of the prediction targets directly like that in MTR-RF2. Therefore, MTR-RF3 has larger variance in the outputs compared to MTR-RF2 as

shown in Table 4-6. When combining the prediction results as the input with the optimization models, the variability can be magnified. Therefore, A3 has the larger variance in the total inspection expertise generated by the optimal assignment decision compared to A2.

An illustration of insights of the superiority of A3 is presented in Appendix B.5. We also present the detailed inspection expertise under each deficiency category of A1, A2, and A3 as shown in Table 4-9.

Table 4-9. Inspection expertise under each deficiency category

Method/ deficiency category	C1: ship safety	C2: ship management	C3: ship condition and structure	C4: communication and navigation
Original test set	630	478	289	407
Best in theory	459.9	356.5	209.5	261
A1	447.24 (97.25%)	309.03 (86.68%)	201.81 (96.33%)	263.84 (101.09%)
A2	446.15 (97.01%)	309.10 (86.70%)	201.25 (96.06%)	263.59 (100.99%)
A3	446.72 (97.13%)	316.39 (88.75%)	198.89 (94.94%)	261.61 (100.23%)

It can be seen from Table 4-9 that A1, A2, and A3 can achieve more than 85% of the inspection expertise compared to the best situation in theory. Especially, except for C2: ship management, more than 95% of the best inspection expertise in theory can be achieved by the three combined prediction and assignment models. The results further indicate that all the three models that match PSCOs' inspection expertise with ship deficiency condition are effective and accurate.

4.5.3 Comparison with other state-of-the-art prediction models

In this section, comparisons of the proposed tree-based prediction models with other state-of-the-art and popular prediction models are made. We select three ML models for prediction: ridge regression, the least absolute shrinkage and selection operator (LASSO) regression, and support vector regression (SVR) for comparison. Their performance of predicting the deficiency number under each deficiency category is presented in Section 4.5.3.1 and the total inspection expertise realized when combining with assignment models is presented in Section 4.5.3.2.

4.5.3.1 Regression performance

All the three models are implemented by using scikit-learn in Python with the hyperparameter tuples tuned by grid search on the validation set. Like the experiments in Section 4.5.2, we also run the three models 10 times with the optimal hyperparameter tuples. Their performance is shown in Table 4-10.

Table 4-10. Prediction model performance

Model	Metric	Min	Mean	Max	Variance
MTR-RF1	MSE	3.9756	4.0173	4.0762	0.0009
MTR-RF2	MSE	15.4953	15.8342	16.1237	0.0437
MTR-RF3	MSE	16.7775	17.1684	17.5571	0.0443
Ridge regression	MSE	15.9990	15.9990	15.9990	0
LASSO regression	MSE	25.0756	25.0756	25.0756	0
SVR	MSE	26.0432	26.0432	26.0432	0

Table 4-10 indicates that the mean MSE of the outputs of ridge regression is smaller than that of MTR-RF3, while the mean MSE of the outputs of LASSO regression and SVR is bigger than that of MTR-RF2 and MTR-RF3. Besides, the outputs of the ridge regression, LASSO regression, and SVR are determined once the hyperparameters of the three models are fixed, therefore their performance is quite stable. While in the tree-based models, randomness in the outputs can still exist even if the hyperparameters are given.

5.3.2 PSCO assignment performance

We combine the prediction results generated by the three regression models with optimization model M2 and make comparison with A3 regarding the total inspection expertise, as A3 has the best performance in PSCO assignment among the proposed models. The assignment decision generated by combining ridge regression with M2, LASSO regression with M2, and SVR with M2 are denoted by A4, A5, and A6, respectively. Comparison results over the 30 groups of ships based on 10 runs are shown in Table 4-11.

Table 4-11. Comparison of PSCO assignment model performance

	Random assignment	A3 (MTR-RF3+M2)	A4 (ridge+M2)	A5 (LASSO+M2)	A6 (SVR+M2)	Best in theory
Average	38.90	**40.79**	40.59	40.36	40.48	42.90
Ratio*	90.68%	**95.08%**	94.62%	94.08%	94.36%	100%

Note*: calculated by $\frac{Average\ of\ mean\ inspection\ expertise}{The\ best\ inspection\ expertise\ in\ theory} \times 100\%$

Table 4-11 shows that A3 achieves the highest inspection expertise among A3, A4, A5, and A6, which indicates the superiority of the combined tree-based prediction model with the structure of optimization model.

4.5.4 Model extension

In the current prediction and assignment models, the importance of the four deficiency categories is viewed as identical. Nevertheless, their importance can be

different under certain situations, e.g., in the concentrated inspection campaign (CIC) where deficiencies in some categories should be paid more attention to in PSC inspections. To extend our models to deal with the situations where the importance of the deficiency categories is different, we attach each deficiency category with a relative importance score, which is denoted by $w_c, c=1,2,3,C$ and is no less than 1. The larger the value is, the more important the deficiency category is. In the current model, $w_c=1, c=1,2,3,C$. For mathematical model M1, we can combine the importance score directly in Equation (4.2) by revising the objective function to be $\max \sum_{p=1}^{P}\sum_{s\in S}\sum_{c=1}^{C} w_c \hat{\alpha}^{sc} u_{pc} z_{ps}$, while the prediction model MTF-RF1 needs not be revised. Then, we denote $\sum_{c=1}^{C} w_c \hat{\alpha}^{sc} u_{pc} = \lambda^{sp}$, which can be viewed as the weighted total deficiency number identified by PSCO $p=1,...,P$ of ship $s\in S$ and can be predicted by using the MTR-RF models developed in Section 4.1.1. The predicted values for λ^{sp} are denoted by $\hat{\lambda}^{sp}$, and the objective function of mathematical model M2 can be revised to $\max \sum_{p=1}^{P}\sum_{s\in S} \hat{\lambda}^{sp} z_{ps}$. Especially, the total inspection expertise generated by the three models where the differences in the deficiency category importance are considered is denoted by A1', A2', and A3' respectively.

We use an example to illustrate the working process and results of the proposed models where C1: ship safety is more important than the other deficiency categories. The relative importance score can be assigned by the ports in practice. In this example, we assume that $w_1=1.5$ and $w_c=1, c=2,3,C$. Mean inspection expertise of the three models is presented in Table 4-12. The performance of random assignment scheme is the mean inspection expertise of 10,000 times of random PSCO assignment. The inspection expertise under each deficiency category is shown in Table 4-13.

Table 4-12. Mean inspection expertise of the three models (considering deficiency category importance)

Group	Inspection expertise of random PSCOs assignment	Mean inspection expertise of A1'	Mean inspection expertise of A2'	Mean inspection expertise of A3'	Best inspection expertise in theory
1	80.56	83.68	82.72	83.77	89.40
2	44.03	46.60	46.29	46.54	48.30
3	49.10	51.47	51.41	51.31	52.40
4	39.24	39.44	39.29	38.86	43.00
5	34.11	37.24	37.26	37.17	38.20
6	25.36	27.23	27.02	26.66	28.30
7	48.15	50.93	51.10	50.96	51.90
8	61.70	64.45	64.41	64.23	67.10
9	32.41	34.60	34.75	34.68	35.40
10	20.22	19.79	19.85	20.09	23.10
11	60.19	63.87	63.93	63.81	65.30
12	33.69	34.56	34.44	34.76	36.80
13	33.84	34.31	34.60	34.11	37.90
14	37.98	39.56	39.20	39.70	41.60
15	22.63	24.23	25.03	24.72	25.70
16	63.36	66.24	66.60	67.34	69.50
17	27.67	28.65	28.62	29.01	30.40
18	38.16	38.90	39.10	38.47	40.80
19	31.75	33.84	33.77	33.06	36.00
20	44.36	47.10	47.41	47.00	49.50
21	32.82	34.13	34.00	34.06	36.20
22	31.22	33.73	33.64	34.45	35.10
23	29.67	31.33	31.07	31.02	34.30
24	33.42	33.87	34.52	34.29	37.00
25	51.16	53.21	53.04	53.39	55.70
26	23.23	24.16	24.48	24.63	26.60
27	25.88	26.70	26.74	26.70	28.70
28	60.52	62.15	62.44	63.17	65.60
29	22.75	23.84	23.66	24.02	26.00
30	27.88	29.73	29.73	29.41	31.10
Average	**38.90**	**40.65**	**40.67**	**40.71**	**42.90**
Ratio*	**(90.68%)**	**(94.76%)**	**(94.80%)**	**(94.90%)**	**(100%)**

Note*: calculated by $\frac{Average\ of\ mean\ inspection\ expertise}{The\ best\ inspection\ expertise\ in\ theory} \times 100\%$.

Table 4-13. Inspection expertise under each deficiency category

Method/ deficiency category	C1: ship safety	C2: ship management	C3: ship condition and structure	C4: communication and navigation
Original test set	630	478	289	407
Best in theory	459.9	356.5	209.5	261
A1	447.24 (97.25%)	309.03 (86.68%)	201.81 (96.33%)	263.84 (101.09%)
A1’	449.58 (97.76%)	303.20 (85.05%)	204.27 (97.50%)	262.49 (100.57%)
A2	446.15 (97.01%)	309.10 (86.70%)	201.25 (96.06%)	263.59 (100.99%)
A2’	449.68 (97.78%)	303.30 (85.08%)	203.57 (97.17%)	263.57 (100.98%)
A3	446.72 (97.13%)	316.39 (88.75%)	198.89 (94.94%)	261.61 (100.23%)
A3’	451.95 (98.27%)	307.29 (86.20%)	201.22 (96.05%)	260.93 (99.97%)

Table 4-12 shows that if different weights of deficiency categories are taken into account, the total inspection expertise achieved by the three inspection strategies is no larger than the situation when the deficiency categories are of the same importance. Moreover, if C1 is regarded to be more important and is attached with a larger importance score, the realized inspection expertise under C1 increases in all the three inspection strategies as presented in Table 4-13.

4.5.5 Sensitivity analysis

In this section, we analyze how the distribution of the expertise of PSCOs would influence the performance of the proposed PSCO assignment models. To be concise, the sensitivity analysis is conducted on A3 as it achieves the maximum mean inspection expertise among A1, A2 and A3. Four groups of sensitivity analyses (SA) are performed: SA1: composition of a group of PSCOs; SA2: divergence in expertise of a PSCO; SA3: adequacy of PSCO resources; SA4: uncertainty in PSCOs’ expertise.

4.5.5.1 SA1: composition of a group of PSCOs

First, we analyze how the composition of a group of PSCOs would influence the results. Suppose there are five groups of PSCOs (denoted by SA1G1 to SA1G5, respectively) of the same total expertise and different expertise distributions while one PSCO has the same expertise to inspect the four deficiency categories. Groups SA1G1 to SA1G4 contain PSCOs with various expertise, i.e. some of them are experienced while some are newcomers. More specifically, the variations of the expertise of each PSCO are increasing from SA1G1 to SA1G4. On the contrary, the four PSCOs in SA1G5 have the same expertise. The expertise of each PSCO to each deficiency category of the five groups is shown in Table 4-14. The analysis results of SA1 are shown in Figure 4-1.

Table 4-14. Expertise of PSCOs in SA1

SA1G1	C1	C2	C3	C4	**SA1G 2**	C1	C2	C3	C4
PSCO 1	0.775	0.775	0.775	0.775	PSCO 1	0.85	0.85	0.85	0.85
PSCO 2	0.725	0.725	0.725	0.725	PSCO 2	0.75	0.75	0.75	0.75
PSCO 3	0.675	0.675	0.675	0.675	PSCO 3	0.65	0.65	0.65	0.65
PSCO 4	0.625	0.625	0.625	0.625	PSCO 4	0.55	0.55	0.55	0.55
SA1G3	C1	C2	C3	C4	**SA1G4**	C1	C2	C3	C4
PSCO 1	0.925	0.925	0.925	0.925	PSCO 1	1.0	1.0	1.0	1.0
PSCO 2	0.775	0.775	0.775	0.775	PSCO 2	0.8	0.8	0.8	0.8
PSCO 3	0.625	0.625	0.625	0.625	PSCO 3	0.6	0.6	0.6	0.6
PSCO 4	0.475	0.475	0.475	0.475	PSCO 4	0.4	0.4	0.4	0.4
SA1G5	C1	C2	C3	C4					
PSCO 1	0.7	0.7	0.7	0.7					
PSCO 2	0.7	0.7	0.7	0.7					
PSCO 3	0.7	0.7	0.7	0.7					
PSCO 4	0.7	0.7	0.7	0.7					

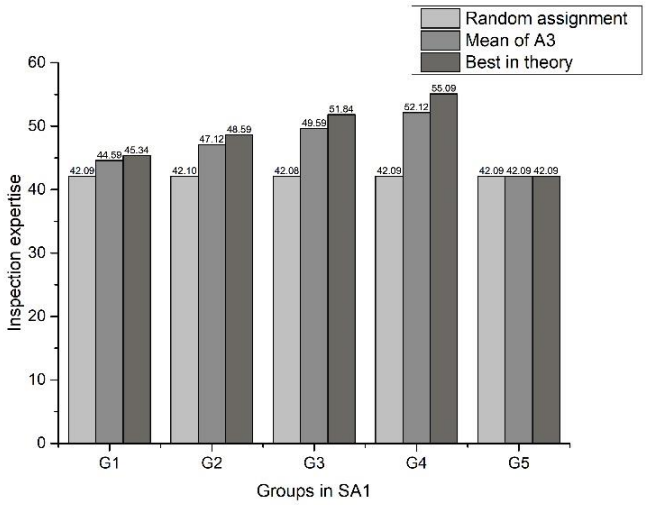


Figure 4-1. Analysis results of SA1

Several conclusions can be drawn from Figure 4-1. First, as the divergence of expertise of the group of PSCOs become larger, both the best inspection expertise in theory and the mean inspection expertise achieved by using A3 increase, as the diverse conditions of the inspected ships can be better matched with the more varied expertise of the group of PSCOs. Second, the superiority of the PSCO assignment scheme generated by A3 over random PSCO assignment becomes more obvious when the inspection expertise of the PSCOs gets more diverse. Third, the mean inspection expertise achieved by A3 is equal to the best inspection expertise in theory when all the PSCOs have the same expertise. However, as the expertise of the group of PSCOs gets more varied, the gap between mean inspection expertise and the best inspection expertise in theory gets larger. This indicates that predicting errors of the prediction model (i.e., MTR-RF3) have a larger influence on the final assignment model when

the expertise of PSCOs becomes more diverse, as the assignment scheme relies more on the predicted number of deficiencies of a ship that can be identified if assigned to a PSCO.

The extreme situation is that when all the PSCOs have the same expertise, the mean inspection expertise achieved by A3 equals the best inspection expertise in theory and random PSCO assignment, as the PSCO assignment is totally random under this condition and has nothing to do with the prediction results of MTR-RF3. Nevertheless, it should also be noted that even in SA1G4, where the expertise of the PSCOs is most varied, the mean inspection expertise is approaching 95% of the best inspection expertise in theory, and the PSCO assignment performance of A3 is 24% better than random PSCO assignment. The results indicate that our model is more suitable to be applied than random PSCO assignment scheme when the expertise of the PSCOs is divergent. When the expertise of all the PSCOs is the same, our model is equal to random PSCO assignment.

4.5.5.2 SA2: divergence in expertise of a PSCO

Second, we analyze how various expertise of a PSCO in different deficiency categories would influence the results. We consider four groups of PSCOs (denoted by SA2G1 to SA2G4, respectively) where the total expertise is the same for each PSCO and the total expertise to inspect one deficiency category is the same for each group (i.e., the sum of each row and the sum of each column are the same in all groups). In SA2G1 to SA2G3, the PSCOs have different expertise to inspect different deficiency categories, while in SA2G4, all PSCOs have the same expertise in different deficiency categories. More specifically, the variations of the PSCOs are increasing from SA2G1 to SA2G3: the sum of absolute variations of all PSCOs in SA2G1, SA2G2 and SA2G3 is 1.8, 2.2 and 2.6, respectively. The expertise of each PSCO in each deficiency category of the four groups is shown in Table 4-15. The results of the analyses are presented in Figure 4-2.

Table 4-15. Expertise of PSCOs in SA2

SA2G1	C1	C2	C3	C4	**SA2G2**	C1	C2	C3	C4
PSCO 1	0.9	0.8	0.6	0.5	PSCO 1	0.8	0.5	1.0	0.5
PSCO 2	0.6	0.8	0.7	0.7	PSCO 2	0.9	0.8	0.5	0.6
PSCO 3	0.5	0.6	0.9	0.8	PSCO 3	0.6	0.7	0.6	0.9
PSCO 4	0.8	0.6	0.6	0.8	PSCO 4	0.5	0.8	0.7	0.8
SA2G3	C1	C2	C3	C4	**SA1G4**	C1	C2	C3	C4
PSCO 1	0.8	0.9	0.7	0.4	PSCO 1	0.7	0.7	0.7	0.7
PSCO 2	0.8	0.4	0.8	0.8	PSCO 2	0.7	0.7	0.7	0.7
PSCO 3	0.8	0.5	0.8	0.7	PSCO 3	0.7	0.7	0.7	0.7
PSCO 4	0.4	1.0	0.5	0.9	PSCO 4	0.7	0.7	0.7	0.7

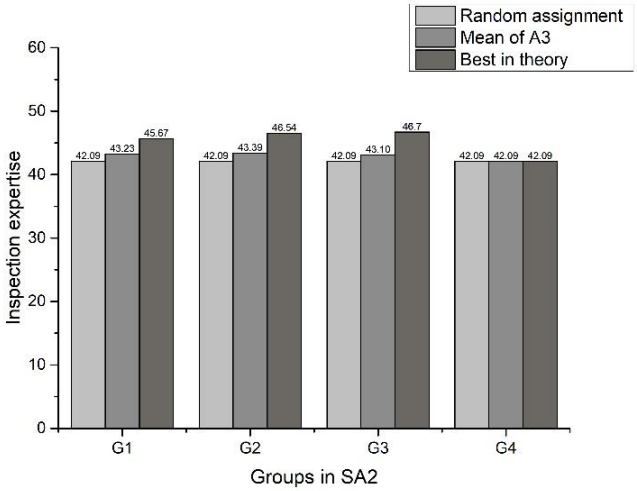


Figure 4-2. Analysis results of SA2

As shown in Figure 4-2, when the total expertise of each PSCO is the same and the total expertise to inspect one deficiency category for each group is the same, the best inspection expertise in theory shows gentle increase when the divergence of the PSCOs' expertise increases. Nevertheless, due to the randomness in the dataset and the model performance, the predicted mean inspection expertise does not show this trend: when the variations in the expertise of the PSCOs increase, the predicted mean inspection expertise can either increase or decrease modestly.

4.5.5.3 SA3: adequacy of PSCO resources

Third, we analyze the influence of the adequacy of PSCO resources on the inspection results. In our problem, four PSCOs are assigned to inspect 10 ships coming to the port state every day (benchmark, denoted by SA4G3). The maximum number of ships that can be inspected by one PSCO is three. We consider other situations where there are 8 (SA3G1), 9 (SA3G2), 11 (SA3G4), and 12 (SA3G5) ships coming to the port state every day while keeping the other settings unchanged and compare the mean

inspection expertise of a single ship to identify the influence of PSCO resources. The results are shown in Figure 4-3.

