A method to measure flag performance for the shipping industry

Mikhail Perepelkin[1] , Sabine Knapp[2*], German Perepelkin[3], Michiel de Pooter[4*]

Econometric Institute Report 2009-04

Abstract

The subject of measuring the performance of registries has been a topic of policy discussions in recent years on the regional level due to the recast of the European Union (EU) port state control (PSC) directive which introduces incentives for flags which perform better. Since the current method used in the EU region entails some shortcomings, it has therefore been the subject of substantial scrutiny. Furthermore, the International Maritime Organization (IMO) developed a set of performance indicators which however lacks the ability to measure compliance as set out in one of its strategic directions towards fostering global compliance. In this article, we develop and test a methodology to measure flag state performance which can be applied to the regional or global level and to other areas of legislative interest (e.g. recognized organizations, Document of Compliance Companies). Our proposed methodology overcomes some of the shortcomings of the present method and presents a more refined, less biased approach of measuring performance. To demonstrate its usefulness, we apply it to a sample of 207,821 observations for a 3 year time frame and compare it to the best know current method in the industry.

[1] Kondratyevsky pr. 57/63, 195197, Saint Petersburg, Russia, email: mihail.perepelkin@gmail.com, tel: +79-045513566

[2] Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands, email: knapp@ese.eur.nl, tel: +44-7786309171

[3] Kondratyevsky pr. 57/63, 195197, Saint Petersburg, Russia, email: german@mcs.ru

[4] Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands, email: depooter@ese.eur.nl

* Disclaimer for Knapp and de Pooter: The views expressed in this article present those of the authors and do not necessarily represent those of the International Maritime Organization (IMO) nor those of the Board of the Governors of the Federal Reserve System or any other employee of the Federal Reserve System.

1. Introduction

Following a series of major oil tanker accidents in the 1970s, the concept of port state control (PSC) evolved to allow port states to conduct safety inspections on foreign flagged vessels entering its ports. The countries grouped themselves into PSC regimes based on Memoranda of Understanding (MoU) and today, ten PSC regimes exist, covering most port states. These regional MoU's enforce international legislation and act as a second line of defense against substandard shipping where the first line of defense is the flag state itself.

The topic of measuring flag state performance was first introduced by the oldest PSC regime, the Paris MoU[5] and was later adopted by the Tokyo MoU[6]. Each year, the "Black/Grey/White List (BGW-list) is published which is compiled using a specific method to classify registries into three groups – black, grey and white – where black listed flags perform worse than average and white listed flags perform better. In order to classify for any preferred treatment under the new recast EU directive, a registry needs to be on the white list. Given this new incentive, the current method of calculation has come under scrutiny recently because of its perceived inaccuracy in correctly determining each registry's classification and, consequently, a debate has started with the aim of developing a more accurate method. Since its introduction, the list however has become the industry benchmarking standard for flag performance although it is only applied at the regional level.

Despite the development of a complex legislative framework in the shipping industry, enforcement can be weak due to its international nature and can vary greatly on the flag state level. The legislative framework of about 50 conventions is developed by the International Maritime Organization (IMO) which is the regulator of the shipping industry but which lacks enforcement powers and does not monitor performance of its member states directly.

Notwithstanding the lack of enforcement, the IMO, through its technical cooperation committee (TCC), provides training and support to member states of developing countries and promotes harmonized enforcement of the legislative framework. In this respect, the IMO developed the Voluntary Member States Audit Scheme (VMSAS) which, at this stage, is voluntary but provides a mechanism to foster compliance.

Furthermore, one of the IMO's latest developments at council level encompasses the Strategic Plan for the Organization, of which the latest for the period 2008-2013 is based on Assembly Resolution A.989(25) [1] and sets out 13 broad strategic directions. Resolution A.990(25) [2] provides the corresponding High Level Action Plan (HLAP) for the 2008-2009 biennium. The IMO also developed a set of 42 performance indicators (PI) to measure progress made towards the 13 strategic directions. Strategic direction 2 deals with fostering global compliance but the level of compliance is only measured as a global average and not on an individual member state level. The current PIs further only measure the willingness of countries to undergo a voluntary audit or present aggregated non-compliance rates. No individual measurement of flag states or countries is made which limits the IMO's ability to identify weaknesses in the system.

[5] The Paris MoU covers the EU, parts of Canada and the Russian Federation

[6] The Tokyo MoU covers Asia, Australia, Chile and parts of the Russian Federation

Given this situation, this article analyzes the current method for compiling the BGW-list in section 2 and presents its major shortcomings. Section 3 develops a new method and applies it to a unique dataset of combined PSC inspection results and incident data. Section 4 summarizes the advantages of the method and provides a set of recommendations to the regulators on the regional level (e.g. PSC regimes) and the global level (IMO). The article ends with identifying other areas of application such as improved targeting for inspections and the measurement of performance of recognized organizations (RO) or Document of Compliance companies (DoC) on the regional or global level.

