Marine Policy 60 (2015) 162–170

Contents lists available at ScienceDirect

Marine Policy

journal homepage: www.elsevier.com/locate/marpol

A method to measure enforcement effort in shipping with incomplete information

Xichen Ji [a], Jan Brinkhuis [b], Sabine Knapp [b,*]

[a] Gibbs Quantitative Research & Consulting, Westersingel 6, 3014 GM Rotterdam, The Netherlands
[b] Econometric Institute, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 8 December 2014
Received in revised form
11 June 2015
Accepted 12 June 2015
Available online 25 June 2015

Keywords:
Performance measurement
Small sample sizes
Inspections
Detentions
Incidents
Measuring effort under incomplete information
Accounting for bad luck
Incentives

ABSTRACT

Current methods in the shipping industry to evaluate performance do not account for differences in fleet profiles of registries such as age, size or ship type and not for bad luck. This can lead to unfair evaluation of enforcement efforts of the international standards. Furthermore, incentives to improve performance are concentrated on decreasing detentions rather than incidents. This article proposes a new method to a longstanding problem to evaluate performance that rectifies shortcomings of the method currently used. The proposed method measures the enforcement effort by means of proxy variables and introduces incentives for improvement that go beyond the currently used 'detention'. The aim is to provide a fair and transparent way. The proposed method is applied and results are compared with methods currently used to demonstrate how the rankings change. The method can be adapted to other areas of the shipping industry such as classification societies or ship management companies.

1. Introduction

The shipping industry is characterized by a complex legislative framework of over 50 conventions of the International Maritime Organization (IMO), which lacks enforcement powers due to its international nature. Since enforcement at the flag state level is not directly monitored, port states have created port state control regimes (PSC) that enforce internationally agreed standards on vessels entering their territory, by exercising their right to perform PSC inspections. If a vessel is found to be not compliant, it can be detained. Two PSC regimes (the Paris MoU and the Tokyo MoU) publish each year a list of flags according to their performance during inspections, the so-called Black/Gray/White List (BGW-list), where black listed flags perform worst. The Paris MoU covers the European Union, parts of Canada and the Russian Federation while the Tokyo MoU covers Asia, Australia, Chile and parts of the Russian Federation. The list has become the industry standard and is often interpreted as a rank of list according to flag performance.

It has been a longstanding problem to find a better method to measure the performance of registries. Perepelkin et al. [1] have proposed a method that deals with some of the shortcomings of the current method, giving a common criterion. Indeed, the criterion used at the moment is defined in terms of the excess factor. The value of which depends on the BGW-list and for each of the three, black/gray/white, it is defined by a different procedure (Perepelkin et al. [1]). Perepelkin et al. [1] have considered incident data and deficiencies besides the current standard of using detention data.

Given this situation, this article builds on some aspects of the method developed by Perepelkin et al. [1], and in particular it tries to address the lack of any common criterion that depicts the effort of a flag. The proposed method introduces the concept of the '*enforcement effort*' which cannot be directly observed. The number of undesirable events is counted that are the result of insufficient effort such as weighted numbers of detainments, very serious accidents and serious accidents. The outcome is taken as a proxy for the effort of a registry. The method can be extended in the future to include other quantities that can measure enforcement effort or implementation effort. Data from the Member States audit scheme of the International Maritime Organization might perhaps be useful to integrate in the future.

In principle, other factors might also be relevant, such as the age of the vessel and the sizes or the ship type, as these have an

* Corresponding author. Alternative address: A Schurichtstraße 4, A-2380 Perchtoldsdorf, Austria.
E-mail addresses: xichen.ji@gibbsqrc.com (X. Ji), brinkhuis@ese.eur.nl (J. Brinkhuis), knapp@ese.eur.nl, sknapp@tpg.com.au (S. Knapp).

http://dx.doi.org/10.1016/j.marpol.2015.06.015
0308-597X/

influence on the safety quality of ships (Bijwaard and Knapp [2]). The reason for this is that the major shipping markets have different characteristics. These differences are due to the varying commercial conditions of the shipping markets and are best reflected by ship types. Ship types are not considered in methods currently used to measure performance. Moreover, it is also difficult to evaluate a registry with a small fleet fairly by means of currently used methods. One reason for this is that for small fleet, the performance is more prone to bad luck. Therefore, the concept of *'sympathy'* is introduced into the measure, giving each flag the benefit of the doubt, but not more. Registries with smaller fleets get more sympathy, as desired.

The proposed method also addresses the lack of use of combined data sources (Knapp [3], Knapp and Franses [4], Bijwaard and Knapp [2]). Currently, the PSC authorities believe that their target is different from the target to avoid incident and only PSC data is used. The authors of this article do not necessarily agree with this perception and believe that enforcement influences the safety quality which then influences incident type risk. In addition, earlier research (Knapp [3], Heij and Knapp [5]) confirms that substandard ships that are targeted for inspections benefit from an inspection and the probability of casualty is therefore decreased for certain time period after the inspection (Bijwaard and Knapp [2]). In order to measure enforcements, it is therefore more realistic to use combined data to better capture the level of noncompliance and to reduce biases due to targeting of ships for detentions only. This work builds on the idea of Perepelkin et al. [1] and extends the use of incident data to include two degrees of seriousness – very serious and serious incidents as proxy to safety quality in addition to detention. A measurement including both datasets provides a more balanced approach by taking the level of noncompliance into account. Moreover, taking also incident data into account provides an additional incentive to flags to decrease incidents, not just detentions or deficiencies.