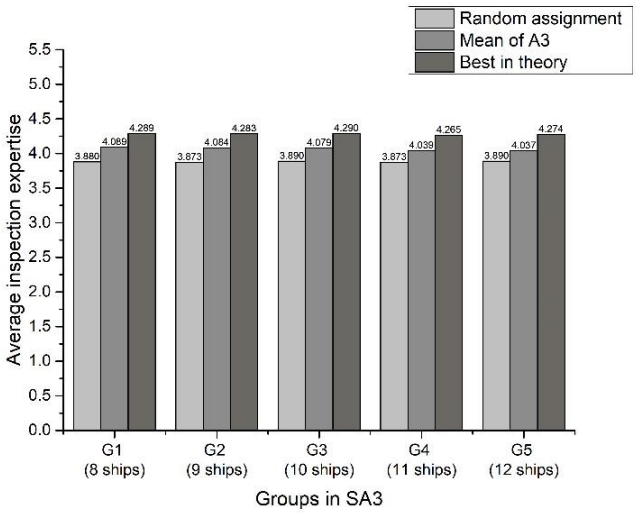


Figure 4-3. Analysis results of SA3

As shown in Figure 4-3, when the number of ships in a group grows while the number of PSCOs and the maximum number of ships can be inspected by one PSCO remain unchanged, the average inspection expertise of one ship remains stable. This indicates that the performance of the proposed models is not heavily influenced by the adequacy of the resources of PSCOs, which also shows that the model performs robustly. Besides, our model performs much better than random PSCO assignment in all situations.

5.5.4 SA4: uncertainty in PSCOs' expertise

Fourth, we analyze the uncertainty in PSCO expertise in each deficiency category. Although the expertise of PSCOs could be measured by tests, interviews, and questionnaires, uncertainties can exist, which means that the expertise we obtained may not be the exact expertise in reality. The expertise values presented in Table 4-4 are the measured inspection expertise and we suppose that the real inspection expertise is within 10% more or less than the measured inspection expertise. For example, the expertise of PSCO 1 for deficiency category C1 is 0.8, and we suppose that the real inspection expertise is uniformly distributed from 0.72 to 0.88 (accurate to two digits). We randomly select a value within this interval for each inspection expertise value and form a new expertise table of each PSCO in each deficiency category for ten times, and we can obtain ten possible real inspection expertise tables. The inspection

expertise of random PSCO assignment, the best inspection in theory, and the mean inspection expertise of the ten groups are shown in Figure 4-4.

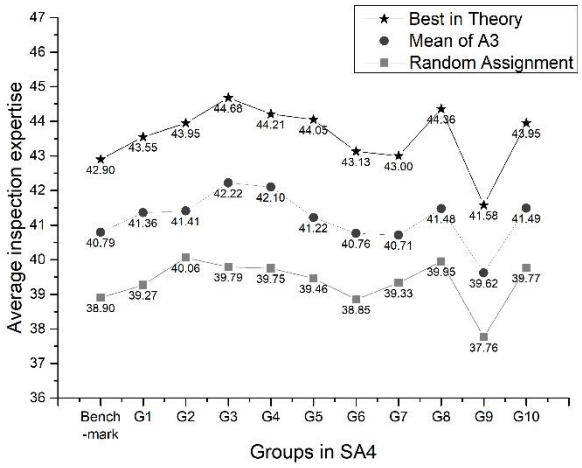


Figure 4-4. Analysis results of SA4

Under the assumption that the real inspection expertise of each PSCO to each deficiency category is within 10% more or less than the measured inspection expertise presented in Table 4-5, the variance of the best inspection expertise in theory of the 10 groups in SA4 is 0.8298. The range of the best inspection expertise in theory is 3.1 (the maximum inspection expertise of the 10 groups is 44.68 and the minimum inspection expertise is 41.58). Compared with the benchmark, which is generated by using the measured inspection expertise shown in Table 4-4, the differences are between －3.08% and ＋4.15% and are much smaller than ±10%.

As for the predicted inspection expertise, the variance of the 10 groups in SA4 is 0.4999, and the range of mean inspection expertise is 2.6 (the maximum inspection expertise of the 10 groups is 42.22 and the minimum inspection expertise is 39.62). The differences between the benchmark and the 10 groups range from －2.87% to ＋3.51% and are also much smaller than ±10%. We also compare the differences between predicted mean inspection expertise of the benchmark with the best inspection expertise of the 10 groups in SA4. The difference is from ＋1.94% to ＋9.54%, which indicates that the proposed models are robust even if there are some uncertainties in measuring the inspection expertise of each PSCO to each deficiency category. The average best inspection expertise of the 10 groups is 43.646 and the average predicted inspection expertise of the 10 groups is 41.237, which indicates that the proposed model can identify about 94.5% of the total deficiencies and that it always performs

better than random PSCO assignment in all groups of tests as shown in the two lower lines of Figure 4-4.

4.6 DISCUSSION AND FUTURE RESEARCH

As indicated in Section 4.5.2.1, the inspection expertise of each PSCO in each deficiency category shown in Table 4-4 is assumed by the authors as there is no such standard tests or questionnaires at the moment in the Tokyo MoU. The assumption of the inspection expertise table is that each PSCO has more expertise in one or two deficiency categories than the other PSCOs and we just use the assumed inspection expertise to illustrate the working process of the proposed models. Although massive sensitivity analysis has been conducted to evaluate the performance of the proposed models, in future research, accurate assessments would be developed to evaluate the real expertise of the PSCOs for different deficiency categories. For example, a test consisting of a theoretical part and a practical part of all the four deficiency categories, or an interview regarding the background, experience, and self-evaluation of the PSCOs, or a questionnaire for collecting the PSCOs' own preference and expertise can be held. The results of the test, interview, and questionnaire can be considered simultaneously to comprehensively evaluate the inspection expertise of the PSCOs. For the convenience of MoU management, we propose another way to evaluate the performance of the PSCOs. Suppose that there are several PSCOs at a port, and we let them to inspect a group of ships (say 10 ships or 20 ships) in a certain amount of time. Then, we compare the total number of detected deficiencies under each deficiency category of the PSCOs regarding all the ships. For the PSCO(s) who can identify the most deficiency number of a category, we denote her/his expertise to be "1". The expertise of the other PSCOs regarding this deficiency category is calculated by dividing her/his number of detected deficiencies in this category by the largest number of detected deficiencies of this category of all PSCOs. For example, the detected deficiency number for deficiency category C1 is 20, 18, 16, and 14 for PSCO1, PSCO2, PSCO3, and PSCO4, and their inspection expertise for C1 should be 1, 0.9, 0.8, and 0.7, respectively. In this way, evaluating and updating the inspection expertise of the PSCOs can be more convenient for the ports.

Another thing should be noted is that the expertise of the PSCOs can be updated over time and experience. Therefore, reevaluations should be carried out for updating the expertise of the PSCOs. We suggest that the reevaluation to be carried out once a

year for the following reasons. First, the committee meeting of Tokyo MoU is held once a year. In the committee meeting, several important discussions and decisions are made, such as the application for co-operating member status, actions relating to harmonization of PSC, and approval of the final report of the concentrated inspection campaign (CIC) in the 29th committee meeting in 2018 (Tokyo MoU, 2018b). Therefore, it should be a good chance to discuss the details of reevaluations at the committee meeting. Second, setting the time interval of two reevaluations to be one year is a result of a trade-off: for one thing, the inspection expertise for the PSCOs could remain unchanged for only a period of time; for another, it can be time-consuming to prepare for the reevaluations.

Given the fact that the inspection expertise of the PSCOs can improve over time and experience and it is also a goal to improve the PSCOs' inspection expertise to be as close as possible to 1, we propose two ways to achieve the comprehensive development of the PSCO if the proposed models are applied. First, except for the assigned PSCO who is responsible to conduct the PSC inspection, other PSCOs can get onboard during the PSC inspections to learn from the PSCO with more inspection expertise in certain deficiency categories to achieve self-improvement. Second, during the regular trainings and seminars, the PSCOs can share their experience and expertise as well as discuss the difficulties they meet during the inspections with each other to achieve co-operation and progress.

As the main goal of PSC is to identify substandard ships and detain them if necessary to protect the maritime safety and protect the marine environment, ship detention probability can also be incorporated in the prediction and assignment models in the future research for better applicability and practicability. Meanwhile, PSCOs' expertise in targeting ships with high detention probability should also be evaluated and considered.

4.7 CONCLUSION

Maritime safety and the marine environmental protection are gaining increasing concern in recent years. PSC inspection is a widely-believed effective and efficient way to safeguard the sea. To improve the efficiency of PSC inspections, one of the key points is to identify as many deficiencies as possible using limited inspection resources. At the ports with less experienced and divergent PSCOs, this requires matching the

PSCOs of different expertise and the deficiency conditions of the inspected ships, e.g. the deficiency number under each deficiency category. To achieve this goal, this chapter proposes three ML models: MTR-RF1, MTR-RF2, and MTR-RF3 and two PSCO assignment models M1 and M2 to match the diverse ship deficiency conditions with the expertise of PSCOs. More specifically, MTR-RF1 predicts the number of deficiencies in each deficiency category for each ship in a way that minimizes the MSE between actual and predicted numbers of deficiencies; MTR-RF2 predicts the number of deficiencies each PSCO can identify for each ship by minimizing the MSE between actual and predicted deficiency numbers; MTR-RF3 predicts the number of deficiencies each PSCO can identify for each ship while adopting a loss function motivated by the structure of the optimization problem, i.e. minimizing the MSO in the numbers of deficiencies that can be detected among the PSCOs for each ship. Numerical experiments show that the performance of combination of MTR-RF3 and M2 (i.e., A3) is the best among the three proposed models, while all the three models perform much better than the currently used random PSCO assignment as they can identify about 95% of all the deficiencies compared to the best inspection expertise in theory.

By conducting sensitivity analyses, several managerial insights can be drawn. First, our model is more suitable to be applied when the expertise of the PSCOs is divergent as the superiority of the proposed models becomes more obvious when the divergence of the PSCOs increases. Second, the adequacy of the PSCO resources would not heavily influence the performance of the proposed models. Besides, even if uncertainty may exist in measuring the expertise of the PSCOs to each deficiency category, the robustness of our model is validated.

In this chapter, both prediction and optimization are required to generate the decision for PSCO assignment. Meanwhile, both prediction and optimization are challenging tasks, as errors cannot be avoided in the prediction problem, while the unknown parameters in the optimization model are determined by the outputs of the prediction model. For A1 and A2, the ML models for parameter prediction totally ignore the downstream optimization problem and only aim to minimize the prediction error which is evaluated by MSE. Although the objective is to make the predicted outputs as close as to the real outputs, inaccuracy always exists, and thus minimizing the output error cannot guarantee the best decision in theory generated by the following

optimization model or generate the decision as close as to the best decision in theory. Moreover, inaccuracy in the predicted results is highly likely to be magnified when combining with the downstream optimization model and thus make the final generated decision far away from the best decision in theory. On the contrary, MSO used in MTR-RF3 (and thus in A3) is highly related to the structure and property of the downstream optimization model, as the prediction model is designed to generate outputs that make the generated decisions of the following optimization model as close as to the best decision in theory by aiming to maintain the property of the parameters in the optimization model for generating the best decision in theory.

Theoretically, the proposed MTR-RF1 and MTR-RF2 treat prediction and optimization models as sequential steps while the proposed MTR-RF3 partially combines prediction and optimization models by considering the structure and property of the optimization model when constructing the ML model. The numerical experiments show that although MTR-RF3 performs much worse than MTR-RF2 as a regression model evaluated by the metric of MSE, the performance of MTR-RF3 is better than MTR-RF2 when combining with the following optimization models. Practically, the proposed models help to address a meaningful practical problem in PSC inspection. Compared with random assignment of PSCOs, the proposed three models can help to detect 4.70%, 4.55%, and 4.86% more deficiencies after inspecting the same groups of ships by using the same PSCO recourses. Meanwhile, the performance of the three models is stable and their performance would achieve 95% of the best inspection expertise in theory.

Chapter 5: Efficient and Explainable Ship Selection Planning[5]

The prediction models proposed in Chapter 3 and Chapter 4 are of black-box nature, which means that the prediction results and the working mechanism of the prediction models are unexplainable to model users. As a result, their popularity and reliability are likely to be weakened. This Chapter aims to develop explainable and interpretable prediction models of ship total deficiency number. It first develops a data-driven ship risk prediction framework using features the same as the current ship selection scheme. Like the existing ship risk prediction models, the proposed framework is of black-box nature whose working mechanism is opaque. To improve model explainability, local explanation of the prediction of individual ships by the Shapley additive explanations (SHAP) with the properties of local accuracy and consistency is provided. Furthermore, we innovatively extend the local SHAP model to a fully-explainable near linear-form global surrogate model of the original black-box data-driven model by deriving feature coefficients and fitting curves of feature values and SHAP values from the SHAP value matrix. This demonstrates that the behavior of black-box data-driven models can be as interpretable as white-box models while retaining their prediction accuracy. Numerical experiments demonstrate that the white-box global surrogate model can accurately present the behavior of the original black-box model, shedding light on model validation, fairness verification, and prediction explanation, and hence promote their acceptance and application among maritime stakeholders.

5.1 INTRODUCTION

Given a large number of foreign visiting ships and the limited inspection resources at a port, only a small proportion of the ships can be inspected by PSC. For example, only 13.05% of all the foreign ships visiting the Hong Kong Port were inspected in 2019 (Tokyo MoU, 2020). Meanwhile, globally, less than half of the inspected ships were with deficiency detected during 2018 and 2020, while only 2.50%

[5] Yan, R., Wu, S., Jin, Y., Cao, J., Wang, S., 2022. Efficient and explainable ship selection planning in port state control. Transportation Research Part C: Emerging Technologies, under review.

of them were detained (i.e., with very serious deficiency or deficiencies detected) during this period (Marine Department, 2021). This indicates that accurate identification of substandard ships and rational allocation of the scarce inspection resources at port is the key to improve the effectiveness of PSC while reducing the delay of the fast turnover of the maritime logistics systems brought about by non-essential inspections. Moreover, the cost of a PSC inspection can be very high: according to Tokyo MoU, the charge of the first hour of follow-up inspection at the Hong Kong Port is 3,270 HKD (about 420 USD) and that of the subsequent hours is 1,115 HKD (about 143 USD) per hour, and the documentation fee is 1,115 HKD (about 143 USD) per hour (Tokyo MoU, 2016). Therefore, correct identification and inspection of high-risk ships can not only improve inspection efficiency, but also save resources and reduce costs.

This chapter aims to develop and explain a state-of-the-art data-driven ML based ship risk prediction model to assist port states in identifying and selecting high-risk foreign visiting ships. We use six years' PSC inspection records at the Hong Kong Port to develop a ship risk prediction framework based on a gradient boost regression trees (GBRTs) to predict ship deficiency number. Features in the proposed framework are the same as those considered in the current ship selection scheme applied by the Tokyo MoU. Post-hoc, model-agnostic, and local explanations are then given by Shapley additive explanations (SHAP) method, aiming to explain the prediction of individual ships. We further extend the local SHAP method to a global explanation method taking a near linear form by calculating the average SHAP values of different states for categorical features and fitting curves of feature values and SHAP values for integer and continuous features. Thorough analysis of model explanations is given to draw policy insights and managerial recommendations for both port authorities as well as ship owners and managers. To be more specific, this study makes the following contributions.

From theoretical perspective, we extend the local SHAP method to a global explanation method in an intuitive and succinct way, showing that the prediction behavior of black-box models (e.g., the GBRT models to predict ship deficiency number in this study) can be presented by white-box models (e.g., the extended SHAP model taking a near linear form) without compromising their prediction performance under arbitrary problem setting. The near linear-form global surrogate model is derived

directly from local explanations, and thus can illustrate the average contribution of each feature value to the final prediction on the whole dataset. Such unification of local and global explanations can make the interpretation of black-box model more comprehensive and consistent. Furthermore, we demonstrate that the black-box ship risk prediction model is of satisfactory accuracy, and the difference between the predictions given by the original black-box model and that given by the near linear-form global surrogate model is minor. In addition, model explanations given by local and global feature importance scores, beeswarm plots, and near linear-form global surrogate model are comprehensible to port authorities and ship owners, operators, and management companies. The explanations are also essential for them to trust and apply the proposed frameworks. Therefore, the explanations can be validated to follow the 'predictive, descriptive, and relevant' framework for black-box model explanation evaluation (Murdoch et al., 2019).

From practical perspective, to the best of the authors' knowledge, this is the very first study that explores explanations of black-box prediction models in maritime transport and thus paves the way of adopting ML models (which is a typical type of black-box model) to address maritime transport problems. Especially, a critical problem in a major international shipping policy is addressed in this study, i.e., high-risk ship selection in PSC. Only the factors considered in the current ship selection scheme are used for developing the ship selection framework, making it more applicable to port authorities. Thorough explanations of the black-box prediction model further make it more comprehensible and acceptable by port authorities as well as ship owners and managers. Numerical experiments show that the proposed ship selection framework is more efficient in identifying high-risk ships compared to the current ship selection scheme.

From policy making point of view, the comprehensive and consistent explanations provided in this study make an initial step to bridge the gap between making a prediction and making a decision in maritime transport area from at least three perspectives: trustworthiness, fairness, and informativeness. Disclosing the inner working mechanism and decision process of a black-box prediction model can help to verify whether the predictions given by a black-box model comply with domain knowledge. If yes, the proposed black-box prediction model can be expected to be more trustable and acceptable by decision makers. Fairness of the recommendations

made by black-box prediction models is a main concern of policy makers, which can also be validated by investigating the coefficients and curves of the features in the global surrogate models developed in this study. Insights extracted from practical data can shed light on policy and decision makings in the future, and thus enhance the informativeness of model explanation.

5.2 LITERATURE REVIEW AND RESEARCH GAP

5.2.1 Explainability of ship risk prediction model

Based on the literature reviewed in Chapter 2, the dataset used, features considered, risk indicators and prediction model developed, and model explainability of the studies on ship risk prediction are provided in Table 5-1.