2. The current method for measuring flag state performance (BGW-list)

2.1. General concept of the current method

The current method in force is presented in Equation 1(a),(b) [3] and is used to construct a confidence interval (the *grey* area) for the allowable number of detentions, centered by the allowable number of detentions (the yardstick) with a range on the left hand side from the lower limit to the yardstick *(black to grey)* and a range on the right hand side from the yardstick to the upper limit *(grey to white)*. The underlying assumption to the formula is that the number of detentions follows a binomial distribution as detentions are counted in discrete numbers. The formula itself is derived from approximating this discrete binomial distribution by the continuous normal distribution. The reason for this approximation is simply because it simplifies the calculation of the limits. For calculating the discrete lower and upper limit a correction factor for continuity is added to each side (either +0.5 or -0.5). The lower and upper limits are calculated as

$$u_{blacktogrey} = N.p + 0.5 + z\sqrt{(N.p.(1-p)} \quad (1a)$$

$$u_{greytowhite} = N.p - 0.5 - z\sqrt{(N.p.(1-p)} \quad (1b)$$

where

$u_{blacktogrey}$= *lower (left) limit of detentions for the grey area*
$u_{greytowhite}$= *upper (right) limit of detentions for the grey area*
N = number of inspections
p = 0.07 (the set yardstick probably of a detention)
z = 1.645 critical value of the normal distribution at confidence level of 95%

The formula is applied each year to inspection data of the most recent 3 year period and with a minimum sample size of 30 or more inspections. Data from several PSC regimes are not combined and the lists are generated separately for the two regimes that currently use it (the Paris MoU and the Tokyo MoU). In order to make flags comparable, the concept of the *excess factor (EF)* is introduced with incremental steps of 0.03 for each EF - point. A flag with EF = 0 and below is placed on the white list while EF=1 and higher is used for the black list. Grey listed flags are between 0 and 1. The following section will analyze the various shortcomings on the current method used.

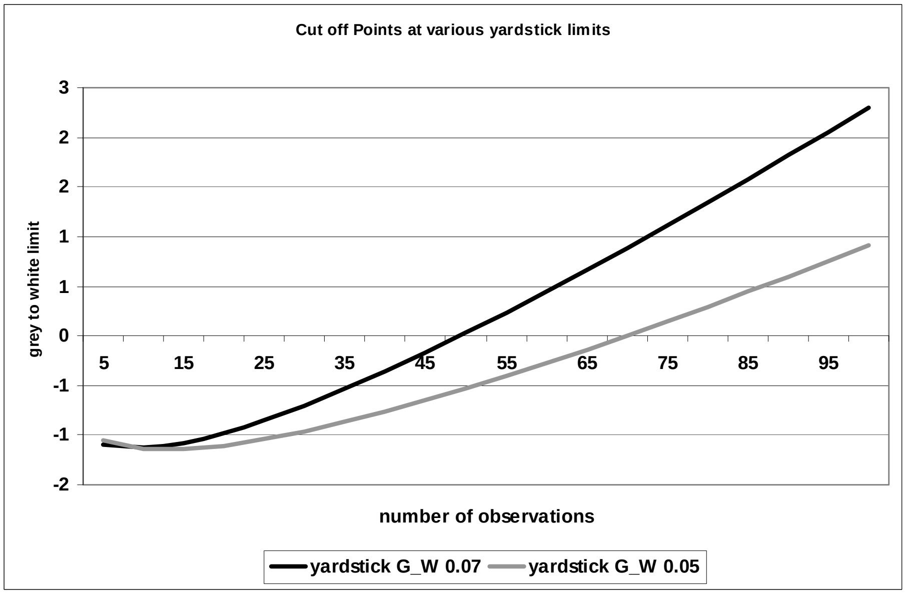
2.2. Critique 1: Inability to handle small sample sizes and inaccurate approximation

The use of the normal distribution to approximate the binomial distribution is based on the central limit theorem[7] and is therefore only accurate if N (the number of inspections) is sufficiently large. Consequently, for relatively small N the situation can occur that the medium to high limit for inspections is negative which is problematic since the number of inspections cannot be negative. Furthermore, the accuracy of approximation can be very low.

Figure 1 visualizes the concept of negative values for detentions for two yardsticks (p=0.07 and 0.05). One can see that the limit becomes positive at approx. 50 and 70 inspections respectively. These cut off points, however, do not represent the minimum amount of observations one should use for a relative accurate approximation of the binomial distribution by the normal distribution which is the underlying concept of the formula on hand.