Underreporting in the Global Integrated Ship Information System (GISIS) does not occur for very serious incidents but it can occur for serious incidents. Taking serious incidents into account could therefore provide an undesirable incentive for underreporting by flags and ranks using the proposed method are therefore compared either with or without serious incidents. It turns out that this hardly influences the ranking of a registry. That is, the new methodology turns out to be robust to underreporting of serious incidents. However to capture – as comprehensive as available data allow – the enforcement effort, it is better to include serious incidents. Moreover, it is expected that the possibility of underreporting serious incidents will reduce in the future. Other data such as data from the IMO Member State Audit could be included in the future but was not available for this project.

The challenge is to find a method to measure enforcement effort of international standards, providing leniency to registries that have smaller fleets or that have more challenging fleet profiles. The second challenge is to provide fair incentives to improve. It is reasonable to expect that in case of sufficient effort by a flag, for the ships under this flag, certain undesirable events will be rare. For example, inspections of ships will rarely lead to detention, and very serious incidents will be rare. This suggests to count some well-chosen types of undesirable events, detentions and very serious accidents, and to use the outcome as a performance measurement that is proxy for the effort: a low respectively high outcome is interpreted as a good respectively inadequate effort by the flag.

The new methodology is kept on purpose simple in order to make acceptance more feasible for policy makers. It is refined enough to be realistic and to overcome the challenges mentioned above. The method is developed free of arbitrary elements and emphasis is placed on the main underlying idea – that is to measure the 'enforcement effort' of each flag based on inspection and incident data. Major challenge is to take into account that, given a fixed level of effort and given that ships can be selected randomly for inspection, the probability for detention for one inspected ship is larger for some flags than for others. For example, it is larger if the fleet of a flag consists of old vessels or if this fleet is small (a flag is considered small when it is small in world fleet statistics and not in terms of PSC inspections). The new methodology takes this into account in a satisfactory way without the need for making arbitrary choices; everything flows naturally from general principles.

The main benefit of this proposal is that it is exactly the same for all flags and is based on a simple intuitive idea: subtract from the weighted number of undesirable events the maximum number of events that could be attributed to bad luck (in reason). Technical derivations (given in Appendix A) lead to an explicit formula for this maximum number. A further technical analysis then shows that the 'sympathy' is more for small flags than for large flags. This agrees with how this intuitively should be. The only arbitrary elements in the new methodology are two weights to be chosen by policy makers. These represent the very intuitive decisions how heavy one wants to count (very) serious incidents versus detentions. It is expected that decision makers will be comfortable with determining these weights and this approach has already investigated and compared some 'reasonable' choices for these weights.

The method is applied and results are compared with methods developed by Perepelkin et al. [1] and with the excess factor methods currently used by the industry in order to demonstrate how the ranking of flags changes by introducing the 'enforcement effort' and 'sympathy' to registries with smaller fleet or with more challenging fleet profiles.

The proposed method is not restricted to the use of registries but could be extended to recognized organizations (RO) or Document of Compliance Companies or any other agent where the principal cannot be directly observe the effort, but only certain undesirable events that must be ascribed to a mixture of chance and inadequate effort, that is, in many moral hazard problems (see Laffont and Martimort [6] for this type of problem).

2. Derivation of proposed method

2.1. General concept

The development of the alternative method starts with the introduction of two numbers for each flag F, d_F , the quotient of the proportion of inspections of vessels under flag F that lead to detention and this proportion for all vessels, and z_F, the quotient of the proportion of the vessels under flag F that has been involved in a very serious accident and this proportion for all vessels. Thus, one gets that for d_F, as well as for z_F, the value 1 is a benchmark. For example, z_F is smaller, respectively larger, than 1 precisely if the proportion of vessels under flag F that has been involved in a very serious accident is smaller, respectively larger, than this proportion for all flags. It follows that if two flags are compared, F_1 and F_2, for which $d_{F_1} \geq d_{F_2}$ and $z_{F_1} \geq z_{F_2}$: in this case, one considers that the effort of F_2 is at least as good as that of F_1.

This idea is now extended in order to be able to compare the effort for each pair of flags. To this end, a weight factor c is introduced which is to be chosen by policy makers. As such, one can consider that the effort of F_2 is at least as good as that of F_1 precisely if $d_{F_1} + cz_{F_1} \geq d_{F_2} + cz_{F_2}$. That is, a first attempt is made for measuring the performance of a flag F:

$$Q_F^{\backprime} = d_F + cz_F \qquad \text{'crude performance measure'} \tag{1}$$