Table 5-1. Summary of studies on ship risk prediction for PSC inspection

Literature	Dataset	Features considered	Risk indicator	Risk prediction model	Explainability
Xu et al. (2007a)	5,000 ships with more than 4 inspection records in the Paris MoU from January 2003 to January 2007	Generic factors: ship age, type, tonnage, flag, classification society, company, History factors: the number of deficiencies, outstanding deficiencies, duplicate deficiencies, duplicate outstanding deficiencies, and detentions in past 4 inspections, time since last initial inspection	Ship detention	SVM	No
Xu et al. (2007b)	The same as Xu et al. (2007a)	Numbers of non-lasting and lasting equipment/operation deficiencies, number of outstanding non-lasting and lasting equipment/operation deficiencies, numbers of deficiencies/outstanding deficiencies in areas 1 to 8 in past 4 inspections and the features considered by Xu et al. (2007a)	Ship detention	SVM	No
Gao et al. (2007)	140,000 inspection records in the Tokyo MoU	15 features including ship generic factors, dynamic factors, and history factors	Ship detention	KNN-SVM	No
Wu et al. (2021)	Inspection records of general cargo ship from 2014 to 2018 in the Tokyo MoU	Ship age, number of deficiencies, and 5 types of deficiencies selected by AHP and GRA	Ship detention	SVM	No
Yang et al. (2018a)	Inspection records of bulk carriers from 2005 to 2008 in the Paris MoU	Ship flag, RO, deadweight tonnage, age, inspection type, inspection port, and the number of deficiencies detected	Ship detention	BN model	Partially explainable, presented by conditional probability
Yang et al. (2018b)	Inspection records of bulk carriers from 2015 to 2017 in the Paris MoU	Ship flag, age, company performance, inspection type, inspection port, inspection date, and the number of deficiencies detected	Ship detention	BN model	Partially explainable, presented by conditional probability
Wang et al. (2019)	Inspection records in 2017 at the Hong Kong Port in the Tokyo MoU	Ship age, GT, type, flag performance, company performance, RO performance, last inspection time, the number of deficiencies in last inspection, the number of previous detentions, and the number of times of changing flag	Ship deficiency number	BN model	Partially explainable, presented by conditional probability
Yan et al. (2020)	Inspection records from 2016 to 2018 at the Hong Kong Port in the Tokyo MoU	Ship age, GT, length, depth, beam, type, the number of times of changing flag, total detention times, casualties in last five years, ship flag, RO, and company performance, last inspection time, last deficiency number, follow-up inspection rate	Ship deficiency number under each deficiency category	RF models consisting of multi-target regression trees	No

Yan et al. (2021b)	Inspection records from 2016 to 2018 at the Hong Kong Port in the Tokyo MoU	Ship age, GT, type, depth, length, beam, the number of times of changing flag, casualties in the last 5 years, total detentions, ship flag, RO, and company performance, last inspection time, last deficiency number, and follow-up inspection rate	Ship detention	BRF model	No
Yan et al. (2021a)	Inspection records from 2016 to 2018 at the Hong Kong Port in the Tokyo MoU	Ship age, GT, length, depth, beam, type, ship flag, RO, and company performance, last inspection date, last deficiency number, total detentions, the number of flag changes, and casualty in last 5 years	Ship deficiency number	XGBoost model	No
Degré (2007)	IMO casualty records from 1998 to 2003	Ship type, size, and age	Ship risk evaluated by the probability of the occurrence of casualties and their potential consequences	A statistical model	Yes
Degré (2008)	Casualty descriptive statistics and world merchant fleet descriptive statistics	Ship type, size, and age	Black-grey-white lists of categories of ships	A binomial calculation method	Yes
Heij and Knapp (2019)	IHS Markit for ship-particular data, ship incident database from 2010 to 2014, and ship inspection database from 2010 to 2014	A total of thirty factors with more than 500 variables, such as flag, owner, engine designer and builder are contained in the initial model, while only significant variables are contained in sub-models	Ship inspections, detentions, and very serious and serious incidents	A logit model	Yes
Knapp and Heij (2020)	IHS Markit for ship-particular data, ship incident database from 2010 to 2014, and ship inspection database from 2010 to 2014 for estimating risk formulas and probabilities, and quarterly data of incidents, inspection and ship particular data for estimating probabilities	Over 500 variables are contained in the initial model, and 16 to 172 variables are contained in the sub-models	Ship inspections, detentions, and very serious and serious incidents	A combination of logit model and percentage rank model	Yes
Dinis et al. (2020)	Inspection records of 136 ships at the port of Lisbon in the Paris MoU in 2018, and AIS data of 25 ships that have entered the same port	Ship type, age, flag, RO, company, deficiency and detention within the last 3 years	SRP with more detailed states	BN	Partially explainable, presented by conditional probability

The above analysis indicates that one of the largest gaps in current literature is the lack of model explainability. On the one hand, except for BN, which is partially explainable, all the other models for direct ship risk prediction are in a total black-box nature. It is also noted that although Naive Bayes, which is the most basic type of BN model, can be viewed as a type of interpretable model (Molnar, 2020), its interpretability is due to the underlying independence assumption, and thus the contribution of each feature towards the prediction target is clearly presented by the conditional probability tables. However, as Naive Bayes models usually oversimplify the reality, their accuracy is highly compromised. Therefore, none of the BN models developed in the abovementioned research is Naive Bayes model. Instead, BNs with more complex structures, especially those taking interdependencies among the variables into account, were developed for ship risk prediction. Consequently, interpretability of these BNs is largely weakened, especially those containing

intermediate variables such as Yang et al. (2018a, 2018b) and Dinis et al. (2020). On the other hand, although the statistical models employed for indirect ship risk prediction are interpretable to a certain degree, their predictive power could be weaker than that of the state-of-the-art ML models (Murdoch et al., 2019).

Another gap in existing literature is the features used for ship risk prediction. Table 5-1 indicates that except for Dinis et al. (2020) where only factors in the SRP are considered to predict ship risk, external databases with different degrees of difficulty in obtaining are used by all the other studies. Consequently, these models might be hard to be adopted by the conservative port authorities as such external datasets may not be trusted by them and much more time, efforts, and money might be spent on obtaining and processing the required data. Meanwhile, although only the factors in the SRP were considered to develop ship risk prediction models in Dinis et al. (2020), the prediction target of more detailed ship risk profile (a total of 14 risk levels) is abstract and might hard to be verified.

5.2.2 Explainable artificial intelligence in transportation research

To make the literature review more comprehensive, we also briefly review the existing literature on exploring explainable artificial intelligence (XAI) in transportation research. ML and deep learning approaches have been adopted by a large number of works in the transportation field, but only quite a few have addressed model explainability issue (Kalatian and Farooq, 2021), and most of the related studies are published in recent three years. These explanation methods can be divided into two types: global explanation, which aims to explain the entire model behavior, and local explanation, which aims to explain an individual prediction. Features' relative importance to the prediction target is the most common way of global explanation, which can be found in Zhang and Haghani (2015) for highway travel time prediction, Hagenauer and Helbich (2017) for travel mode choice prediction, and Chen et al. (2017) for passengers' ridesplitting behavior prediction. In addition, Wang et al. (2020) developed a decision tree as a surrogate of the black-box prediction model for congestion attack prediction. Meanwhile, local explanation is achieved by SHAP in Barredo-Arrieta et al. (2019) for traffic flow prediction, Veran et al. (2020) for crash prediction, and Kalatian and Farooq (2021) for pedestrians' wait time prediction. Other common methods for local explanation, such as partial dependence plot (PDP),

individual conditional expectation (ICE), and accumulated local effect (ALE) were used in Khoda Bakhshi and Ahmed (2021) for road crash probability prediction.

Both global and local explanations are provided by some studies via separate approaches. Especially, global explanation was also mainly achieved by deriving feature importance or feature interactions, while local explanation was reached by PDP in Zhao et al. (2018) for travel mode switching behavior prediction, by SHAP in Parmar et al. (2021) for parking duration prediction, by ALE in Kim et al. (2020) for passenger transit purpose prediction, and in Kim (2021) for travel mode choice prediction, by local interpretable model-agnostic explanations (LIME) in Bukhsh et al. (2019) for rail maintenance need prediction and management, and by PDP and ALE in Xu et al. (2021) for ridesplitting adoption prediction.

The studies covered in this subsection so far mainly adopt existing explanation methods. Besides, researchers have also proposed innovative explanation methods in specific problem settings. Zhao et al. (2019) extended the PDP method to conditional PDP and conditional individual PDP for travel mode switching behavior analysis. The key idea was to group instances into subpopulations first based on some features, and then conduct analysis in each subpopulation. Kim et al. (2020) developed a two-stage framework consisting of a linear regression (LR) part for model interpretability and a long short-term memory (LSTM) part for model accuracy to predict taxi demand. Wang et al. (2020, 2021a, 2021b) tried to explain and extend deep neural networks (DNNs) to analyze travel mode choice. Particularly, Wang et al. (2020) demonstrated that DNNs could provide economic interpretation as complete as classical discrete choice methods (DCMs). Wang et al. (2021a) further substantiated the interpretability of DNN by formulating the function approximation loss to measure interpretation quality. Considering the shared utility interpretation of DCMs and DNNs, Wang et al. (2021b) synergized both models to a unified framework to achieve mutual benefits for travel behavior modeling.

Although some pioneering efforts have been made to disclose the black-box prediction models applied in transportation research, there are still some limitations. First, although some studies aim to provide more comprehensive model explanations by giving both global and local explanations, the two types of explanations are derived from separate methods, and thus inconsistency in explanation might occur. Second, although there are some extensions of current explanation methods, such extensions

are problem-specific, making them hard to be applied to other problems. Third, none of these studies are within the context of maritime transport area, where interpretation and explanation of black-box models are urgently needed to facilitate their successful application in the traditional and conservative shipping industries.

To bridge these gaps, a highly accurate ML based ship risk prediction framework using features from the current ship selection scheme is developed in this chapter. Local explanation and analysis are then given by SHAP. To go one step further, we extend SHAP to a global method in a near linear form by formulating a global surrogate model of the original ML model, and such extension can be applied to arbitrary problem other than the ship selection problem in maritime transport. Contribution of each feature value to the final prediction can be derived from the parameters in the surrogate model similar to a near linear regression model, and thus the black-box prediction model can be considered as explainable as white-box models while its high accuracy can be fully retained.

5.3 DEVELOPMENT OF ML BASED SHIP RISK PREDICTION FRAMEWORK FOR PSC

As the domain knowledge based SRP applied by the Tokyo MoU has several drawbacks which reduce its effectiveness in high-risk ship identification, data-driven ship risk prediction framework based on ML model is developed in this study to achieve efficient ship selection. Data sources and features used for model calibration are first overviewed, and the data-driven framework for ship risk prediction is then introduced, and finally the prediction performance is comprehensively compared and analyzed.

5.3.1 Data

A total of 3,672 initial PSC inspection records at the Hong Kong Port from 1 January 2015 to 31 December 2020 constitute the case dataset of this study. The whole dataset is randomly split into training set (80%, 2,937 samples) and test set (20%, 735 samples). To make the ship risk prediction frameworks developed more consistent with the current ship selection scheme at the Hong Kong Port, i.e., the SRP, and to avoid imposing extra burden of data acquisition and model understanding on the model users, we adopt the same parameters and their encoding method used in the SRP within

the Tokyo MoU as the features to develop the ML model, which is denoted by T-SRP. Detailed parameter decoding method is presented in Table 5-2.

Table 5-2. Feature processing methods in T-SRP

Type	Parameters in SRP	Criteria in SRP	SRP	T-SRP	
			Weighting points	Feature decoding method	Feature type after decoding
1	Ship type	Chemical tanker, gas carrier, oil tanker, bulk carrier, passenger ship, container ship	2	If ship type within the criteria of SRP: 1_ship_type_concerned = 1; else: 1_ship_type_concerned = 0	Binary
2	Ship age	All types with age > 12 years	1	If ship age more than 12: 2_ship_age_12+ = 1; else: 2_ship_age_12+ = 0	Binary
3	Flag performance in Black-Grey-White list of Tokyo MoU	Black	1	If ship flag performance black: 3_flag_black = 1; else: 3_flag_black = 0	Binary
4	RO performance in Tokyo MoU	Low/very low	1	If ship RO performance low or very low: 4_RO_low = 1; else: 4_RO_low = 0	Binary
5	Company performance in Tokyo MoU	Low/very low/no inspection within previous 36 months [unknown]	2	If ship company performance low, very low, or unknown: 5_company_low = 1; else: 5_company_low = 0	Binary
6	Number of deficiencies recorded in each inspection within previous 36 months	How many inspections were there which recorded over 5 deficiencies?	Number of inspections which recorded over 5 deficiencies	The number of inspections with over 5 deficiencies in previous 36 months: 6_deficiency_no_last_36	Integer
7	Number of detentions within previous 36 months	3 or more detentions	1	If involved in 3 or more detentions within previous 36 months: 7_deficiency_last_36 = 1; else: 7_deficiency_last_36 = 0	Binary

5.3.2 Introduction of GBRT

Boosting is one of the most powerful learning methods in the ML community (Friedman et al., 2001). The basic idea of boosting is to develop a procedure that combines the outputs of many less accurate but diverse weak learners in an additive manner to produce a power ensembled model (Friedman et al., 2001). Flexible CARTs are popular weak learners in boosting models. In GBRT for regression tasks, one CART is fit on the negative gradient value of the given loss function in each iteration. Denote a dataset with n samples and m features by $D=\{(\mathbf{x}_i, y_i), i=1,...,n\}, \mathbf{x}_i \in R^m, y_i \in R$, and the prediction of sample $(\mathbf{x}_i, y_i)$ by $f(\mathbf{x}_i)$. If the squared loss in Eq. (5.1) is used as the loss function, least squares is applied,

$$L(y_i, f(\mathbf{x}_i)) = \frac{1}{2}[y_i - f(\mathbf{x}_i)]^2, \tag{5.1}$$

and the negative gradient value of the loss function for sample i is the ordinary residual represented by

$$-g_i = -\frac{\partial L(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)} = y_i - f(\mathbf{x}_i). \tag{5.2}$$

The main hyperparameters of a GBRT model are listed in Table 5-3.

Table 5-3. Main hyperparameters of GBRT

Hyperparameter	Meaning	Value space
n_estimators (K)	The number of iterations (weak learners) constituting a GBRT model	integer, $[1, +\infty)$
learning_rate (ε)	This hyperparameter aiming to shrink the contribution of each tree to the whole ensemble model to reduce overfit	decimal, $(0,1]$
max_depth	The maximum depth of each regression tree	integer, $[1, +\infty)$
min_samples_leaf	The minimum number of training samples required to be at a leaf node	integer, [1, the number of samples]
sub_sample	The fraction of training samples to be randomly selected to construct each regression tree	decimal, $(0,1]$
sub_feature	The fraction of features to be randomly selected to construct each regression tree	decimal, $(0,1]$

The detailed procedure to construct a GBRT model is presented in Procedure 1 (Friedman et al., 2001).

Procedure 1. Construction of a GBRT model	
Input	Training set D; the number of iterations/regression trees K; the loss function L; *max_depth* and *min_samples_leaf* as the stopping criteria of one tree; learning rate ε; *sub_sample*; *sub_feature*
Output	A GBRT model denoted by $f(\mathbf{x})$
Step 1	Initialize $f_0(\mathbf{x}) = \arg\min_{c_0} \sum_{i=1}^{n} L(y_i, c_0)$, where c_0 is the initial predicted target value.
Step 2	for $k = 1, ..., K$: Randomly select $n' = sub\_sample \times n$ training samples and $m' = sub\_feature \times m$ features to construct the kth tree.
Step 2.1	for $i = 1, ..., n'$: Calculate the residual of sample i in iteration k by $r_{ki} = -g_{ki} = y_i - f_{k-1}(\mathbf{x}_i)$. Set r_{ki} as the new prediction target value for sample i by updating the ith sample to $(\mathbf{x}_i, r_{ki})$.
Step 2.2	Use the new training set $D' = \{(\mathbf{x}_i, r_{ki}), i = 1, ..., n'\}$ with m' features to train an ordinary regression tree using the CART algorithm as the kth tree in the ensemble. Especially, all the m' features and their corresponding values should be traversed to select the feature value pair leading to the minimum sum of losses in the left and right child nodes when splitting one node in the tree. The tree grows in a depth-first and recursive manner and stops growing if either of the stop criteria evaluated by *max_depth* and *min_samples_leaf* is reached. Denote the total number of leaf nodes contained in the constructed regression tree by J_k, with one leaf node denoted by $R_{kj}, j = 1, ..., J_k$.
Step 2.3	for $j = 1, ..., J_k$: Calculate the optimal output value of leaf j denoted by c_{kj} by $c_{kj} = \arg\min_{\tilde{c}_{kj}} \sum_{\mathbf{x}_i \in R_{kj}} L(y_i, f_{k-1}(\mathbf{x}_i) + \tilde{c}_{kj})$. Under our problem setting, c_{kj} is the mean of r_{ki} falling in this leaf node.
Step 2.4	Update the current GBRT model to $f_k(\mathbf{x}) = f_{k-1}(\mathbf{x}) + \varepsilon \sum_{j=1}^{J_k} c_{kj} I(\mathbf{x} \in R_{kj})$.
Step 3	The final GBRT model can be expressed by $f(\mathbf{x}) = f_K(\mathbf{x}) = f_{K-1}(\mathbf{x}) + \varepsilon \sum_{j=1}^{J_K} c_{Kj} I(\mathbf{x} \in R_{Kj})$.

In the first step, the optimal c_0 can be obtained by calculating the derivative of $\sum_{i=1}^{n} L(y_i, c_0)$ regarding c_0 and then set it to zero, i.e.

$$\sum_{i=1}^{n} \frac{\partial L(y_i, c_0)}{\partial c_0} = \sum_{i=1}^{n} \frac{\partial [\frac{1}{2}(y_i - c_0)]^2}{\partial c_0} = \sum_{i=1}^{n} (c_0 - y_i) = 0, \tag{5.3}$$

and we can have $c_0 = \frac{\sum_{i=1}^{n} y_i}{n}$, which is the average target value of all the samples. Similarly, c_{kj} is the average target value of all the samples contained in leaf j in the kth iteration. Recall that the target value of sample i in the kth iteration is the residual r_{ki} instead of the original target value y_i.

After the construction of the GBRT models, three typical regression model performance metrics are used to demonstrate model performance: MSE, root mean squared error (RMSE), and MAE. The definitions of the metrics are given as follows.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}[f(\mathbf{x}_i) - y_i]^2, \tag{5.4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[f(\mathbf{x}_i) - y_i]^2}, \tag{5.5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f(\mathbf{x}_i) - y_i|. \tag{5.6}$$

5.3.3 A ship risk prediction framework based on GBRT

A GBRT model is developed for the T-SRP framework for ship risk prediction using the features shown in Table 5-2. The searching spaces of the hyperparameters are given in Table 5-4, and they are tuned based on 5-fold cross-validation on the training set with MSE as the metric.

Table 5-4. Hyperparameter tuning in T-SRP

Hyperparameter	T-SRP	
	Searching space	Value adopted
n_estimators	[200, 1000] with 200 as the interval	200
learning_rate	{0.01,0.02,0.05,0.1,0.2}	0.02
max_depth	[3, 13] with 2 as the interval	9
min_samples_leaf	[1, 9] with 2 as the interval	7
sub_sample	{0.4,0.5,0.6,0.7,0.8}	0.4
sub_feature	{0.3,0.4,0.5,0.6,0.7}	0.3

The framework is finally developed using the hyperparameter values found by hyperparameter tuning on the whole training set, and the model performance is validated on the test set. The MSE, RMSE, and MAE of the T-SRP is 17.9821, 4.2405, and 2.7564, respectively.

5.3.4 Comparison of the new framework and the SRP for ship risk prediction

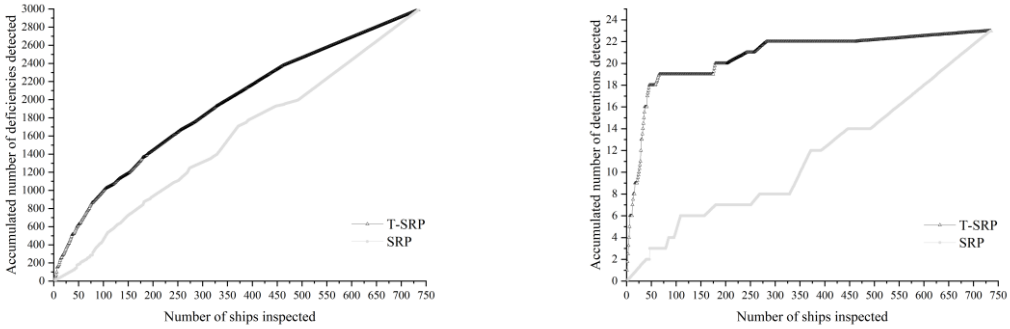
We compare the newly proposed T-SRP framework and the SRP framework under three comparison schemes. In scheme I, ship inspection priority in the SRP is ignored. In other words, the ship inspection sequence is purely dependent on the ship risk scores generated by each framework and the ships are inspected from high risk score to low risk score. In scheme II, ship inspection priority in the SRP is considered. Specifically, ship inspection priority from the highest to the lowest is as follows: ships with no previous inspection (P1), ships with the last inspection time beyond the upper

bound of the time window (where the time window attached to each risk profile is specified in Table 5-1) (P2), ships with the last inspection time within the time window (P3), and ships with the last inspection time below the lower bound of the time window (P4). In comparison scheme I and scheme II, we use the total number of deficiencies and detentions detected after inspecting a certain number of ships as the performance metrics. In scheme III, we first divide the ships in the test set into high-risk, standard-risk, and low-risk types considering their predicted risk scores in T-SRP with the same ratios as those generated by the SRP. Specifically, the number of ships belonging to HRS, SRS, and LRS is 225, 337, and 173 in the test set, respectively. Then, we calculate the average ship deficiency number and detention within each risk type.