Several rules of thumbs can be used to test the accuracy of the approximation and three of these are presented in Equation 2a-c (Schader and Schmid [4], Wackerly *et al* [5]) and were applied for p=0.07. A conservative approach was used and the results are presented in Table 1 for various sample sizes. One can easily see that sample sizes 50 and 70 are too small while sample sizes 90 and 100 are on the limit.

Figure 1: Cut off points at various yard sticks



rule of thumb 1*:* $N*P*(1-p)$ *should be* $>= 9(10)$ *(2a)*

rule of thumb 2*:* $N*p$ *or* $N(1-p) >5$[8] *(2b)*

rule of thumb 3*:* $p+-2*Sqrt(p*(1-p)/N)$ *should be between 0 and 1* *(2c)*

[7] The central limit theorem states that the larger the sample size, the more closely the sampling distribution will resemble the normal distribution.

[8] N*p is used for 0 <=p<=0.5 which is applicable here in this case while N(1-p) is used for 0.5<p<1.

Table 1: Rule of Thumb (RoT) values for approximating the binomial distribution with the normal distribution (with *p*=0.07)

		RoT 1	RoT 2	RoT 3	
N	**results**	**N*p**	**N*p*(1-p)**	**p+-2*Sqrt(p*(1-p)/N**	
Rule of Thumb		*> 5 (10)*	*> 9 (10)*	*Between 0 and 1*	
50	too small	3.50	3.26	0.142	-0.002
70	too small	4.90	4.56	0.131	0.009
90	On limit	6.30	5.86	0.124	0.016
100	On limit	7.00	6.51	0.121	0.019
150	Ok	10.50	9.77	0.112	0.028
170	Ok	11.90	11.07	0.109	0.031
200	Ok	14.00	13.02	0.106	0.034

In summary, the recommended amount of minimum observations for *p*=0.07 has been calculated to be 90-150 observations in order to provide a reasonable approximation of the discrete binominal distribution by means of the normal distribution.

2.2. Critique 2: Use of biased samples to measure performance

The world fleet eligible for port state control inspections can be estimated to be around 47% (Knapp [6]) of the world fleet or approx. 44,000 vessels (above 400 gt, based on data for 2005). In addition, based on these vessels, more than 50% of vessels get inspected in more than one regime and 20% in three or more regimes. The same split up per flag further reveals the approximate figures [7]: 33% of all registries get inspected in six or more regions, 12% in five regions, 9 % in four, 10% in three, 11% in 2 regions and 19% in one region. About 6% do not get inspected at all. These figures clearly show that there is an overlap in inspection efforts across various regimes which can not and should not be ignored. By choosing only one dataset (one regime) for the evaluation of performance and ignoring the information contained in the samples of other regimes, one could list the following implications:

- ***Important information of a vessel is omitted*** such as additional detentions and deficiencies.

- ***The effect of an inspection performed outside one region is ignored*** and risk profiling of this particular vessel could therefore be overestimated and the wrong vessel be targeted for inspection. According to Knapp [6], the average effect of a PSC inspection (depending on the overall ship risk profile) is estimated to be around a 5% to 10% decrease in the probability of a very serious casualty per inspection.

- ***One sample can only be seen as product of its own target factor and could be biased towards it.*** Each PSC regime uses target factors to decided whether a ship should be inspected or not in the effort to concentrate inspection efforts towards high risk vessels. Some regimes (e.g. USCG) performs a certain percentage of random inspections (about 5%) each year. In order to balance the sample towards increased randomness, it is recommended to combine the samples of various regimes into one dataset to measure the overall performance of a flag.

A possible remedy to rectify the shortcomings listed above is to combine inspection data from various PSC regimes as the basis to measure flag state performance. A possible criticism to this approach is the idea that the data cannot be combined due to technical reasons such as the incompatibility of the data or the different level of quality of inspections. However, the common practice of combining these types of datasets in the industry suggests otherwise and PSC data is combined by various industry players (e.g. Equasis, oil majors, maritime administrations) to provide a better picture of the performance of a fleet. Based on the legal framework of each MoU in existence, some differences can be uncovered but no significant differences with respect to the common goal which is the elimination of substandard vessels as a second line of defense.

Based on standard econometric techniques, Knapp and Franses [8] examined the effect of inspections performed by various port state control regimes and showed that these reduce the probability of a casualty. While this effect might change across regimes, it is still found to reduce the probability of a casualty which means that in essence, each regime does eliminate substandard vessels. Similar results were found for port state control and industry inspections by Bijwaard and Knapp [9] using duration analysis and applying the models to ship life cycles over a 29 year period.