The ship risk scores given by the T-SRP are represented by the number of deficiencies predicted by the corresponding GBRT models. The ship risk score given by the SRP is calculated using the risk calculation matrix presented in Table 5-5 (Wang et al., 2019). As there might be ties in ship risk scores, we run each framework in each comparison scheme 1,000 times and use the mean as the result. The performance of each framework in comparison schemes I, II, and III are shown in Figure 5-1, Figure 5-2, and Figure 5-3. The overall comparison of SRP and T-SRP under each comparison scheme is summarized in Table 5-6.

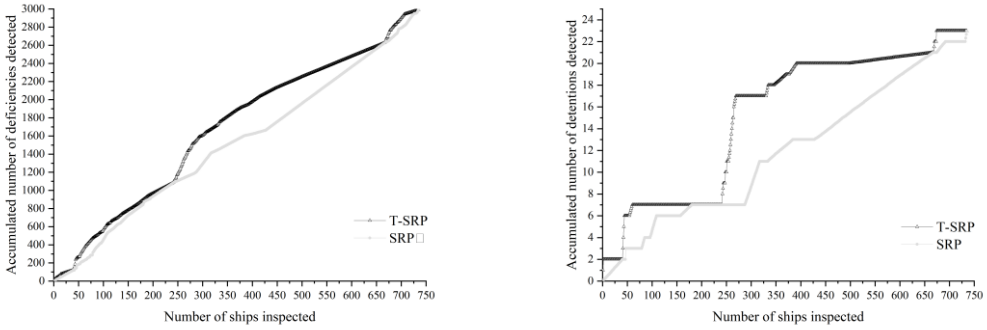
Table 5-5. Calculation of ship risk score in SRP

SRP	Time window (months)	Relationship between the last inspection time (T_l) and the time window		
		T_l beyond the upper bound of the time window	T_l within the time window	T_l below the lower bound of the time window
LRS	9 to 18	$\frac{T_l}{18}$	$\frac{T_l-9}{18-9}$	$\frac{T_l}{9}$
SRS	5 to 8	$\frac{T_l}{8}$	$\frac{T_l-5}{8-5}$	$\frac{T_l}{5}$
HRS	2 to 4	$\frac{T_l}{4}$	$\frac{T_l-2}{4-2}$	$\frac{T_l}{2}$



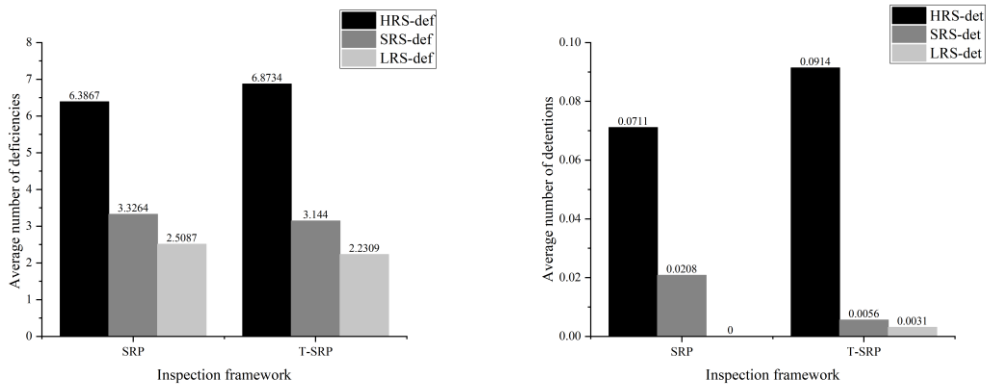
(a) Comparison of deficiency (b) Comparison of detention

Figure 5-1. Comparison results in scheme I



(a) Comparison of deficiency (b) Comparison of detention

Figure 5-2. Comparison results in scheme II



(a) Comparison of deficiency (b) Comparison of detention

Figure 5-3. Comparison results in scheme III

Table 5-6. Summary of the comparison of SRP and T-SRP

Comparison scheme	Improvement regarding the number of deficiencies detected (represented by ratio, a total of 2,992 deficiencies)	Improvement regarding the number of detentions detected (represented by absolute value, a total of 23 detentions)
Scheme I		
T-SRP over SRP	63.71 %	9.36
Scheme II		
T-SRP over SRP	17.48%	3.36
Scheme III	**SRP**	**T-SRP**
Average no. of deficiencies among 'HRS'	6.3867	6.8734
Average no. of deficiencies among 'SRS'	3.3264	3.1440
Average no. of deficiencies among 'LRS'	2.5087	2.2309
Average no. of detentions among 'HRS'	0.0711	0.0914
Average no. of detentions among 'SRS'	0.0208	0.0056
Average no. of detentions among 'LRS'	0	0.0031

It is shown that when the inspection priority in the SRP is considered, the T-SRP is much better than the SRP, with over 60% more deficiencies and over 9 more detentions detected on average given certain inspection resources. When the inspection priority is considered, the superiority of the T-SRP over the SRP is heavily reduced, with over 17% more deficiencies and over 3 more detentions detected.

By comparing Figure 5-1 (a) with Figure 5-2 (a), it can be found that the slope of the line representing the performance of the newly proposed framework in Figure 5-1 (a) gradually reduces as the number of inspected ships increases. This indicates that ships with a larger deficiency number can be distinguished from those with less deficiencies in the new framework. The line of the newly proposed framework in Figure 5-2 (a) is divided into four segments with 40, 241, and 667 inspected ships as the splitting points, which are the thresholds of ship inspection priorities from P1 to P4. The slope gradually decreases in each segment, which also shows that the newly proposed framework is effective within each inspection priority, although its effectiveness is highly compromised when considering such inspection priority. In addition, the SRP is always much better than the SRP ship selection scheme. A similar pattern can be found in Figure 5-1 (b) and Figure 5-2 (b). Among all the 23 detentions, 19 of them can be identified after inspecting 66 of the 735 ships by the T-SRP, which is much more effective than the SRP. In contrast, in scheme II, segmentations of the new framework also exist, and the number of detentions identified increases significantly at the beginning of each segment, as is the case in Figure 5-2 (a).

Finally, Figure 5-3 and Table 5-6 show that the proposed framework is effective in classifying ships into HRS, SRS, and LRS types, as the average numbers of deficiencies and detentions gradually decrease from HRS to LRS. Particularly, the average deficiency number of the HRS identified by the T-SRP is much higher than that identified by the SRP, while the average deficiency number of the LRS identified by the T-SRP is much lower than that identified by the SRP. This indicate that the T-SRP framework is more efficient in identifying high-risk ships. Similarly, the average detention rate of the HRS identified by the T-SRP is much higher than that identified by the SRP, while none of the ships belonging to LRS indicated by the SRP is detained, but the detention rate of the LRS identified by the T-SRP is 0.0021.

5.4 XAI AND ITS IMPORTANCE IN MARITIME TRANSPORT

In addition to developing the highly-efficient ML based ship risk prediction framework for high-risk ship selection in PSC, we further try to explain the predictions given by it from various aspects. In this section, we first clarify the definition of XAI and common approaches to achieve it as well as its benefits. Factors making XAI essential in marine policy making as well as in PSC are then analyzed.

5.4.1 Introduction of XAI

Despite the success of ML models to address real-world problems, the most significant drawback of ML models is their lack of transparency (Du et al., 2020). As a matter of fact, ML models do not explicitly show its internal mechanisms and cannot be understood by looking at their parameters. In addition, the intermediate computation process of the output is opaque. To make the black-box ML models understandable by humans, the area of XAI gained a rapid development in recent years (Doshi-Velez and Kim, 2017). One widely used definition of XAI is given by Arrieta et al. (2020): “Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand.” This definition covers three key points. First, explainability should be presented to ‘a certain audience’, as different audiences pose different requirements for an explainable system due to different background knowledge and communication styles. Second, ‘details’ should answer a ‘how’ question: the more explainable a model is, the more detailed information about its internal structure and working process should be disclosed to an audience. Third, ‘reasons’ should answer a ‘why’ question: the more explainable a

model is, the easier for a human to understand why certain predictions or recommendations have been made.

Using XAI models brings several advantages. For model developers, XAI models can help verify model accuracy and robustness with the assist of domain knowledge. If any irrationality is found, the XAI models can further contribute to model debugging. For model users, XAI models are more likely to be trusted and accepted than black-box models with similar accuracy. Actually, model explainability is even considered as a prerequisite for the adoption of AI systems in high stakes or traditional and conservative domains where reliability, safety, and fairness are required. In addition, explicit decision rules can be extracted from the explanations, and thus shed light on future judgments and decisions for the users.

XAI techniques can be divided into two categories depending on the time when explainability is obtained: one is to develop an interpretable ML model directly, and the other is to use post-hoc explainability techniques after developing a (usually uninterpretable) ML model. Particularly, interpretable ML models are by themselves understandable, such as linear/logistic regression, decision trees, k-nearest neighbors, rule-based learning, general additive models, and Naive Bayes models. In contrast, post-hoc explainability techniques are used to explain the output of an ML model, which are further divided into model-agnostic techniques that can be applied to any ML model disregarding its inner structures and mechanisms, and model-specific techniques that are designed to explain certain ML models considering their internal structure. Popular model-agnostic techniques include PDP, ICE, ALE plot, and SHAP (Molnar, 2020). Particularly, the first two are global methods considering all samples and give a global relationship between a feature and the predicted outcome in one explanation. The others are local methods, where only part of the instances is covered in one explanation. Model-specific techniques have been designed for neural networks and tree-based models.

The major advantage of interpretable ML models is that their explainability is inherent and the prediction and explainability are consistent as both are derived from the ML model directly. However, model accuracy and interpretability need to be balanced. Usually, the higher the prediction accuracy achieved, the lower the model interpretability (Du et al., 2019; Arrieta et al., 2020; Burkart and Huber, 2021). In contrast, post-hoc explanation developed after model construction can help to ease this

problem by using a white-box surrogate model of the black-box prediction model to gain explanation while keeping its high accuracy. Nevertheless, the post-hoc surrogate models might cause inconsistency due to their approximation nature (Du et al., 2019; Babic et al., 2021).

5.4.2 The necessity of XAI to facilitate marine policy making

When black-box ML models are used to assist policy making, a detailed understanding of the prediction model and its output are as important as the prediction accuracy. According to Doshi-Velez and Kim (2017), explainability of black-box model can only be omitted in two situations: 1) no significant consequences will be caused by unacceptable prediction results, and 2) the problem is sufficiently well-studied and the system's decision are trusted even if it is not perfect. Unfortunately, neither condition is satisfied in the context of critical marine policy making. This is mainly because there are several heterogeneous and conservative stakeholders involving and the decisions are heavily dependent on long-term experience while seldom on recommendations given by data-driven models. Consequently, policy recommendations generated by black-box models without convincing explainability provided are seldom accepted, even if they could be much more efficient than recommendations made from naive but transparent rules or expert systems. One example is ship selection models in PSC: although various accurate and efficient ship selection models for substandard ship identification are proposed in several studies, they are rarely adopted by any port authority at the moment. Instead, intuitive and comprehensible ship selection schemes based on domain knowledge are preferred.

In sum, the main reasons for requiring XAI models applied to assist marine policy making are as follows:

a) Trust: conservative practitioners in the traditional maritime industry are reluctant to trust any black-box model to guide policy making. Only when they understand and verify the prediction model's internal schemes, working processes, and strengths and weaknesses, can they trust and thus use the model.

b) Transferability: Only when the policy makers know how well the prediction model generalizes, or in which context it generalizes well, can this prediction model be put in charge of policy making.

c) Fairness: As various stakeholders, such as ship owners, operators, management companies, port authorities, and shipping service providers, are influenced by maritime conventions, fairness is the key to the successful implementation of any critical marine policy. Explanations generated by XAI can help to verify the recommendations given by black-box models to be fair and compliant to ethical standards.

d) Extensibility: On the one hand, XAI enables the developers to improve the prediction model by adjusting its parameters and hyperparameters and by integrating domain knowledge. On the other hand, policy makers can extract new knowledge from massive data by the XAI models and thus to obtain insights for future decision making.

The above-mentioned points are also essential for developing XAI models for ship selection in PSC (Adadi and Berrada, 2018). Similar to the situation discussed by Kleinberg et al. (2015) and Athey (2017), black-box models for ship risk prediction without explanation provided are not enough, as they cannot answer more complex question of why a certain ship should be given a higher inspection priority or what properties would increase ship risk. With the assistance of the tailored explanations given to these black-box models, the above question can be addressed to a large extent, making the models more likely to be adopted in practice and thus a larger number of substandard ships can be inspected by PSC. Therefore, the ports as well as the PSC inspection can better fulfill their responsibility to enhance the maritime safety, to protect the marine environment, and to guarantee decent living and working conditions of seafarers. For ship owners, operators, and managers, they will be more willing to accept explainable ship selection methods as both time and monetary costs can be high if their ships are frequently involved in PSC inspections. Meanwhile, fair ship selection can in turn motive them to keep their ships in satisfactory condition to reduce future inspections. For shipping service providers, they can provide tailored services by considering a ship's PSC inspection results, and thus to reduce maritime risks and pollutions.

5.5 BLACK-BOX MODEL EXPLANATION USING SHAP

Prediction given by the black-box GBRT model in the T-SRP framework is explained from both local and global perspectives in this section based on SHAP. We first introduce the concept of SHAP and give local explanation for the prediction of

individual ships given by the GBRT model. SHAP values in the T-SRP framework are also visualized and analyzed. Then, we go one step further to extend the local SHAP method to a global method by formulating a near linear-form global surrogate model that can closely approximate the output of the GBRT model with full explainability. Validation of the explanation performance of the global surrogate model is finally presented.

5.5.1 Introduction of SHAP

SHAP is proposed by Lundberg and Lee (2017) aiming to explain the output of an individual prediction (i.e., local method) of any ML model (i.e., model agnostic) and it is applied after model construction (i.e., post-hoc). SHAP is based on the Shapley value from coalition game theory first developed by Shapley (1953). It assigns an additive importance value (which can be negative, zero, or positive) to each feature as its contribution to the prediction. Therefore, the prediction is similar to a near linear model by summing the base value, which is the mean of the outputs in the training set denoted by $\bar{y}$, and the contributions of all the features. In the context of XAI, the 'game' refers to the prediction task of a sample, the 'players' are the features included in the model, and the 'gain' is the difference between the actual prediction and the base value.

The basic idea of applying Shapley value to XAI is that the marginal contribution of a single feature concerned is determined by the differences in the outputs of the possible combinations of features with and without this feature. There are several algorithms to calculate SHAP values with different tricks to reduce the computational burden. Here we briefly introduce a basic but easy-to-digest one. To begin with, a power set of features with different feature coalitions ranging from no feature contained to all features contained presented by a tree structure are formulated as shown in Figure 5-4, where each node represents a coalition of features, and each edge indicates adding a feature excluded in the coalition at the head to the coalition at the tail. l is the depth of the tree. Given the dataset in our problem with m features, we can have a total of 2^m coalitions of features, and thus 2^m nodes in the tree.

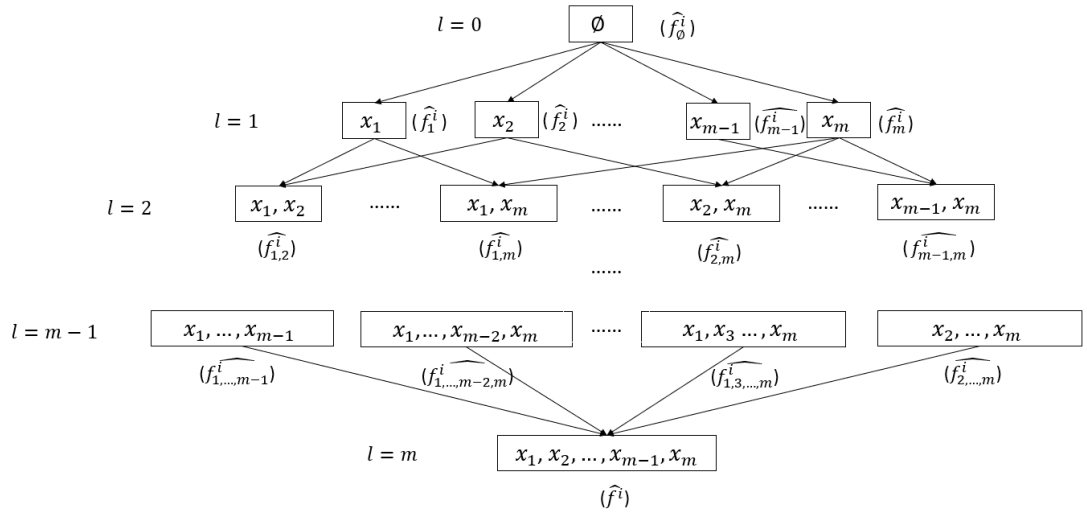


Figure 5-4. An illustration of feature coalitions

Suppose we want to explain the prediction of sample D_i given by the developed ML model. After deciding the feature coalitions, the next step is to decide the predicted target value of D_i given by the ML model using the feature coalition contained in each node. The feature(s) contained in each node is(are) input to the developed ML model, while the absent feature(s) is(are) replaced by a random feature value from the data. The predicted target value of each of the 2^m feature coalitions is presented on the right of or below the corresponding node in Figure 5-4. Particularly, the output of the node containing no feature at the root of the tree is $\hat{f}_{\varnothing}$, which is the average target values in the training set called the base value. As shown in Figure 5-4, the difference between two nodes lies in just one feature. Therefore, the prediction difference between these two nodes connected by an edge can be regarded as the effect, or the marginal contribution, brought by that additional feature (Lundberg and Lee, 2017). For example, if we only consider the first two layers, the marginal contribution of feature x_1 regarding sample D_i can be presented by $\hat{f}_1^i - \hat{f}_{\varnothing}$.