2.3. Critique 3: Limited amount of factors to measure performance

The current method is based on the number of inspections and detentions and omits other information of relevance such as the number the types of deficiencies which are found during an inspection. It also only takes port state control inspections into account and omits other types of failures such as maritime incidents. In principle, other factors might be relevant such as the age of the vessel or the ship type. The new proposed method will address some of these factors.

2.4. Summary of critique of current method

The main drawbacks of the current method can be summarized as follows:

- The current sample size of 30 is small compared to a recommended number of 90-150 observations to ensure adequate approximation
- The method cannot accommodate small sample sizes (flag states)
- Data from several regimes are not combined and valuable information is omitted in measuring performance
- Other factors such as the number and type of deficiencies found during inspections are not taken into account.
- Other relevant information such as casualties are not taken into consideration
- Applicable to large number of inspections can also lead to complications, as the larger N is, the narrower the grey list becomes and consequently, the grey list is incomparably smaller that the black and white ones.
- The black, grey and white lists have no common criterion depicting the quality of a fleet. The role of this criterion is performed by the excess factor (EF), but its value depends on the list in question and is defined by different

procedures specific for each list. The boundary EF values are tailored for each list in order to ensure smooth transition from one list to another.

Given the list of shortcomings of the current method to determine performance of a flag, we suggest a revised method which should address at least some of these shortcomings. The new method is referred to as the *list of flag performance (LFP)* instead of BGW-list used by the port state control regimes. This is due to the fact that besides port state control inspections, the new method also account for deficiencies and casualties as defined by the IMO [10] and therefore presents a wider concept of measuring performance and, consequently, the quality of a flag.

3. An improved method to measure performance of flags

3.1. General concept of the method

Our proposed method ranks flag states based on one characteristic (*Q; quality of a flag)* which is common for all lists which facilitates comparisons across flags. By focusing on the ratio of the number of detentions over the number of inspections, smaller flag states can also be included in the BGW-list, which in this article we will denote by the list of flag performance (LFP). A further strong point of our proposal is, however, that the quality of a flag can be based on more information than just the number of inspections and the number of detentions as is the case in the current method. In particular, Q can be adjusted to take into account for example deficiencies or casualties, thereby allowing for a more accurate classification of flag states.

Q reflects the quality of a flag and is defined as the expected ratio of the number of detentions (D) to the number of inspections (N) but corrected in such a way that also deficiencies and casualties are taken into account. In addition, the calculation of Q can in principle be easily extended further to also take into account a rich set of additional factors such as ship type, ship size, among other variables.

The method is based on the ratio of detentions over inspections, in the sense that $D/\underline{N}$ is treated as an estimate of the (unobserved) probability of detention, p, for which in turn a normal probability distribution is assumed. Based partially on the latter assumption, the lower and upper limits of a confidence interval for p are derived which in turn are then combined with information on the number of deficiencies and casualties for each individual flag state. The method consists of adjusting D/N to a final value of Q which then reflects the quality of the flag.

The method assumes that the probability of detention, p, is an *unobserved* number (contrary to the current method where it is imposed to be equal to 0.07) and therefore constructs a confidence interval which will contain p with a certain required level of probability (e.g. 95% which is denoted by β in the proposal). This confidence interval is derived under the assumption that p is normally distributed and that D/N can be used as an estimate for the unknown value of p. Given the set-up, the lower and upper limits of the confidence interval for p can be derived which are denoted by L and H respectively.

The assumption of normality that underlies the method is an often-used assumption in statistics and is general very accurate. The derivation of the lower and upper limits of the confidence interval does, however, not crucially depend on this assumption of normality. The same limits result if the (weaker) assumption of just a symmetrical distribution for p is used. An alternative distribution would only result in a different factor t_β in the expressions for L and H[9].

Given L and H, the value of Q is then determined as D/N corrected with additional information regarding flag state quality and by incorporating deficiencies and casualties. In particular, the number of casualties, as well as the number of deficiencies of a flag state (the latter relative to the total number of deficiencies across all flag-states), will shift the value of Q away from D/N where the amount of the shift is directly proportional to the range between the limits L and H. As this range is larger for small flag states relative to larger flag states, this adjustment will have a bigger impact of the perceived flag quality of smaller flags.

Practically our proposal implies that a characteristic should be developed, which is common to all three lists (the Black/Grey/White list) and which would not be based solely on the number of inspections and number of detentions, but also take into account deficiencies and casualties as defined by the IMO in MSC/Circ. 953 and MEPC/Circ. 372, Reports on Marine Casualties and Incidents, (14th December 2000) [10][10]. The IMO currently distinguished between very serious, serious and less serious casualties but there are currently no mandatory reporting requirements for the latter two categories.