One last question is how to combine the marginal contribution of each feature presented by different node pairs connected by the edges where the feature is not contained in the node at the head but is contained in the node at the tail in Figure 5-4. The weights connecting all the node pairs in consecutive layers l and $l+1$, $l \in [0, m-1]$, are required to be equal and are denoted by $w_{l,l+1}$. For feature x_1, the overall effect of its marginal contribution, which is also called the SHAP value of feature x_1, is denoted by ϕ_1^i and can be calculated by

$$
\begin{aligned}
&\phi_1^i = w_{0,1} \times (\hat{f}_1^i - \hat{f}_\varnothing) + \\
&[w_{1,2} \times (\hat{f}_{1,2}^i - \hat{f}_2^i) + \ldots + w_{1,2} \times (\hat{f}_{1,m}^i - \hat{f}_m^i)] + \\
&[w_{2,3} \times (\hat{f}_{1,2,3}^i - \hat{f}_{2,3}^i) + \ldots + w_{2,3} \times (\hat{f}_{1,m-1,m}^i - \hat{f}_{m-1,m}^i)] + , \\
&\ldots \\
&+[w_{m-1,m} \times (\hat{f}^i - \hat{f}_{2,\ldots,m}^i)]
\end{aligned} \tag{5.7}
$$

where the sum of all the weights is 1. The sum of weights connecting each two consecutive layers is further required to be equal, and thus the weights connecting layer l and $l+1$ is $w_{l,l+1} = [(l+1) \times C_m^{l+1}]^{-1}, l \in [0, m-1]$, where $C_m^{l+1} = \binom{m}{l+1} = \frac{m!}{(l+1)!(m-l-1)!}$. The SHAP value or the feature importance of a feature m', $m' \in [1, m]$ regarding sample i, can therefore be calculated by

$$
\phi_{m'}^i = \sum_{S \subseteq M \setminus \{m'\}} \frac{|S|!(m-|S|-1)!}{m!} [\hat{f}_{S \cup \{m'\}}^i - \hat{f}_S^i], \tag{5.8}
$$

where M is the set of all features. Finally, according to the 'local accuracy' property of SHAP indicated by Lundberg and Lee (2017), summing the Shapley values of all features of sample D_i yields the difference between its predicted output and the base value, where the sum of Shapley values can be regarded as the effects of all the features on the output of this sample. Therefore, the predicted output of sample D_i can also be represented in an additive linear function form as follows:

$$
f(\mathbf{x}_i) = \bar{y} + \sum_{m'=1}^{m} \phi_{m'}^i . \tag{5.9}
$$

The above algorithm for SHAP value calculation is computationally expensive as its needs to predict the targets for a total of 2^m times using different feature coalitions. Fortunately, efficient implementations to calculate SHAP values are proposed by several studies such as Lundberg and Lee (2017) and Lundberg et al. (2019) which can be found from the SHAP API for Python (Lundberg, 2021). The SHAP values are calculated based on the implementation of Lundberg et al. (2019) for tree-based models in this study.

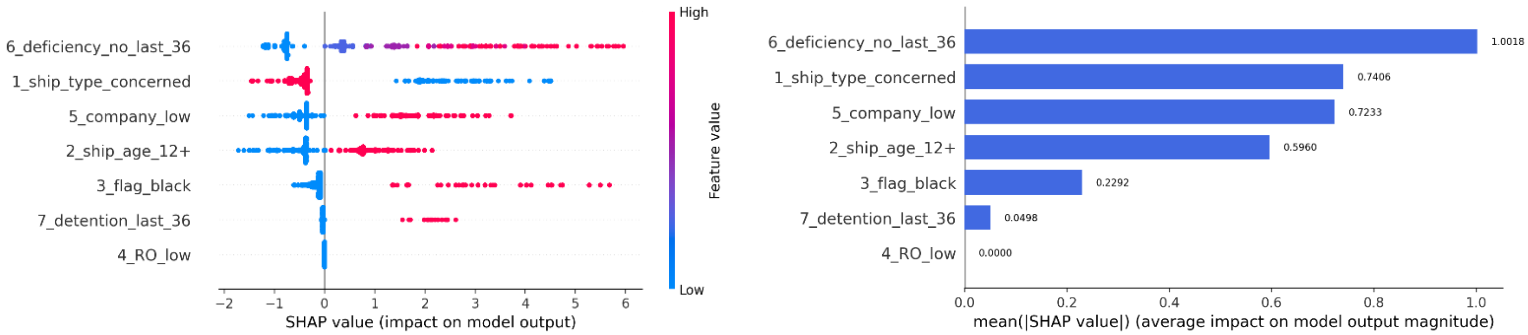
5.5.2 Explanation of GBRT via SHAP

SHAP is originally designed for local explanation, which aims at knowing the reasons for a specific prediction (such as why a particular ship is predicted to have a certain number of deficiencies). In the following subsections, we first give an overview

of local feature effects in the T-SRP framework and the global feature importance derived from the local explanations. Then, we explain specific predictions in the test set.

5.5.2.1 Model explanation based on the training set

Feature SHAP values in the training set and the global feature importance in the T-SRP framework are presented in Figure 5-5.



(a) Local explanation summary (b) Global feature importance

Figure 5-5. Local explanation summary and global feature importance in the T-SRP framework

Figure 5-6 (a) is a set of beeswarm plot with *y*-axis representing each feature and *x*-axis representing the features' SHAP values, while each dot in a figure represents a single ship in the training set. Feature values from low to high are shown by gradient colors as illustrated by the chromatographic on the right side, and the dot's position on the *x*-axis shows the impact that feature value has on the ship's predicted deficiency number given by the GBRT model, i.e., the SHAP value of the feature value for each ship. When multiple dots land at the same *x* position, they pile up to show the density. Figure 5-6 (b) is a bar chart showing the importance of each feature calculated by the mean absolute SHAP values of a feature among all the samples in the training set. The larger a feature's mean absolute SHAP value, the greater influence the feature has on the prediction as it can change the predicted target more.

Figure 5-6 (b) shows that in the T-SRP framework, the only integer variable, i.e., the number of inspections with over 5 deficiencies within previous 36 months, has the highest feature importance. Figure 5-6 (a) indicates that a larger value of this feature leads to a larger predicted deficiency number. Especially, the highest feature values (e.g., more than 20) can increase the final prediction by more than 6. In contrast, if there is no inspection with over 5 deficiencies in previous 36 months, the final

prediction will be reduced by 0 to 2. Among the binary features, whether a ship is of a certain ship type concerned has the largest feature importance, followed by whether a ship has low performance management company. It is interesting to find that if a ship is of a type of concern, less deficiencies will be found; otherwise, much more deficiencies ranging from 1 to 5 will be found. This finding shows that the other features override the feature of ship type when deciding ship risk level.

Moreover, if the performance of a ship's management company is evaluated to be low, very low, or its performance is not listed by the Tokyo MoU, up to 4 more deficiencies can be found compared to the base value. Regarding ship age, it is not surprising to find that if ship age is more than 12, much more deficiencies will be detected; if not, up to 2 less deficiencies will be found compared to the base value. Figure 5-6 also indicates that although features 3_flag_black and 7_detention_last_36 are less important in the T-SRP framework, if a ship's flag is on the black-list or it is detained 3 times or more within the previous 36 months, its predicted number of deficiencies will be increased by 1 to 6 and 1 to 3, respectively. Finally, as there is no ship with low or very low RO performance in the training set, i.e., 4_RO_low is 0 for all the samples, this feature will not influence the prediction results and thus it has zero feature importance.

The explanations based on feature SHAP values in the T-SRP framework indicate that port authorities should pay more attention to ships with worse performance in the last 36 months, especially those with a larger number of deficiencies detected. In addition, older ships, ships of certain types (e.g., bulk carrier, other type, and gas carrier), and ships with worse performance management organizations especially the ISM company should also receive more attention.

5.5.2.2 Model explanation in the test set

This section aims to explore feature contributions in specific samples. The feature values and the SHAP values as well as the prediction results of two samples in the test set are shown in Tables 5-7 and 5-8. Visualization of major features' contribution in the T-SRP is given in Figures 5-6 to 5-7.

Table 5-7. Feature values and the corresponding SHAP values of sample ship 1

Parameters	Ship feature	Feature in T-SRP	Feature value in T-SRP	SHAP value in T-SRP
Ship type	Oil tanker	1_ship_type_concerned	1	–0.511285
Ship age	16	2_ship_age_12+	1	0.685112
Flag performance in Black-Grey-White list of Tokyo MoU	White	3_flag_black	0	–0.114320
RO performance in Tokyo MoU	High	4_RO_low	0	0
Company performance in Tokyo MoU	Medium	5_company_low	0	–0.456442
Number of deficiencies in each inspection within previous 36 months	0 0 8 0 7 3	6_deficiency_no_last_36	2	1.387823
Detention condition in inspections within previous 36 months	No no no no no	7_detention_last_36	0	–0.033159
Real deficiency number		4		
Base value		4.112735		
Sum of feature SHAP values		0.957728		
Predicted deficiency number		5.070463		
Difference between real and predicted deficiency number		–1.070463		

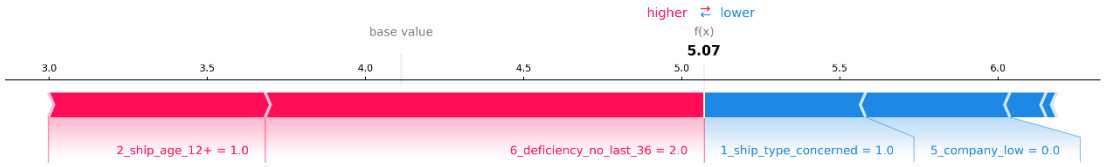


Figure 5-6. Major feature contribution of sample ship 1 in the T-SRP framework

In the T-SRP framework, compared to the base value at 4.11, the increase of the predicted deficiency number is mainly caused by 2 inspections with 5 or more deficiencies in the last 36 months and ship age more than 12 by 2.07, while the final prediction is mainly reduced by being the type of ship concerned and with ship company performance not low, very low, or undefined by 0.97. The sum of all the feature SHAP values is 0.96, and thus the final prediction is 5.07.

Table 5-8. Feature values and the corresponding SHAP values of sample ship 2

Parameters	Ship feature	Feature in T-SRP	Feature value in T-SRP	SHAP value in T-SRP
Ship type	Bulk carrier	1_ship_type_concerned	1	–0.349309
Ship age	6	2_ship_age_12+	0	–0.377828
Flag performance in Black-Grey-White list of Tokyo MoU	White	3_flag_black	0	–0.084827
RO performance in Tokyo MoU	High	4_RO_low	0	0
Company performance in Tokyo MoU	Medium	5_company_low	0	–0.357464
Number of deficiencies in each inspection within previous 36 months	0 0 3	6_deficiency_no_last_36	0	–0.747952
Detention condition in inspections within previous 36 months	no no no	7_detention_last_36	0	–0.028229
Real deficiency number		3		
Base value		4.112735		
Sum of feature SHAP values		–1.945609		
Predicted deficiency number		2.167127		
Difference between real and predicted deficiency number		0.832873		

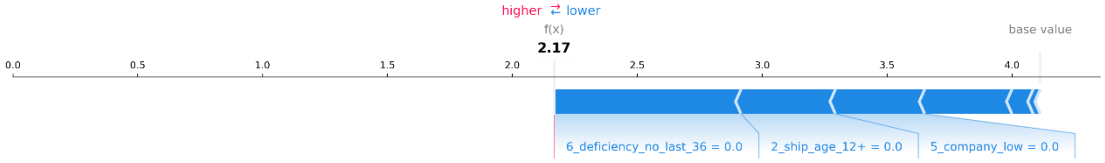


Figure 5-7. Major feature contribution of sample ship 2 in the T-SRP framework

Table 5-8 indicates that in the T-SRP framework, there is no feature with increasing effects on the predicted number of deficiencies compared to the base value. Especially, ship features of no inspection with over 5 deficiencies in previous 36 months, ship age less than 12, and with company performance not low, very low, or undefined (actually medium) contribute the most to the difference between the final prediction and the base value. The total contribution of these features is −1.95, and hence the final predicted deficiency number is 2.17 given the base value 4.11.

Several findings can be drawn after analyzing sample ships 1 and 2. First, explaining the final prediction of a single ship using feature SHAP values makes the decision-making process of the black-box GBRT model transparent. Such explanation makes the new ship selection framework more convinced by the PSCOs. Second, the same feature value can have quite different effects on different samples, and the determinant features of the final prediction are varied among different samples. This is mainly because the features considered interact with each other. As the number of

features increases, such interaction effects become more complex. Third, as T-SRP adopts the same feature encoding method as the current SRP which is in a binary manner, the explanation can be intuitive and understandable. However, such processing simplifies the original features, and thus there will be many samples with the same feature values in the SRP even if the original samples are very different from each other. Consequently, the black-box model's predictive power might be mitigated.

5.5.3 Development of an interpretable global surrogate model based on SHAP: one step further

The above analysis is focused on generating local model explanation, which is the original target of SHAP. To explain the overall performance of the GBRT model from a global perspective, we innovatively extend the local SHAP method to a global method by fitting a near linear-form global surrogate model where the parameters are derived from the SHAP value matrix of the samples in the training set.

5.5.3.1 Main parts of the interpretable global surrogate model

As shown in Table 5-2, the T-SRP framework contains 6 binary features and 1 integer feature. For each binary feature, we calculate the average SHAP values when it takes the value 0 or 1 in the training set as its coefficient in the surrogate model. Specifically, denote a binary feature by $b_{\hat{m}}$, the value of $b_{\hat{m}}$ in sample i is $b_{\hat{m}}^i$ and the corresponding SHAP value is $\phi_{\hat{m}}^i$. The average Shapely value of $b_{\hat{m}}$ when it takes 1 (denoted by $\phi_{\hat{m}\_1}$) and when it takes 0 (denoted by $\phi_{\hat{m}\_0}$) in the whole training set can be calculated by Eq. (5.10) and Eq. (5.11), respectively:

$$\phi_{\hat{m}\_1} = \frac{\sum_{i=1}^{n} \phi_{\hat{m}}^i \times b_{\hat{m}}^i}{\sum_{i=1}^{n} b_{\hat{m}}^i}, \tag{5.10}$$

$$\phi_{\hat{m}\_0} = \frac{\sum_{i=1}^{n} \phi_{\hat{m}}^i \times (1 - b_{\hat{m}}^i)}{\sum_{i=1}^{n} (1 - b_{\hat{m}}^i)}. \tag{5.11}$$

For an integer or continuous feature, we fit its feature values and the corresponding SHAP values using three types of curves: linear curve, quadratic curve, and the mean squared root (sqrt) curve. Specifically, denote an integer feature by $c_{\hat{m}}$

and the SHAP value calculated by the three modes by $\phi_{c_{\hat{m}}}^{\text{linear}}$, $\phi_{c_{\hat{m}}}^{\text{quadratic}}$, and $\phi_{c_{\hat{m}}}^{\text{sqrt}}$ which are presented by Eq. (5.12) to Eq. (5.14):

$$\phi_{c_{\hat{m}}}^{\text{linear}} = a^{\text{linear}} + b^{\text{linear}} \times c_{\hat{m}}, \tag{5.12}$$

$$\phi_{c_{\hat{m}}}^{\text{quadratic}} = a^{\text{quadratic}} + b^{\text{quadratic}} \times c_{\hat{m}} + c^{\text{quadratic}} \times {c_{\hat{m}}}^2, \tag{5.13}$$

$$\phi_{c_{\hat{m}}}^{\text{sqrt}} = a^{\text{sqrt}} + b^{\text{sqrt}} \times \sqrt{c_{\hat{m}}}\,. \tag{5.14}$$

As quadratic curve has the most complex form (three parameters in contrast to two parameters in linear and sqrt modes) and thus is more likely to overfit the data, it will be selected only when its R^2 is higher than that of linear mode and sqrt mode by no less than 0.1. Otherwise, linear or sqrt curve with a higher R^2 will be selected. Finally, the prediction of sample i by the global surrogate model can be presented by

$$\hat{y}_i' = \bar{y} + \sum_{b_{\hat{m}} \in B} [\phi_{\hat{m}\_1} \times b_{\hat{m}}^i + \phi_{\hat{m}\_0} \times (1 - b_{\hat{m}}^i)] + \sum_{c_{\hat{m}} \in C} \sum_{\text{mode} \in \{\text{linear}, \text{quadratic}, \text{sqrt}\}} z_{c_{\hat{m}}}^{\text{mode}} \times \phi_{c_{\hat{m}}^i}^{\text{mode}}, \tag{5.15}$$

where B and C are the set of binary features and the set of integer or continuous features, respectively, $c_{\hat{m}}^i$ is the feature value of $c_{\hat{m}}$ of sample i, $z_{c_{\hat{m}}}^{\text{mode}} \in \{0,1\}$ indicates the fitting mode of feature $c_{\hat{m}}$ and $\sum_{\text{mode} \in \{\text{linear}, \text{quadratic}, \text{sqrt}\}} z_{c_{\hat{m}}}^{\text{mode}} = 1, \forall c_{\hat{m}} \in C$. It should also be mentioned that Eq. (5.15) can easily be extended to contain classification features taking more than 2 values by treating them as continuous or integer values and then fitting the curves of feature values and the corresponding SHAP values. Alternatively, the values can also be treated separately by calculating the average SHAP value of each feature value of the classification feature.

5.5.3.2 Construction of an interpretable global surrogate model for T-SRP

The average binary feature effects of the T-SRP framework are shown in Table 5-9. The relationship between the feature values and the SHAP values of the integer feature 6_deficiency_no_last_36 is shown in Figure 5-8. The fitting curve form and the fitting performance are summarized in Table 5-10.

Table 5-9. Average SHAP values of the binary features in T-SRP

Binary feature	Average SHAP of value 1	Average SHAP of value 0
1_ship_type_concerned (x_1^T)	–0.4454	2.3722
2_ship_age_12+ (x_2^T)	0.8501	–0.4540
3_flag_black (x_3^T)	3.1336	–0.1342
4_RO_low (x_4^T)	0	0
5_company_low (x_5^T)	1.7802	–0.4434
7_detention_last_36 (x_7^T)	2.1108	–0.0307

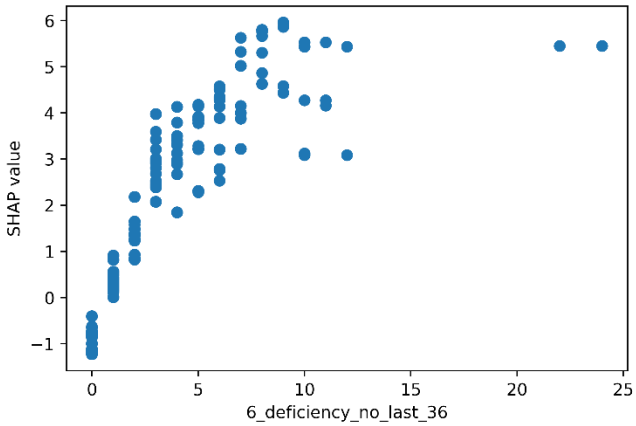


Figure 5-8. Relationship between feature value and SHAP value of feature '6_deficiency_no_last_36'

Table 5-10. Curve fitting performance of feature '6_deficiency_no_last_36'

Integer feature	linear mode	quadratic mode	sqrt mode
6_deficiency_no_last_36 (x_6^T)	$\phi_{x_6^T}^{\text{linear}} = -0.6036 + 0.7462 \times x_6^T$	$\phi_{x_6^T}^{\text{quadratic}} = -0.7553 + 1.0864 \times x_6^T - 0.0403 \times (x_6^T)^2$	$\phi_{x_6^T}^{\text{sqrt}} = -0.8871 + 1.6953 \times \sqrt{x_6^T}$
R^2	0.8383	0.9477	**0.9200**

The sqrt mode is selected to fit the curve of the feature values and their SHAP values of 6_deficiency_no_last_36 in the T-SRP. As the base value of the T-SRP is 4.112735, the near linear form global surrogate model of the T-SRP framework, which is denoted by T-SRP-XAI, can be presented by

$$\hat{y}_i^T = 4.112735 + \left\{ \begin{array}{l} x_1^{T,i} \times (-0.4454) + (1 - x_1^{T,i}) \times 2.3722 + x_2^{T,i} \times 0.8501 + (1 - x_2^{T,i}) \times (-0.4540) + \\ x_3^{T,i} \times 3.1336 + (1 - x_3^{T,i}) \times (-0.1342) + x_5^{T,i} \times 1.7802 + (1 - x_5^{T,i}) \times (-0.4434) + \\ x_7^{T,i} \times 2.1108 + (1 - x_7^{T,i}) \times (-0.0307) + \left[-0.8871 + 1.6953 \times \sqrt{x_6^T} \right] \end{array} \right\}, \quad (5.16)$$

where the term in the curly brackets is the sum of feature effects. Eq. (5.16) is a fully white-box model in a near linear form[6] showing the decision process of the T-SRP

[6] We are fully noted that strictly speaking, $\hat{y}_i^T$ does not take a linear form as it contains a squared root item. Nevertheless, the squared root item can easily be transformed to a linear item by converting all the values of x_6^T into their arithmetic squared root and then feeding to Eq. (5.16).

framework, which is basically consistent with shipping domain knowledge, i.e., older ships, ships with flag on the black-list, low/very low RO performance, low/very low/undefined company performance, larger number of deficiencies and more detentions in recent inspections are more likely to lead to a larger number of deficiencies in the current inspection. This verification has greatly increased the transparency and credibility of the T-SRP framework, and thus makes it more acceptable by shipping practitioners. However, it is also noted that the only difference between the T-SRP and the SRP is that ships of certain types of concern, i.e., chemical tanker, gas carrier, oil tanker, bulk carrier, passenger ship, and container ship are instead with much smaller deficiency number than other types.