For the proposed method, we therefore only take very serious casualties into account. In the future, the method could be adapted to also include other categories once the IMO casualty investigation code becomes mandatory and reporting requirements for casualty investigations change. For the purpose of this article, the following definition for very serious casualties is used [11]:

'A very serious casualty means a marine casualty involving the total loss of the ship or a death or severe damage to the environment'.

It is worth noticing that the classification of casualty data from commercial data providers presents certain limitations. While seriousness can be identified and reclassified, some categories of the first events of a casualty are not readily available (e.g. in collisions, two ships are normally involved but not necessarily at fault). The data quality will hopefully improve in the future with the development of the Global Integrated Ship Information System (GISIS) at IMO.

We prefer to base the measurement of performance or quality on port state control data in conjunction with casualty data in order to better capture the level of non-compliance (given by the results of the port state control inspections) and the number

[9] Note in particular that the notation of t_β is typically used for the Student-*t* distribution which, although it resembles the normal distribution by being bell-shaped and symmetrical, has more probability mass in the tails compared to the normal distribution.

[10] The Maritime Safety Committee (MSC84) adopted MSC Resolution 255(84) on 16 May 2008 where the definitions were slightly changed and no longer distinguish between serious and less serious casualties. The definition for very serious casualty remains unchanged however. The reporting requirements will also change in the future.

of very serious casualties during a certain time period. If one would only base the analysis on casualty data, the outcome of the inspections are ignored including deficiency and detention information. Ships which are targeted for inspection can benefit from an inspection and the probability of casualty is therefore decreased (see Knapp [6] for a discussion on the topic where the decrease is estimated to be 6% per inspection). A measurement including both datasets provides a more balanced approach by taking the level of non-compliance into account.

Finally, the limits of values of the characteristics for each list (B/G/W) should be defined aiming, in particular, to size all three lists in a similar way so that the new lists have comparable length. We will denominate this characteristic as *Q*. Let us consider *Q* as the expected ratio of the number of detentions (*D*) to the number of inspections (*N*) corrected by means of some (variable) number which takes account of deficiencies and casualties.

3.2. Mathematical derivation of the proposed new method

Mathematically our method could be interpreted as follows (at this stage, without accounting for deficiencies and casualties). An estimate of the unknown probability *p* of an event (detention in our case) could be obtained on the basis of its observed frequency *p'* in *N* independent trials. The result of an independent trial X_i *(i=1. N)* is taken as 1 in case of a detention and 0 if the inspection went successfully. The frequency *p'* is the mean value of the outcomes of X_i and is given by:

$$p' = \frac{1}{N}\sum_{i=1}^{N} X_i$$

In this setup the population mean of the random variable *X* is equal to *p* and its dispersion equal to *p*q*, where *q=1-p*. The population mean of the frequency *p'* is also equal to *p*, i.e., the estimate of *p* is a fixed (non-moveable) value. The dispersion of *p'* (denoted by *DIS*) is equal to:

$$DIS[p'] = \frac{p\cdot(1-p)}{N}$$

In theory, with the number of inspections going to infinity, *p'* converges in probability to *p*, meaning that the dispersion of *p'* around *p* reduces to zero. In practical terms, *p'* may be used in all cases to estimate the unknown value of *p*. However, correctness and reliability of this estimate are of importance – a so-called confidence interval (CI) for *p* should be obtained which length positively depends on *DIS*.

Let us consider first a simple case when *N* is comparatively large and the true probability *p* is neither too low nor too high. We may then consider that the frequency *p'* is a random value and its distribution is close to the normal (or Gaussian) distribution. Calculations demonstrate that such an assumption may be used even with relatively low *N* values; it is practically acceptable when *D* exceeds 4 (Ventzel [12]). Let us assume this condition is fulfilled, and that the distribution of the frequency *p'* corresponds to the normal distribution. The population mean and square root of the dispersion (root mean square deviation, $\sigma_{p'}$) of the distribution of p' will then be defined by:

$$m_{p'} = p \,, \qquad \sigma_{p'} = \sqrt{\frac{p \cdot (1-p)}{N}}$$

Let us assume that also the value of p is known. Let us fix a confidence level β and define the interval ($p - \varepsilon_\beta$, $p + \varepsilon_\beta$) such that p' is contained therein with probabilityβ:

$$P\left(\left|p' - p\right| < \varepsilon_\beta\right) = \beta$$

Since p' value is normally distributed, then

$$P\left(\left|p' - p\right| < \varepsilon_\beta\right) = 2\Phi\left(\frac{\varepsilon_\beta}{\sigma_{p'}}\right) - 1 = \beta \quad \text{from which} \quad \varepsilon_\beta = \sigma_{p'} \cdot \Phi^{-1}\left(\frac{1+\beta}{2}\right),$$

where Φ^{-1} is the inverse of the cumulative normal distribution function Φ. To define β, let us denominate

$$t_\beta = \Phi^{-1}\left(\frac{1+\beta}{2}\right) \quad \text{then} \quad \varepsilon_\beta = t_\beta \cdot \sigma_{p'},$$

where t_β is to be found in standard normal probability tables[11]. Therefore we may, with probability β, assert that

$$\left|p' - p\right| < t_\beta \cdot \sqrt{\frac{p \cdot (1-p)}{N}} \qquad (3)$$

i.e. that the probability of detention p really lies within the limits of such an interval. In practice, we do not know the true value of p. Nevertheless, the expression in equation (3) will remain valid at probability β irrespective of whether the probability p is known or not. Having obtained from experience a specific p' value it is possible, by using (3), to define an interval which would contain the true value of p value with probabilityβ. By rewriting equation (3) we can then define the lower (L) and upper (H) limits of the interval (range) and by replacing the value of p' with that of D/N, we obtain the following expressions given in equation (4):

$$L = \frac{\frac{D}{N} + \frac{1}{2} \cdot \frac{t_\beta^2}{N} - t_\beta \cdot \sqrt{\frac{D \cdot (N-D)}{N^3} + \frac{1}{4} \cdot \frac{t_\beta^2}{N^2}}}{1 + \frac{t_\beta^2}{N}} \,. \qquad 4(a)$$

[11] The values for t_β are as follows: 1.65, 1.96 and 3.) for a 90%, 95% or 99% confidence level

$$H=\frac{\frac{D}{N}+\frac{1}{2}\cdot\frac{t_\beta^2}{N}+t_\beta\cdot\sqrt{\frac{D\cdot(N-D)}{N^3}+\frac{1}{4}\cdot\frac{t_\beta^2}{N^2}}}{1+\frac{t_\beta^2}{N}} \qquad 4(b)$$

It is worth noticing that in case *D* is smaller than 4 (Ventzel [12]) the task may be resolved by using numerical methods, when CI (confidence interval) is to be defined by a direct calculation of the difference of the values of the integral distribution function at the points *(D/N – ε)* and (*D/N* + *ε*). In this case CI may be defined exactly (without assumption on the normal distribution of *p'*), from the following equation:

$$\int_{\max\left\{\frac{D}{N}-\varepsilon,0\right\}}^{\frac{D}{N}+\varepsilon}(N+1)\cdot C_N^D\cdot x^D\cdot(1-x)^{N-D}\cdot dx=\beta\,. \qquad (5)$$

However, such an exact analytic solution is impossible with high numbers of inspections and detentions, as in such case the function of distribution of detention characteristic cannot be expressed through simple functions, and with fixed *N* and *D* it represents a polynomial expression of power *N*.

In this article, however, we concentrate on sample sizes above 30 since by combining data from various port state control regimes, there are not many flags left with very small samples sizes. The extension of using equation (5) for very small sample sizes might be important if each regime continues to evaluate flags only on their own inspection data where the overall sample size is considerably smaller. Furthermore, we use the cut off point of 30 in order to be able to compare the results of the EF (as applied currently) and the new proposed method.

For small fleets the width of the interval from *L* to *H* are substantially wide, and in their case the influence of the number and weight of deficiencies and casualties on *Q* is significant. For large fleets the range of *Q* variation is narrow, and deficiencies will not significantly influence *Q*. This is intuitive since with a high number of inspections the statistics of detention will describe the quality of a fleet sufficiently well. With a low number of inspections, the fleet quality characteristic *Q* should be adjusted through the statistics on deficiencies. This adjustment can be made by shifting the calculated value within the range obtained or beyond it. Therefore a fleet having a lower number of deficiencies with the same *D*/*N* ratio will be considered as a 'better' one in terms of *Q*, and the lower *N* is, the greater role will be attributed to the statistics of deficiencies.

To take account of the presence of casualties of a flag state, *Q* may be adjusted (increased) by a value proportional to the length of the *L* to *H* range and some factor which is denominated as k_z. This parameter acts like a kind of penalty factor.

To take into consideration the number of deficiencies (*D*/*NC*) it could be proposed to compare the number of *D*/*NC*s for each included flag state (taking into consideration their 'weight'), to a mean value of *D*/*NC*s across all flag states and, depending on the results of such a comparison, to shift *Q* by a value proportional to the length of the *L* to *H* range. The resulting formula for *Q* is given in equation (6):

$$Q = \frac{D}{N} + k_Z \cdot Z \cdot (H - L) + \left(\frac{Def}{N \cdot Def_{Mean}} - 1 \right) \left(\frac{H - L}{2} \right) \tag{6}$$

where

D and N are the numbers of detentions and inspections
k_z – casualty 'significance' factor,
Z – number of very serious casualties,
Def – total weight of D/NCs revealed,
Def_{Mean} – mean weight of D/NCs for all flags per one inspection during a reference period (one year in our case),
H, L – bounds of the (Low , High) range calculated as shown above.