Furthermore, the T-SRP-XAI also offers insights into high-risk ship identification from a qualitative perspective. For example, ships with flag on black-list, not of the type concerned, and with no less than 3 detentions in the last 36 months should receive more attention. Finally, it is interesting to find that different values of the same binary feature can have different effects on the final prediction. For example, 0.85 more deficiency will be detected if a ship is more than 12 years old, while 0.45 less deficiency will be detected, otherwise. The absolute difference between the two average SHAP values is 1.3. In contrast, for binary feature such as 3_flag_black, the difference reaches 3.0, indicating that the situations of ships with flag performance not on the black-list is complex and hence their effects can be divergent, that is, ships with flag on the white-list and grey-list can be quite different.

We then apply Eq. (5.16) to predict the deficiency number of the samples in the test set. The MSE, RMSE, and MAE on the test is 18.4831, 4.2992, and 2.7909. Compared to the T-SRP framework, whose MSE, RMSE, and MAE is 17.9821, 4.2405, and 2.7564, the accuracy of the T-SRP-XAI is lower due to the approximation of feature effects. However, the sacrifice of model accuracy results in a globally fully-interpretable model presented in a near linear form, enabling the recommendations given by the black-box GBRT model of the T-SRP framework totally transparent and verifiable. Further experiments show that the MSE and MAE between the prediction of T-SRP and the prediction of T-SRP-XAI are only 0.9262 and 0.6233, respectively.

5.5.4 Comparison of the SRP and the global surrogate model

We compare the performance of the SRP and the T-SRP-XAI regarding the number of deficiencies detected and the ship detentions identified on the test set. The results are summarized in Table 5-11.

Table 5-11. Comparison results of SRP and T-SRP-XAI

Comparison scheme	Improvement regarding the number of deficiencies detected represented by ratio (a total of 2,992 deficiencies)	Improvement regarding the number of detentions detected represented by absolute value (a total of 23 detentions)
Scheme I		
T-SRP-XAI over SRP	65.19%	9.27
Scheme II		
T-SRP-XAI over SRP	17.46%	3.35
Scheme III	**SRP**	**T-SRP-XAI**
Average no. of deficiencies among 'HRS'	6.3867	6.8207
Average no. of deficiencies among 'SRS'	3.3264	3.1610
Average no. of deficiencies among 'LRS'	2.5087	2.2665
Average no. of detentions among 'HRS'	0.0711	0.0911
Average no. of detentions among 'SRS'	0.0208	0.0045
Average no. of detentions among 'LRS'	0	0.0058

Tables 5-6 and 5-11 indicate that the difference in the performance between the near linear-form global surrogate model and the corresponding black-box model regarding the number of deficiencies and detentions detected is minor in Scheme I and Scheme II, even though the original black-box model is more accurate than its global surrogate model. Regarding the deficiency and detention conditions of the ships in three risk levels, results of comparison Scheme III show that T-SRP-XAI can identify ships in 'HRS' more efficiently as evaluated by both deficiency and detention conditions. Based on the above findings, it can be concluded that the near linear-form global surrogate model is almost as efficient as its original black-box model regarding the ability to identify high-risk ships, although its accuracy is slightly worse than the original model. Therefore, it is justifiable to go one step further to extend the local SHAP method to a near linear-form global surrogate model.

5.6 CONCLUSION

To identify high-risk ships more efficiently, this chapter first proposes and validates a ship risk prediction framework based on the state-of-the-art GBRT model, namely T-SRP, by using six years' inspection records at the Hong Kong Port. To make

the new frameworks more comprehensible and acceptable by the port authority part and the ship part, features used and their processing methods in the T-SRP are the same as those in the SRP. In addition, predictions given by the black-box model are thoroughly explained from both local and global perspectives using the post-hoc, model-agnostic, and local SHAP method by explaining the prediction of individual ships, calculating global feature importance scores, and formulating white-box global surrogate models in a near linear form of the original ML model denoted by T-SRP-XAI. The analysis of model explanations is given, and policy implications are drawn from various perspectives.

Comprehensive numerical experiments show that the predictions given by the T-SRP are accurate. When applying it to predict ship risk and identify high-risk ships, more than 60% more deficiencies and nearly 40% more detentions can be detected by the new framework when ignoring ship inspection priority compared to the current SRP. When the inspection priority is considered, nearly 20% more deficiencies and over 10% more detentions can be detected compared to the SRP. The new framework is also more efficient in identifying the type of HRS ship compared to the SRP. Meanwhile, its while-box global surrogate model taking a near linear form follows the PDR model explanation evaluation framework and can provide accurate and comprehensive explanations to decisions makers and practitioners in the shipping industry, so as to enhance their applicability to the conservative maritime transport area.

To the best of the authors' knowledge, this study makes the very first attempt to disclose and explain the working process of the black-box prediction models in the maritime transport research. It also innovatively extends the local SHAP method to a global method by formulating a white-box global surrogate model of the original black-box model. From practical aspect, this study addresses the critical ship selection issues in PSC, which is one of the most important international marine policies. It can help to fulfill IMO's goal of realizing 'safe, secure and efficient shipping on clean oceans'.

Chapter 6: Conclusions and Future Research

6.1 CONCLUSIONS

This thesis has developed several ship risk prediction models from different aspects to facilitate high-risk ship identification and selection in PSC inspection, followed by PSCO assignment and scheduling models to rationalize inspection resource allocation. It comprises four main parts. In the first part, current ship selection methods adopted by different MoUs are summarized, and existing studies on improving PSC efficiency, including ship selection and onboard inspection planning, are reviewed.

In the second part, an XGBoost based ML model is constructed to predict ship deficiency number, where shipping domain knowledge regarding vessel operation conditions is incorporated by modifying model structure and property. The predicted deficiency number is then input to the downstream PSCO scheduling model. To improve model efficiency, the concepts of *inspection template* and *un-dominated inspection template* are proposed, and the optimization model is modified accordingly. Results of numerical experiments show that the XGBoost model's MSE is 12.5 and its MAE is 2.4 on the test set. The combined ship risk prediction and PSCO scheduling model is better than the current inspection procedure by more than 20%, while its gap over the model under perfect-forecast policy is about 8%. The PSCO scheduling model is stable under various conditions as validated by the extensive sensitivity analysis.

In the third part, ship inspection efficiency is improved by matching ship condition with PSCOs' expertise by developing three ML models with different prediction targets and structures as well as two PSCO assignment models. The first two prediction models have normal prediction targets, i.e., the number of deficiencies under each deficiency category and the total number of deficiencies of each ship, and the objective is to minimize the prediction error compared to the real targets. The third prediction model also aims to predict the total number of deficiencies of each ship, while it adopts a loss function motivated by the structure of the optimization problem, i.e., minimizing the MSO in the numbers of deficiencies that can be detected among

the PSCOs for each ship. The two PSCO assignment models are equivalent to each other, and their difference is caused by the different prediction targets as the input. Numerical experiments show that the combination of the third prediction model and the second PSCO assignment model has the best performance, while all the three combined models perform much better than the currently used random PSCO assignment. Several insights are generated through sensitivity analyses.

In the fourth part, a ship risk prediction model using the same features as the current SRP ship selection method based on GBRT called T-SRP is developed. To improved model explainability and transparency, predictions given by the GBRT-based black-box model are thoroughly explained from both local and global perspectives using the post-hoc, model-agnostic, and local SHAP method. The SHAP method is further extended to a global method by formulating a white-box global surrogate model in a near linear form called T-SRP-XAI. The analysis of model explanations is given, and policy implications are drawn from various perspectives. Comprehensive numerical experiments show that the predictions given by the T-SRP are accurate and is much more efficient than the SRP. Meanwhile, its while-box global surrogate model follows the PDR model explanation evaluation framework and can provide accurate and comprehensive explanations to decisions makers and practitioners in the shipping industry, so as to enhance their applicability to the conservative maritime transport area.

This thesis addresses an important practical problem in maritime industry, i.e., improving the efficiency and effectiveness of ship inspection by PSC. The models proposed can help port states to identify high-risk ships more accurately and to assign and schedule the scare inspection resources more efficiently. It can help to fulfill IMO's goal of realizing 'safe, secure and efficient shipping on clean oceans'.

6.2 FUTURE RESEARCH

First, regarding the problem studied, the three studies all aim to predict ship deficiency condition as the risk indicator. Other risk indicators, such as the detention probability and future accident involvement or their combinations can be considered as the prediction target in future research, which could represent the concept of 'ship risk' more properly. Furthermore, how to link ship inspection performance with ship accidents and incidents could be further explored. In addition, apart from improving

the efficiency of PSC from high-risk ship selection and PSCO assignment and scheduling, it can also be improved from the aspects of optimizing onboard inspection sequence. For example, association rules among ship specification and deficiency items can be first mined, and then used to optimize onboard inspection sequence. The efficiency of inspection resource assignment can also be improved by conducting joint PSCO routing and scheduling planning considering more sophisticated scenarios, such as vessel berthing places and periods, the geographical locations of terminals, and the sea and weather conditions along the route, etc.

Second, from the perspective of research data, ship inspection data from only the Hong Kong port are used in this thesis. Inspection data from other authorities, such as other port states in the Tokyo MoU and even those from other MoUs can be used for ship risk prediction, as these inspection results can also provide valuable information on ship risk condition. This can be achieved by constructing an inspection profile of each single ship in the world merchandise fleet, and then predict the inspection performance of a visiting ship based on recommender system. Furthermore, as PSCOs can be different even at the same port, how their differences would influence the inspection results can be explored by comparing the inspection records from different ports and MoUs. Besides, a wider range of data, such as more types of ship specification information, data on ship inspection by flag state, recognized organization, and management company, can be used for ship risk prediction. For model extension, how to apply the prediction models developed in this thesis using the inspection data at the Hong Kong Port to other ports around the world can be further investigated.

Third, from the perspective of research method, more advanced prediction models especially deep learning models can be adopted. As we only use ship inspection records at the Hong Kong Port in recent years where the total number is quite limited (actually, only three to five inspections are conducted for one working day), the limited data prevents us from using highly complex prediction models where the risk of overfitting might be too high. If much more data can be collected, more sophisticated models can then be applied. In addition, more tailored models for ship risk prediction and resource allocation can be developed. For example, a typical situation in ship selection for PSC inspection is that among all the visiting ships to a port on one day, a certain number of them can be inspected considering the available

inspection resources, and the decision to be generated is which set of them should be inspected. To optimize the final decision such that as many ship deficiencies as possible can be found, the following ship selection problem can be incorporated in the development of the ship risk prediction model by e.g., assigning different weights to the trees in an RF model considering the decision quality their prediction results generated, or by tailoring the learning rates of a gradient boosting model considering the downstream decision quality.

Bibliography

Abuja MoU, 2012. Memorandum of understanding on port state control for West and Central African Region, 1999. Accessed 16 Aug 2020. http://www.abujamou.org/post/90.pdf.

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6, 5213852160.

Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila R. Herrera, F., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

Athey, S., 2017. Beyond prediction: Using big data for policy problems. Science 355(6324), 483–485.

Babic, B., Gerke, S., Evgeniou, T., Cohen, I., 2021. Beware explanations from AI in health care. Science 373(6552), 284–286.

Barredo-Arrieta, A., Laña, I., Del Ser, J., 2019. What lies beneath: A note on the explainability of black-box machine learning models for road traffic forecasting. In Proceedings of 2019 IEEE Intelligent Transportation Systems Conference, 2232–2237.

Bateman, S., 2012. Maritime security and port state control in the Indian Ocean Region. Journal of the Indian Ocean Region 8(2), 188–201.

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25(2), 197–227.

Black Sea MoU, 2016. Information sheet of the BS MoU new inspection regime. Accessed 16 Aug 2020. http://www.bsmou.org/downloads/info-sheets/InfoSheetBSMoUNewInspectionRegime.pdf.

Blockeel, H., 1998. Top-down induction of first-order logical decision trees. PhD Thesis of Catholic University of Leuven.

Blockeel, H., De Raedt, L.,1998. Top-down induction of first-order logical decision trees. Artificial Intelligence 101, 285–297.

Breiman, L., 1996. Bagging predictors. Machine Learning 24(2), 123–140.

Breiman, L., 2001. Random forests. Machine learning 45(1), 5–32.

Breiman, L., 2017. Classification and regression trees. Routledge, Abingdon, United Kingdom.

Breiman, L., Friedman, J., Stone C. J., Olshen R. A., 1984. Classification and Regression Trees. Taylor & Francis, Abingdon.

Bukhsh, Z., Saeed, A., Stipanovic, I., Doree, A., 2019. Predictive maintenance using tree-based classification techniques: A case of railway switches. Transportation Research Part C: Emerging Technologies 101, 35–54.

Burkart, N., Huber, M., 2021. A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research 70, 245–317.

Cariou, P., Mejia, M. Q., Wolff, F. C., 2007. An econometric analysis of deficiencies noted in port state control inspections. Maritime Policy & Management 34(3), 243–258.

Chen, T., 2014. Introduction of boosted trees. Accessed 14 July 2020. https://web.njit.edu/~usman/courses/cs675_fall16/BoostedTree.pdf.

Chen, T., 2016. Monotonic Constraints in Tree Construction. Accessed 16 March 2021. https://github.com/dmlc/xgboost/issues/1514.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Chen, X., Zahiri, M., Zhang, S., 2017. Understanding ridesplitting behavior of on-demand ride services: An ensemble learning approach. Transportation Research Part C: Emerging Technologies 76, 51–70.

Chung, W. H., Kao, S. L., Chang, C. M., Yuan, C. C., 2020. Association rule learning to improve deficiency inspection in port state control. Maritime Policy & Management 47(3), 332–351.

Daniels, H., Velikova, M., 2010. Monotone and partially monotone neural networks. IEEE Transactions on Neural Networks 21(6), 906–917.

Degré, T., 2007. The use of risk concept to characterize and select high risk vessels for ship inspections. WMU Journal of Maritime Affairs 6(1), 37–49.

Degré, T., 2008. From black-grey-white detention-based lists of flags to black-grey-white casualty-based lists of categories of vessels? The Journal of Navigation 61(3), 485–497.

Demirović, E., Stuckey, P. J. , Bailey, J. , Chan, J., Leckie, C., Ramamohanarao, K., Guns, T.,2019. An investigation into prediction+ optimisation for the knapsack problem. In Proceedings of International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research, 241–257.

Dinis, D., Teixeira, A. P., Soares, C. G., 2020. Probabilistic approach for characterising the static risk of ships using Bayesian networks. Reliability Engineering & System Safety, 107073.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V., 1996. Support vector regression machines. Advances in Neural Information Processing Systems 9, 155–161.

Du, M., Liu, N., Hu, X., 2019. Techniques for interpretable machine learning. Communications of the ACM 63(1), 68–77.

Duivesteijn, W., Feelders, A., 2008. Nearest neighbour classification with monotonicity constraints. In proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 301–316.

Elmachtoub, A. N., Grigas, P., 2017. Smart "predict, then optimize". Management Science (Forthcoming).

Elmachtoub, A. N., Liang, J. C. N., McNellis, R., 2020. Decision trees for decision-making under the predict-then-optimize framework. arXiv preprint arXiv:2003.00360.

European Commission, 2017. Ex-post evaluation of Directive 2009/16/EC on port state control. Accessed 29 Aug 2020. https://ec.europa.eu/transport/sites/transport/files/2018-final-report-psc-evaluation.pdf.

Freund, Y., Schapire, R. E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 55(1), 119–139.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: A statistical view of boosting. The Annals of Statistics 28(2), 337–407.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning. Springer Publisher, Berlin.

Fu, J., Chen, X., Wu, S., Shi, C., Wu, H., Zhao, J., Xiong, P., 2020. Mining ship

deficiency correlations from historical port state control (PSC) inspection data. PLoS one 15(2), e0229211.
Gao, Z., Lu, G., Liu, M., Cui, M., 2008. A novel risk assessment system for port state control inspection. In Proceedings of 2008 IEEE International Conference on Intelligence and Security Informatics, 242–244.
Graziano, A., Cariou, P., Wolff, F. C., Mejia Jr, M. Q., Schröder–Hinrichs, J. U., 2018a. Port state control inspections in the European Union: Do inspector's number and background matter? Marine Policy 88, 230–241.
Graziano, A., Mejia Jr, M. Q., Schröder–Hinrichs, J. U., 2018b. Achievements and challenges on the implementation of the European Directive on port state control. Transport Policy 72, 97–108.
Graziano, A., Schröder–Hinrichs, J. U., Ölcer, A. I., 2017. After 40 years of regional and coordinated ship safety inspections: destination reached or new point of departure? Ocean Engineering 143, 217–226.
Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Systems with Applications 78, 273–282.
Harrington, P., 2012. Machine learning in action. Manning Publications Co., New York, USA.
Heij, C., Bijwaard, G. E., Knapp, S., 2011. Ship inspection strategies: Effects on maritime safety and environmental protection. Transportation Research Part D 16(1), 42–48.
Heij, C., Knapp, S., 2019. Shipping inspections, detentions, and incidents: An empirical analysis of risk dimensions. Maritime Policy & Management 46(7), 866–883.
Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1), 55–67.
Huisman, D., 2007. A column generation approach for the rail crew re-scheduling problem. European Journal of Operational Research 180(1), 163–173.
IMO, 2017. Resolution A.1119(30): Procedure for port state control, 2017. Accessed 17 May 2019, http://www.imo.org/en/KnowledgeCentre/IndexofIMOResolutions/Assembly/Documents/A.1119%2830%29.pdf.
Janacek, J., Kohani, M., Koniorczyk, M., Marton, P., 2017. Optimization of periodic crew schedules with application of column generation method. Transportation Research Part C: Emerging Technologies 83, 165–178.
Kalatian, A., Farooq, B., 2021. Decoding pedestrian and automated vehicle interactions using immersive virtual reality and interpretable deep learning. Transportation Research Part C: Emerging Technologies 124, 102962.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30, 3146–3154.
Khoda Bakhshi, A., Ahmed, M., 2021. Utilizing black-box visualization tools to interpret non-parametric real-time risk assessment models. Transportmetrica A: Transport Science 17(4), 739–765.
Kim, E., 2021. Analysis of travel mode choice in Seoul using an interpretable machine learning approach. Journal of Advanced Transportation, in press.
Kim, E., Kim, Y., Kim, D., 2021. Interpretable machine-learning models for estimating trip purpose in smart card data. In Proceedings of the Institution of Civil Engineers-Municipal Engineer 174(2), 108–117.
Kim, T., Sharda, S., Zhou, X., Pendyala, R., 2020. A stepwise interpretable machine

learning framework using linear regression (LR) and long short-term memory (LSTM): City-wide demand-side prediction of yellow taxi and for-hire vehicle (FHV) service. Transportation Research Part C: Emerging Technologies 120, 102786.
Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z., 2015. Prediction policy problems. American Economic Review 105(5), 491–95.
Knapp, S., Franses, P. H., 2007. Econometric analysis on the effect of port state control inspections on the probability of casualty: Can targeting of substandard ships for inspections be improved? Marine Policy 31(4), 550–563.
Knapp, S., Heij, C., 2020. Improved strategies for the maritime industry to target vessels for inspection and to select inspection priority areas. Safety 6(2), 1–21.
Knapp, S., Van de Velden, M., 2009. Visualization of differences in treatment of safety inspections across port state control regimes: A case for increased harmonization efforts. Transport Reviews 29(4), 499–514.
Kocev, D., Vens, C., Struyf, J., Dzeroski, S., 2007. Ensembles of multi-objective decision trees. Proceedings of European Conference on Machine Learning 2007, 624–631.
Kulkarni, S., Krishnamoorthy, M., Ranade, A., Ernst, A. T., Patil, R., 2018. A new formulation and a column generation-based heuristic for the multiple depot vehicle scheduling problem. Transportation Research Part B: Methodological 118, 457–487.
Li, K. X., 1999. The safety and quality of open registers and a new approach for classifying risky ships. Transportation Research Part E 35(2), 135–143.
Li, K. X., Zheng, H., 2008. Enforcement of law by the port state control (PSC). Maritime Policy & Management 35(1), 61–71.
Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R news 2(3), 18-22.
Lundberg, S., 2021. API Reference. Accessed 10 May 2021. https://shap-lrjball.readthedocs.io/en/latest/api.html.
Lundberg, S., Lee, S., 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, 4768–4777.
Lundberg, S., Erion, G., Lee, S., 2019. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.
Marine Department, 2021. PSC Information Related to Hong Kong Ships. Accessed 10 August 2021. https://www.mardep.gov.hk/en/faq/pscinfo.html.c
Mediterranean MoU, 2020. Selection of ships for inspection. Accessed 16 Aug 2020. http://197.230.62.214/Basic_Principlse.aspx#3.
Molnar, C., 2020. Interpretable Machine Learning. Accessed 10 May 2021. https://christophm.github.io/interpretable-ml-book/.
Muñoz, G., Espinoza, D., Goycoolea, M., Moreno, E., Queyranne, M., Letelier, O. R., 2018. A study of the Bienstock–Zuckerberg algorithm: Applications in mining and resource constrained project scheduling. Computational Optimization and Applications 69(2), 501–534.
Murdoch, W., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences 116(44), 22071–22080.
Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics 7, 21.
Paris MoU, 2010. Paris memorandum of understanding on port state control. Accessed

3 Dec 2019. https://www.lesiagroup.com/Resources/Paris%20MOU%20NIR.pdf.
Paris MoU, 2014. Annex 7 of Paris memorandum of understanding on port state control. Accessed 20 July 2019. https://www.parismou.org/system/files/Annex%207%20ship%20risk%20profile.pdf.
Parmar, J., Das, P., Dave, S., 2021. A machine learning approach for modelling parking duration in urban land-use. Physica A: Statistical Mechanics and its Applications 572, 125873.
Pazzani, M. J., Mani, S., Shankle, W. R., 2001. Acceptance of rules generated by machine learning among medical experts. Methods of Information in Medicine 40(05), 380–385.
Pei, S., Hu, Q., Chen, C., 2016. Multivariate decision trees with monotonicity constraints. Knowledge-Based Systems 112, 14–25.
Probst, P., Boulesteix, A. L., 2017. To tune or not to tune the number of trees in random forest. The Journal of Machine Learning Research 18(1), 6673–6690.
Probst, P., Wright, M. N., Boulesteix, A. L., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9(3), 1–19.
Ravira, F. J., Piniella, F., 2016. Evaluating the impact of PSC inspectors' professional profile: A case study of the Spanish Maritime Administration. WMU Journal of Maritime Affairs 15(2), 221–236.
Sampson, H., Bloor, M., 2007. When Jack gets out of the box: The problems of regulating a global industry. Sociology 41(3), 551–569.
Şanlıer, Ş., 2020. Analysis of port state control inspection data: The Black Sea Region. Marine Policy 112, 1–11.
Santosa, F., Symes, W. W., 1986. Linear inversion of band-limited reflection seismograms. SIAM Journal on Scientific and Statistical Computing 7(4), 1307–1330.
Shapely, L., 1953. A value for n-person games. Contributions to the theory of games. Annals of Mathematics Studies (2), 307–318.
Sill, J., 1997. Monotonic networks. Advances in Neural Information Processing Systems 10, 661–667.
Tokyo MoU, 2013. Information sheet of the new inspection region (NIR). Accessed 15 November 2018. http://www.tokyo-mou.org/doc/NIR-information%20sheet-r.pdf.
Tokyo MoU, 2014. Information sheet of the new inspection regime (NIR). Accessed 15 Mar 2019. http://www.tokyo-mou.org/doc/NIR-information%20sheet-r.pdf.
Tokyo MoU, 2016. Information on fees and charges by authorities for follow-up PSC inspection. Accessed 17 August 2021. http://www.tokyo-mou.org/doc/INFORMATION%20FOR%20PSC%20INSPECTION%20FEES%20AND%20CHARGES%20BY%20AUTHORITIES-ver16-r.pdf.
Tokyo MoU, 2018a. Memorandum of understanding on port state control in the Asia-Pacific Region. Accessed 19 October 2019. http://www.tokyo-mou.org/.
Tokyo MoU, 2018b. Tokyo MoU celebrates its 25th anniversary during its 29th committee meeting in Hangzhou, China. Accessed 3 June 2020. http://www.tokyo-mou.org/doc/PSCC29%20PRESS-f.pdf.
Tokyo MoU, 2020. Annual report on port state control in the Asia-Pacific Region 2019. Accessed 1 Aug 2020, http://www.tokyo-mou.org/doc/ANN19-f.pdf.
Tsou, M. C., 2019. Big data analysis of port state control ship detention database. Journal of Marine Engineering & Technology 18(3), 113–121.

UNCTAD, 2021. COVID-19 and maritime transport: Impact and responses. Accessed 10 April 2021. https://unctad.org/webflyer/covid-19-and-maritime-transport-impact-and-responses.
Van Den Akker, M., Hoogeveen, H., Van De Velde, S., 2005. Applying column generation to machine scheduling. Column generation. Springer, Boston, MA.
Veran, T., Portier, P., Fouquet, F., 2020. Crash prediction for a French highway network with an XAI-informed Bayesian hierarchical model. In Proceedings of 2020 IEEE International Conference on Big Data,1256–1265.
Wang, S., Mo, B., Zhao, J., 2021b. Theory-based residual neural networks: A synergy of discrete choice models and deep neural networks. Transportation Research Part B: Methodological 146, 333–358.
Wang, S., Wang, Q., Bailey, N., Zhao, J., 2021a. Deep neural networks for choice analysis: A statistical learning theory perspective. Transportation Research Part B: Methodological 148, 60–81.
Wang, S., Wang, Q., Zhao, J., 2020. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. Transportation Research Part C: Emerging Technologies 118, 102701.
Wang, S., Yan, R., Qu, X., 2019. Development of a non-parametric classifier: Effective identification, algorithm, and applications in port state control for maritime transportation. Transportation Research Part B 128, 129–157.
Wu, S., Chen, X., Shi, C., Fu, J., Yan, Y., Wang, S., 2021. Ship detention prediction via feature selection scheme and support vector machine (SVM). Maritime Policy & Management, in press.
Xu, R., Lu, Q., Li, K. X., Li, W., 2007b. Web mining for improving risk assessment in port state control inspection. In Proceedings of 2007 International Conference on Natural Language Processing and Knowledge Engineering, 427–434.
Xu, R., Lu, Q., Li, W., Li, K. X., Zheng, H., 2007a. A risk assessment system for improving port state control inspection. In Proceedings of 2007 International Conference on Machine Learning and Cybernetics, 818–823.
Xu, Y., Yan, X., Liu, X., Zhao, X., 2021. Identifying key factors associated with ridesplitting adoption rate and modeling their nonlinear relationships. Transportation Research Part A: Policy and Practice 144, 170–188.
Yan, R., Wang, S., 2019. Ship inspection by port state control—review of current research. Smart Transportation Systems 2019, 233–241.
Yan, R., Wang, S., Fagerholt, K., 2020. A semi-“smart predict then optimize”(semi-SPO) method for efficient ship inspection. Transportation Research Part B: Methodological 142, 100–125.
Yan, R., Wang, S., Cao, J., Sun, D., 2021a. Shipping domain knowledge informed prediction and optimization in port state control. Transportation Research Part B: Methodological 149, 52–78.
Yan, R., Wang, S., Peng, C., 2021b. An artificial intelligence model considering data imbalance for ship selection in port state control based on detention probabilities. Journal of Computational Science, in press.
Yan, R., Zhuge D., Wang, S., 2021c. Development of two highly-efficient and innovative inspection schemes for PSC inspection. Asia-Pacific Journal of Operational Research, in press.
Yang, Z., Yang, Z., Teixeira, A. P., 2020. Comparative analysis of the impact of new inspection regime on port state control inspection. Transport Policy 92, 65–80.
Yang, Z., Yang, Z., Yin, J., 2018a. Realising advanced risk-based port state control inspection using data-driven Bayesian networks. Transportation Research Part A

110, 38–56.
Yang, Z., Yang, Z., Yin, J., Qu, Z., 2018b. A risk-based game model for rational inspections in port state control. Transportation Research Part E 118, 477–495.
Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies 58, 308–324.
Zhao, X., Yan, X., Van Hentenryck, P., 2019. Modeling heterogeneity in mode-switching behavior under a mobility-on-demand transit system: An interpretable machine learning approach. arXiv preprint arXiv:1902.02904.
Zhao, X., Yan, X., Yu, A., Van Hentenryck, P., 2018. Modeling stated preference for mobility-on-demand transit: A comparison of machine learning and logit models. arXiv preprint arXiv:1811.01315.

Appendices

Appendix A: Supplementary Material for Chapter 3

This appendix presents the current PSCO scheduling strategy applied at the Hong Kong port. The strategy is in a greedy manner: it aims to assign as many ships as possible to one PSCO for inspection on the morning of each workday. The set of ships assigned to one PSCO should satisfy that (a) they are berthing at the port when inspecting. (b) The PSCO can only inspect one ship in a time unit. (c) The lunch break and off work time of the PSCO should be guaranteed. Denote the number of PSCOs on duty for that day by $|P|$. The procedure of the current scheduling strategy is presented in Figure A1.

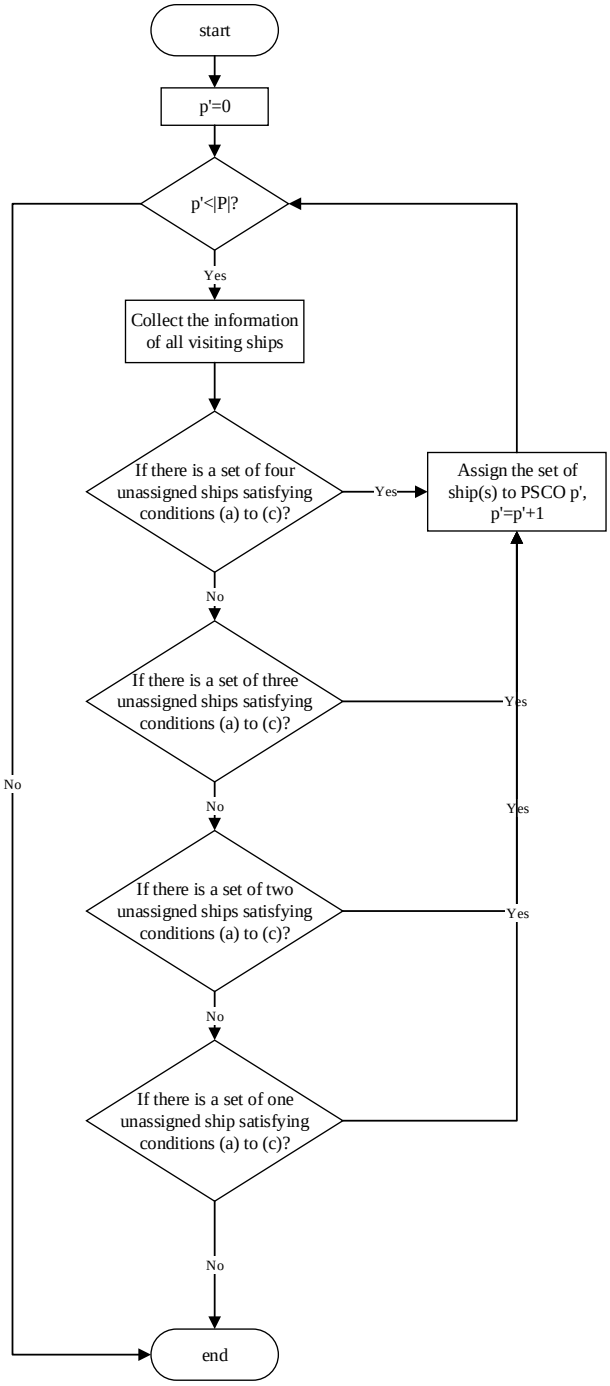


Figure A1. Procedure of current PSCO scheduling strategy

Appendix B: Supplementary Material for Chapter 4

Appendix B1. Procedure of construction of an MTR tree

This appendix presents the procedure of constructing an MTR tree.

Procedure 1: Construction of MTR tree

Input Training dataset D and termination conditions $\Gamma = (\gamma_1, \gamma_2, \gamma_3, ..., \gamma_I)$.

Output MTR tree $f^{MTR}(\mathbf{x})$: for a new example with input features $\mathbf{x}$, the predicted targets are $f^{MTR}(\mathbf{x})$.

Step 0 Construct a root node that contains all the examples in the training dataset (the set of indices for the examples in the root node is denoted by $\{1,...,N\}$). The root is set as the current splitting node.

Step 1 Define R_0 as the set of indices for the examples in the current splitting node. Find the *best split* pair $(j^*, w_{j^*}^*)$ of the current splitting node by enumerating of all possible values of j and w_j.

$$(j^*, w_{j^*}^*) \in \underset{\substack{j\in(x_1,...,x_J)\\ w_j\in\Omega_j}}{\arg\min}\left[\sum_{e\in R_1(j,w_j)}\sum_{k=1}^{K}\left(y^{ek}-\frac{1}{|R_1(j,w_j)|}\sum_{e_1\in R_1(j,w_j)}y^{e_1k}\right)^2+\sum_{e\in R_2(j,w_j)}\sum_{k=1}^{K}\left(y^{ek}-\frac{1}{|R_2(j,w_j)|}\sum_{e_2\in R_2(j,w_j)}y^{e_2k}\right)^2\right]$$

where $R_1(j,w_j)=\{e\in R_0 \mid x^{ej}\le w_j\}$ and $R_2(j,w_j)=\{e\in R_0 \mid x^{ej}> w_j\}$.

Step 2 Use the *best split* $(j^*, w_{j^*}^*)$ to split the current node into two nodes that contain two sub-sets of indices of examples $R_1(j^*, w_{j^*}^*)=\{e\in R_0 \mid x^{ej^*}\le w_{j^*}^*\}$ and $R_2(j^*, w_{j^*}^*)=\{e\in R_0 \mid x^{ej^*}> w_{j^*}^*\}$ with output value for target y_k as $\omega_1^k=\frac{1}{|R_1(j^*,w_{j^*}^*)|}\sum_{e_1\in R_1(j^*,w_{j^*}^*)}y^{e_1k}$ and $\omega_2^k=\frac{1}{|R_2(j^*,w_{j^*}^*)|}\sum_{e_2\in R_2(j^*,w_{j^*}^*)}y^{e_2k}$, respectively, $k=1,...,K$.

Step 3 Repeat *Step* 1 and *Step* 2 in a depth-first manner until coming to a node that reaches one of the preset termination conditions. Then, this node becomes a leaf node and a new node for splitting is found by backtracking.

Step 4 Repeat *Step* 3 until there are no more nodes that can be split. Finally, the total training set is separated into Q mutually exclusive and collectively exhaustive leaf sub-sets $R_1, R_2, ..., R_Q$ according to the input variable vector. The decision tree model can be presented by $f^{MTR}(\mathbf{x})=\sum_{q=1}^{Q}I(\mathbf{x}\in R_q)(\omega_q^1,\omega_q^2,...,\omega_q^K)$, where $I(\mathbf{x}\in R_q)=\begin{cases}1, \mathbf{x}\in R_q\\ 0, \mathbf{x}\notin R_q\end{cases}$.

In Step 1, we have a tree that may not be completed yet, denoted by T, and one of its leaves is the current splitting node. If the current splitting node is split at (j, w_j), we will have a new tree, denoted by T_{j,w_j}, which has two new leaves (the left leave, or called leave 1, and the right leave, or called leave 2) with sets of indices for the examples $R_1(j,w_j)=\{e\in R_0 \mid x^{ej}\le w_j\}$ and $R_2(j,w_j)=\{e\in R_0 \mid x^{ej}> w_j\}$. The predicted value of the kth target for an example e in leave 1 ($e\in R_1(j,w_j)$) is

$\frac{1}{|R_1(j,w_j)|}\sum_{e_1\in R_1(j,w_j)} y^{e_1 k}$, which is the average value of the kth target among all the examples in leave 1. Therefore, it can be seen that Step 1 chooses the best split (j^*, w_{j*}^*) that minimizes the sum of squared error.

Appendix B2. Procedure of construction of an MTR-RF

This appendix presents the procedure of constructing an MTR-RF.