To take account of the weight of a *D*/*NC*, *Def* should be defined as a sum of all *D*/*NC*s calculated with their weights where Def_{Mean} is to be calculated according to the following formula given in equation (7):

$$Def_{Mean} = \frac{\sum_{i=1}^{M} def_i}{N_{\Sigma}} \tag{7}$$

where

M = number of all D/NCs for all Flag States within the reference period and
N_{Σ} =number of all inspections conducted during the reference period.

3.3. Benefits and possible drawbacks of the proposed new method

Two initial benefits of the proposed method relative to the current method are the following. The first benefit concerns the use of the yardstick probability of detention. With the current method, the probability of detention has to be determined exogenously (currently at 0.07). Even when this probability is determined based on data on detentions and inspections, it is not obvious what the 'true' probability of detention is. The new approach constructs confidence intervals for *p*, thereby stating that the true value of *p* will be contained within such intervals, without the need to know or assume the exact value.

The second benefit results from using the ratio of the number of detentions over the number of inspections. A common assumption in the statistical literature is to assume that this quantity follows a normal distribution (provided that *N* is sufficiently large and *p* neither too close to either 0 or 1). Because the *D/N* takes on non-integer values there is no longer the need to approximate an integer-valued distribution with a continuous distribution as is the case in the current method where the focus is entirely on the number of detentions (which is integer-valued). The distributional assumptions and approximations needed for the current method are only valid if the sample size is sufficiently high. The new method is less sensitive to the number of observations and a smaller number of inspections are therefore acceptable in order to apply it to smaller sample sizes.

The main benefit of the proposed method is, however, that information beyond only inspections and detentions can be incorporated to assess the quality of flags. Casualties can be taken into consideration and accounted for where the combination of port state control data and casualty data is of direct interest of the performance indicator measurement at IMO level. The method, when applied to a combined datasets of global PSC and casualty data, can further provide a more refined method to measure performance and the quality of flag.

Although additional analysis will be needed to determine how to best incorporate additional characteristics such as ship type, ship size and age, doing so allows for a fairer comparison of the quality of different flag states and will therefore result in a more accurate ranking of flag states on the *list of flag performance (LFP).* Our current proposed method does not yet include correction factors for these variables.

The next section applies the proposed method to a combined dataset of PSC inspections and casualties and compares the relevant results with the EF method (the current method).

3.2. Application of the new method and comparison to the current method

The dataset used for applying the new method is based on Knapp and Franses [8] and Bijwaard and Knapp [9] and covers a time period from January 2005 to December 2007 for which data on casualties and port state control inspections was extracted. The dataset used in the analysis contains 197,334 inspections, 10,155 detentions (5.1%), 647,254 deficiencies and 332 very serious casualties. Data on casualties were combined from Lloyd's Register Fairplay (LRF) and the IMO where data from LRF was reclassified according to IMO definitions to match the correct seriousness. Observations with unknown flag were grouped into one group denoted 'unknown'.

The current method as given in equation 1(a,b) and the proposed method, as given in equation (6), are applied to the data on hand for all flags with 30 or more observations. We first use the current method as currently applied (with sample size 30) but we also apply the recommended minimum sample size of 90 based on the calculations presented in section 2.2 to ensure adequate approximation for the excess factor method. For the new method, the sample size of 30 is used since it assumes a normal distribution for the detention probability p (where D/N is the estimate of p) and the general concept that if the population distribution is close to normal, then 30 in most cases will be sufficiently large.

For the incorporation of the deficiencies based on equation (7), Table 2 provides the proposed weight factors used for the method indicating the seriousness of a deficiency group. Most weight is assigned to the group dealing with fire fighting appliances and life saving appliances followed by pollution prevention (MARPOL Annex I). Other deficiency groups such as certificates which are administrative deficiencies are given smaller weights compared to deficiencies which more likely provide an indication of the lack of proper maintenance of the vessel and safety management.

Various scenarios are presented and compared to the current excess factor method. The scenarios contain different significant factors for the casualties varying from zero (non-incorporation) to significant factor 2 (most conservative approach).