Procedure 2: Construction of MTR-RF

Input	Training dataset D, termination conditions $\Gamma=(\gamma_1,\gamma_2,\gamma_3,...,\gamma_I)$, the number of trees contained in the RF M, and the maximum number of features considered when splitting each node J', $J'<J$.
Output	MTR-RF $f^{MTR-RF}(\mathbf{x})$: for a new example with input feature $\mathbf{x}$, the predicted targets are $f^{MTR-RF}(\mathbf{x})$.
Step 1: For $m=1,...,M$	Draw a bootstrap sample D' of the whole training set D.
Step 1.0	Construct a root node that contains all the examples in D' (the set of indices for the examples in the root node is denoted by $\{1,...,N\}$). The root is set as the current splitting node.
Step 1.1	Among all the J features, randomly select J' features with each selected feature denoted by j'. Define R_0 as the set of indices for the examples in the current splitting node. Find the *best split* pair (j^*, w_{j*}^*) of the current splitting node by solving the following formula: $(j^*, w_{j^*}^*)\in \underset{\substack{j'\in\{x_1,...,x_{J'}\}\\ w'_{j'}\in\Omega_{j'}}}{\arg\min}\left[\sum_{e\in R_1(j',w'_{j'})}\sum_{k=1}^{K}\left(y^{ek}-\frac{1}{|R_1(j',w'_{j'})|}\sum_{e_1\in R_1(j',w'_{j'})}y^{e_1 k}\right)^2+\sum_{e\in R_2(j',w'_{j'})}\sum_{k=1}^{K}\left(y^{ek}-\frac{1}{|R_2(j',w'_{j'})|}\sum_{e_2\in R_2(j',w'_{j'})}y^{e_2 k}\right)^2\right]$ where $R_1(j',w'_{j'})=\{e\in R_0 \mid x^{ej'}\le w'_{j'}\}$ and $R_2(j',w'_{j'})=\{e\in R_0 \mid x^{ej'}> w'_{j'}\}$.
Step 1.2	Use the *best split* $(j^{'*}, w_{j'*}^*)$ to split the current node into two nodes that contain two sub-sets of indices of examples $R_1(j^*, w_{j*}^*)=\{e\in R_0 \mid x^{ej*}\le w_{j*}^*\}$ and $R_2(j^*, w_{j*}^*)=\{e\in R_0 \mid x^{ej*}> w_{j*}^*\}$ with output value for target y_k as $\omega_1^k=\frac{1}{|R_1(j^*,w_{j^*}^*)|}\sum_{e_1\in R_1(j^*,w_{j^*}^*)}y^{e_1 k}$ and $\omega_2^k=\frac{1}{|R_2(j^*,w_{j^*}^*)|}\sum_{e_2\in R_2(j^*,w_{j^*}^*)}y^{e_2 k}$, respectively, $k=1,...,K$.
Step 1.3	Repeat *Step* 1.1 and *Step* 1.2 in a depth-first manner until coming to a node that reaches one of the preset termination conditions. Then, this node becomes a leaf node and a new node for splitting is found by backtracking.

Step 1.4 Repeat *Step* 1.3 until there are no more nodes that can be split. Finally, the total training dataset is separated into Q^m mutually exclusive and collectively exhaustive leaf sub-sets $R_1, R_2, ..., R_{Q^m}$ according to the input variable vector in decision tree m. The mth decision tree model can be presented by $f_m^{MTR}(\mathbf{x}) = \sum_{q=1}^{Q^m} I(\mathbf{x} \in R_q)(\omega_q^1, \omega_q^2, ..., \omega_q^K)$, where $I(\mathbf{x} \in R_q) = \begin{cases} 1, \mathbf{x} \in R_q \\ 0, \mathbf{x} \notin R_q \end{cases}$. For target $k = 1, ..., K$, the final predicted value generated for $\mathbf{x}$ by tree m is represented by $\hat{y}_k^m$ for short.

Step 2*:* For $k = 1, ..., K$, the final predicted value generated by the RF model is the average regarding all the predicted values of the M DTs, i.e. $\hat{y}_k = \frac{1}{M}\sum_{m=1}^{M} \hat{y}_k^m$. The MTR-RF model can be represented by $f^{MTR-RF}(\mathbf{x}) = (\hat{y}_1, ..., \hat{y}_k, ..., \hat{y}_K)$.

Appendix B3. Proof of Proposition 1

This appendix presents the proof of Proposition 1 in Chapter 3.

Proof: If $\Theta > |S|$, we can safely set $\Theta = |S|$ in model [M1] without losing optimality. Therefore, we can assume that $\Theta \leq |S|$. Since P PSCOs can inspect a maximum of ΘP ships, we add $\Theta P - |S|$ dummy ships to the model, each of which has 0 deficiency in each category. Then model [M1] is equivalent to

[M1']

$$\max \sum_{p=1}^{P}\sum_{s=1}^{\Theta P}\sum_{c=1}^{C} z_{ps}\hat{\alpha}^{sc}u_{pc} \tag{B.1}$$

subject to

$$\sum_{s=1}^{\Theta P} z_{ps} = \Theta,\ p = 1, ..., P \tag{B.2}$$

$$\sum_{p=1}^{P} z_{ps} = 1,\ s = 1, ..., \Theta P \tag{B.3}$$

$$z_{ps} \in \{0,1\},\ p = 1, ..., P,\ s = 1, ..., \Theta P \tag{B.4}$$

where parameters $\hat{\alpha}^{sc} = 0$, $s = |S| + 1, ..., \Theta P$, $c = 1, ..., C$.

Define decision vector $\mathrm{z} = (z_{ps},\ p = 1, ..., P,\ s = 1, ..., \Theta P)$, parameter vector $\mathrm{b} = (\underbrace{\Theta, ..., \Theta}_{P \text{ elements}}, \underbrace{1, ..., 1}_{\Theta P \text{ elements}})$, and parameter matrix $A_{(P+\Theta P)\times\Theta P^2}$ that represents the coefficients in

constraints (B.2) and (B.3). Defining $\mathbb{Z}$ as the set of integers, model [M1'] can be written as

[M1'']

$$\max \sum_{p=1}^{P} \sum_{s=1}^{\Theta P} \sum_{c=1}^{C} z_{ps} \hat{\alpha}^{sc} u_{pc} \tag{B.5}$$

subject to

$$Az = b \tag{B.6}$$

$$0 \le z \le 1 \tag{B.7}$$

$$z \in \mathbb{Z}^{\Theta P^2}. \tag{B.8}$$

We can see that (i) all of the elements in b are integers, (ii) all of the elements in A are 0 or 1, (iii) each column of matrix A has exactly two elements whose values are 1, and (iv) matrix A can be divided into two sub-matrices: the top P rows constitute one matrix and the bottom ΘP rows constitute the other matrix, such that each sub-matrix has exactly one element of 1 in each column. Consequently, A is totally unimodular and all the extreme points are optimal solutions to the linear programming relaxation of model [M1''] are integral. Hence, the integrality constraint in Eq. (B.8) can be dropped. In other words, model [M1''] can be easily solved as a linear programming problem.

Note that the conversion of model [M1] to model [M1''] is polynomial because $\Theta \le |S|$. Since a linear program can be solved in polynomial time of the length of its input parameters, model [M1] can also be solved in polynomial time of the length of its input parameters. □

Appendix B4. Procedure of hyperparameter tuning using revised grid search

This appendix presents the procedure of hyperparameter tuning using the revised grid search method.

Denote the set of hyperparameters (i.e., max_features, max_depth, and min_samples_leaf) to be tuned as $K = \{\kappa_1, \kappa_2, \kappa_3\}$ and one hyperparameter is denoted by κ_i. The minimum and maximum values each hyperparameter can take are denoted by $\kappa_i \in [\kappa_i^{\min}, \kappa_i^{\max}]$, $\kappa_i \in K$. The initial constrained value spaces are denoted by

$R_i = \{\kappa_i^{\min}, \lfloor (\kappa_i^{\min} + \kappa_i^{\max})/2 \rfloor, \kappa_i^{\max}\}$, $\kappa_i \in \mathrm{K}$. The procedure to tune the hyperparameters by revised grid search is as follows:

Procedure 3: Tuning hyperparameters by revised grid search	
Input	The set of hyperparameters to be tuned $\mathrm{K} = \{\kappa_1, \kappa_2, \kappa_3\}$, the minimum and maximum values each hyperparameter can take $\kappa_i \in [\kappa_i^{\min}, \kappa_i^{\max}]$, $\kappa_i \in \mathrm{K}$, the initial constrained value spaces $R_i = \{\kappa_i^{\min}, \lfloor (\kappa_i^{\min} + \kappa_i^{\max})/2 \rfloor, \kappa_i^{\max}\}$ for all $\kappa_i \in \mathrm{K}$.
Output	Hyperparameter value tuple with the best performance on validation set denoted by ψ^*.
Step 1	Set the hyperparameter grid Ψ to $\Psi = R_1 \times R_2 \times R_3$. for each $\psi \in \Psi$: Train MTR-RF model $f_{\psi}^{MTR-RF}(\mathbf{x})$ using the training set and hyperparameter tuple ψ. Measure its performance by calculating the MSE/MSO score m_{ψ} on the validation set.

Step 2	The hyperparameter tuple that yields minimum MSE/MSO score m_{ψ}^* on the validation set is denoted by ψ^*, $\psi^* = \{\kappa_1^*, \kappa_2^*, \kappa_3^*\}$.
	if $\kappa_i^{\min} + 2 \geq \kappa_i^{\max}$ for all $\kappa_i \in \mathrm{K}$:
	Return the optimal hyperparameter tuple $\psi^* = \{\kappa_1^*, \kappa_2^*, \kappa_3^*\}$ and terminate the program.
	else:
	for $\kappa_i \in \mathrm{K}$ with $\kappa_i^{\min} + 2 < \kappa_i^{\max}$:
	if $\kappa_i^* = \kappa_i^{\min}$:
	set $\kappa_i^{\max} = \lfloor (\kappa_i^{\min} + \kappa_i^{\max})/2 \rfloor - 1$, update $R_i = \{\kappa_i^{\min}, \lfloor (\kappa_i^{\min} + \kappa_i^{\max})/2 \rfloor, \kappa_i^{\max}\}$.
	else if $\kappa_i^* = \kappa_i^{\max}$:
	set $\kappa_i^{\min} = \lfloor (\kappa_i^{\min} + \kappa_i^{\max})/2 \rfloor + 1$, update $R_i = \{\kappa_i^{\min}, \lfloor (\kappa_i^{\min} + \kappa_i^{\max})/2 \rfloor, \kappa_i^{\max}\}$.
	else:
	set $\kappa_i^{\min} = \lfloor (\kappa_i^{\min} + \kappa_i^*)/2 \rfloor$ and $\kappa_i^{\max} = \lfloor (\kappa_i^* + \kappa_i^{\max})/2 \rfloor$, update $R_i = \{\kappa_i^{\min}, \lfloor (\kappa_i^{\min} + \kappa_i^{\max})/2 \rfloor, \kappa_i^{\max}\}$.

Step 3	Repeat *Step* 1 and *Step* 2 until termination.

Appendix B5. An illustration of the superiority of A3

We use a randomly selected group of ships in the numerical experiment to illustrate the insights of the superiority of A3. The real inspection expertise of each PSCO to each ship (denoted by a ship-PSCO pair) in the selected group and the best PSC assignment in theory are presented in Table B5.1. For simplicity, we only

compare the performance of A2 (and MTR-RF2) and A3 (and MTR-RF3). The predicted inspection expertise of ship-PSCO pairs generated by MTR-RF2 and MTR-RF3 and the corresponding PSCO assignment are shown in Table B5.2 and Table B5.3.

Table B5.1. Real inspection expertise and best PSCO assignment

PSCO/Ship	PSCO 1	PSCO 2	PSCO 3	PSCO 4	Best PSCO assignment
1	3.2	2.3	3.7	3.3	3
2	0.6	0.5	0.7	0.7	4
3	2.4	2.3	2.8	2.7	4
4	9.3	10.7	9.3	7.6	2
5	5.0	4.6	4.9	3.6	1
6	4.9	3.9	5.2	3.9	3
7	28.3	34.1	31.9	31.1	2
8	0.7	0.4	0.8	0.6	1
9	17.3	17.5	18.8	16.3	3
10	5.8	7.8	6.4	6.4	2
Total inspection expertise					89.4

Table B5.2. Predicted inspection expertise and PSCO assignment of A2

PSCO/Ship	PSCO 1	PSCO 2	PSCO 3	PSCO 4	Assigned PSCO
1	2.71151	2.71333	2.83366	2.37304	2
2	2.19643	2.16080	2.27956	1.88287	1
3	1.42442	1.41698	1.47454	1.21828	4
4	3.16357	3.19042	3.25073	2.66992	2
5	2.94226	2.93073	3.06017	2.54735	1
6	3.46248	3.39896	3.59726	2.95980	3
7	21.94541	22.66446	23.59041	20.79491	3
8	3.30471	3.29913	3.40618	2.80313	1
9	6.76287	6.80433	6.97162	5.77351	3
10	4.58404	4.59511	4.72913	3.90545	2
Total achieved inspection expertise					85.7

Table B5.3. Predicted inspection expertise and PSCO assignment of A3

PSCO/Ship	PSCO 1	PSCO 2	PSCO 3	PSCO 4	Assigned PSCO
1	2.28501	2.31229	2.38390	2.00269	2
2	2.07840	2.06494	2.16828	1.81122	1
3	1.73917	1.73968	1.81308	1.51692	4
4	3.23858	3.27633	3.33146	2.74851	2
5	2.63255	2.61521	2.71647	2.23666	1
6	3.28039	3.24911	3.40201	2.80859	3
7	15.68360	17.17298	16.95479	15.32269	2
8	3.65548	3.67663	3.77489	3.12439	1
9	7.47597	7.52486	7.73092	6.42257	3
10	4.61062	4.56776	4.74119	3.88015	3
Total achieved inspection expertise					86.5

Tables B5.2 and B5.3 show that the performance of A3 is better than A2 by 0.8 inspection expertise, while both A2 and A3 can achieve 95% of the total real inspection expertise. The main differences between A2 and A3 are that PSCO 3 is assigned to

inspect ship 7 in A2 while PSCO 2 is assigned to inspect ship 7 in A3, whereas PSCO 2 is assigned to inspect ship 10 in A2 while PSCO 3 is assigned to inspect ship 10 in A3. Notably, assigning PSCO 2 to inspect ship 7 and PSCO 3 to inspect ship 10 could obtain more inspection expertise, as the difference between assigning PSCO 2 and PSCO 3 to ship 7 is 2.2 while the difference is 1.4 to ship 10. We further compare the squared error of the predicted inspection expertise of each ship-PSCO pair and the MSE score in MTR-RF2 and MTR-RF3 are shown Table B5.4 and Table B5.5. The squared overestimate of the predicted inspection expertise of each ship-PSCO pair and the MSO score in MTR-RF2 and MTR-RF3 are shown in Table B5.6 and Table B5.7.

Table B5.4. Squared error of MTR-RF2

PSCO/Ship	PSCO 1	PSCO 2	PSCO 3	PSCO 4
1	0.23862	0.17084	0.75054	0.85925
2	2.54859	2.75824	2.49501	1.39917
3	0.95175	0.77972	1.75684	2.19551
4	37.65583	56.39386	36.59365	24.30566
5	4.23430	2.78646	3.38496	1.10807
6	2.06646	0.25104	2.56878	0.88398
7	40.38078	130.77153	69.04925	106.19480
8	6.78453	8.40497	6.79218	4.85378
9	111.03115	114.39728	139.91054	110.80693
10	1.47855	10.27133	2.79180	6.22278
MSE (of each pair)				26.4820

Table B5.5. Squared error of MTR-RF3

PSCO/Ship	PSCO 1	PSCO 2	PSCO 3	PSCO 4
1	0.83721	0.00015	1.73213	1.68302
2	2.18567	2.44902	2.15584	1.23480
3	0.43670	0.31396	0.97402	1.39968
4	36.74080	55.11084	35.62344	23.53698
5	5.60484	3.93941	4.76782	1.85870
6	2.62314	0.42366	3.23279	1.19117
7	159.17366	286.52394	223.35935	248.92341
8	8.73485	10.73632	8.84997	6.37257
9	96.51167	99.50336	122.52453	97.56362
10	1.41462	10.44739	2.75165	6.34964
MSE (of each pair)				39.4949

Table B5.6. Squared overestimate of MTR-RF2

PSCO/Ship	PSCO 1& PSCO 2	PSCO 1& PSCO 3	PSCO 1& PSCO 4	PSCO 2& PSCO 3	PSCO 2& PSCO 4	PSCO 3& PSCO 4
1	0.81327	0.14277	0.19226	1.63754	1.79637	0.00367
2	0.00414	0.00028	0.17104	0.00660	0.22842	0.15737
3	0.00857	0.12242	0.25619	0.19575	0.35845	0.02442
4	1.88554	0.00760	1.45530	2.13252	6.65386	1.25259
5	0.15091	0.04749	1.01021	0.02909	0.38022	0.61965
6	0.87699	0.02730	0.24733	1.21374	0.19286	0.43896

7	25.81606	3.82203	15.60644	9.77157	1.27792	3.98201
8	0.08668	0.00000	0.16127	0.08582	0.48442	0.16245
9	0.02513	1.66732	0.00011	1.28304	0.02862	1.69492
10	3.95587	0.20695	1.63481	2.35323	0.50458	0.67845
MSO (of each ship)						10.0031

Table B5.7. Squared overestimate of MTR-RF3

PSCO/Ship	PSCO 1& PSCO 2	PSCO 1& PSCO 3	PSCO 1& PSCO 4	PSCO 2& PSCO 3	PSCO 2& PSCO 4	PSCO 3& PSCO 4
1	0.85986	0.16089	0.14617	1.76463	1.71506	0.00035
2	0.00749	0.00010	0.13482	0.00934	0.20586	0.12749
3	0.01010	0.10634	0.27274	0.18199	0.38783	0.03848
4	1.85572	0.00863	1.46392	2.11740	6.61609	1.24779
5	0.14643	0.03383	1.00824	0.03950	0.38620	0.67272
6	0.93843	0.03182	0.27900	1.31585	0.19406	0.49927
7	18.58139	5.42334	9.99131	3.92756	1.32184	0.69238
8	0.10314	0.00038	0.18583	0.09105	0.56586	0.20295
9	0.02283	1.55014	0.00285	1.19671	0.00955	1.42003
10	4.17329	0.22037	1.77015	2.47569	0.50750	0.74139
MSO (of each ship)						8.0162

Tables B5.4 and B5.5 shows that the MSE of the outputs of MTR-RF3 is much larger than that of MTR-RF2, while Tables B5.6 and B5.7 show that the MSO of the outputs of MTR-RF3 is smaller than that of MTR-RF2. Especially, for ship 7, the MSE is 86.60 for the outputs generated by MTR-RF2 and 229.50 for the outputs generated by MTR-RF3. On the contrary, the MSO of ship 7 is 60.28 in MTR-RF2 and the MSO of ship 7 is 39.94 in MTR-RF3. The differences in the MSE and MSO of MTR-RF2 and MTR-RF3 regarding ship 7 indicate that although MTR-RF3 is less accurate in the prediction values compared to MTR-RF2, it could better predict the "relative relationship" among the four outputs. More specifically, we compare the relative relationship of the outputs in the real situation and the predicted values generated by MTR-RF2 and MTR-RF3 for ship 7 as shown in Figure B5.1.

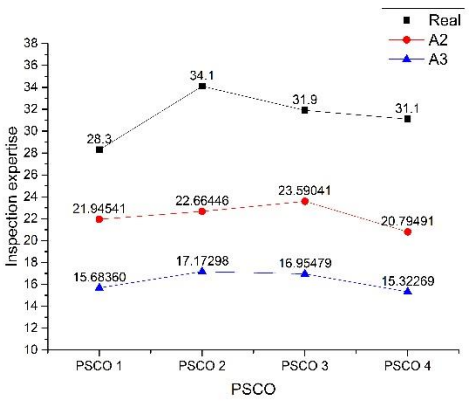


Figure B5.1. Comparison of the predicted inspection expertise of MTR-RF2 and MTR-RF3

Figure B5.1 shows that the relative relationship of the predicted inspection expertise of MTR-RF3 and the real situation is quite the same: PSCO 2 has the largest expertise, following by PSCO 3. Although the predicted relative inspection expertise of PSCO 4 and PSCO 1 is swapped in MTR-RF3, the gap is quite small, and is much smaller than that of MTR-RF2. However, the prediction results of MTR-RF2 suggests that PSCO 3 has the largest inspection expertise followed by PSCO 2, where there is a big gap with the real situation. Therefore, A2 assigns PSCO 3 to inspect ship 7, and A3 assigns PSCO 2 to inspect ship 7 which is the same as the optimal assignment in the real situation.