- *Base Scenario*: current method in force with Rank EF (excess factor)
- *Scenario 1:* Q without very serious casualties with Rank Q (no *Z*)
- *Scenario 2:* Q including very serious casualties with Rank Q (k_z=0.5)
- *Scenario 3:* Q including very serious casualties with Rank Q (k_z=1.0)
- *Scenario 4:* Q including very serious casualties with Rank Q (k_z=1.5)
- *Scenario 5:* Q including very serious casualties with Rank Q (k_z=2.0)

Table 2: Weight factors used for deficiencies

Group	Description	Weight
0100	Certificates and documents	0.95
0200	Crew documents	0.95
0300	Crew and accommodation	0.90
0400	Food and stores	0.85
0500	Working spaces	0.90
0600	Life-saving appliances	1.15
0700	Fire fighting	1.20
0800	Prevention of accidents	1.00
0900	Safety in general	1.00
1000	Alarms	0.98
1100	Carriages of cargoes	0.95
1200	Load Line	1.05
1300	Mooring appliances	0.92
1400	Propulsion and auxiliary engines	1.00
1500	Safety of navigation	1.02
1600	Radio communications	1.05
1700	MARPOL Annex I - Oil	1.20
1800	Oil, chemical and gas tankers	1.05
1900	MARPOL Annex II – Noxious Liquid Substances	1.03
2000	Solas related operational deficiencies	1.00
2100	MARPOL relation operational deficiencies	1.01
2200	MARPOL Annex III - IMDG	1.01
2300	MARPOL Annex V - Garbage	1.01
2500	Intern. Safety Management Code (ISM)	1.05
2600	Bulk Carriers – additional safety measures	1.10
9800	Other deficiencies – with hazard	1.10
9900	Other deficiencies – no hazard	1.00

The results are presented in Appendix 1 where the current method is denoted with *EF* (excess factor) and the results for the new proposed method are denoted by *Q* (Quality). In addition, the total number of inspections (*N*), detentions (*D*), deficiencies (*Total Def*) and very serious casualties (*Z*) for the 3 year period are given for the flags that were evaluated. Flags which were not evaluated are listed at the end of the table. Out of the total 139 flags, 107 are evaluated (21 more than with the current method if the recommended sample size of 90 is used) and 32 are below sample size 30.

It is worth noticing that we keep registries separate from the traditional registries (e.g. the Norwegian International Registry or the Danish International Registries). As indicated in section 3.2, the smaller flags could also be evaluated by using equation (5) but for the purpose of the demonstration of the new method and the comparison to the present method, we use sample size 30 and exclude all flags below sample size 30.

In order to compare the results, the rank of each flag for each scenario is calculated and presented in the table in the last three columns and sorted by the most conservative approach for k=2. In order to evaluate the two methods, we base the comparison on sample size 30 for both methods so that the ranks can be compared. The last two columns present the best and the worst rank of the flag across all methods.

In addition to the ranks, the flags evaluated with the new method are also classified according to the old concept of the BGW-list. However, comparison of this classification with the old method is not straight forward since the length of the lists is not the same for the old method. This classification should only be taken as an overall guidance in comparing the two methods.

4. Evaluation of proposed method and relevant policy implications

This last section of the article provides a discussion of the results and relevant policy implications. In evaluating the proposed method, the following underlying criteria are used:

1. The validity and workability of the method presented from a statistical perspective and its ability to deal with smaller sample sizes
2. The combination of information used to measure performance and the level of access to the requested data
3. The added value of changing the method including its possible impact
4. The variability of the results when applying various methods to the same data

With respect to item 1, the new method is based on general theory of probability. The assumption of normality that underlies the method is an often-used assumption in statistics and is generally very accurate and more accurate than the probabilistic assumptions underlying the current method. An advantage of the new proposed method versus the current method is that more flags can be evaluated due to the smaller required sample sizes. Workability of the method is relatively straight forward and allows running scenarios based on different significant factors (*k* factors) and deficiency weights so that regional and global differences can be taken into consideration.

With respect to the combination of data which is used for the new method, a clear advantage of the proposed method compared to the current method is the incorporation of past deficiency information and casualty information although casualty data from commercial data providers need to be reclassified to match the IMO definitions of seriousness.

It is worth noticing that the identification of casualty first events with respect of collision and contacts should be improved with commercial data providers so that a better distinction between ships at fault and ships that are merely a victim of a collision can made. For the purpose of this analysis, we excluded the category collision of very serious casualties since this distinction at current is not clear in the data available via commercial data providers. A future possible incorporation could be different weight factors for different types of casualties depending on their seriousness.

Access to the data and the improvement of the identification of first events and categorization of seriousness of casualty data will be improved with the development of the Global Integrated Ship Information System (GISIS) at IMO. It is also planned that GISIS will contain port state control information from all regional MoU's and will therefore combine both data sources. For the time being, commercial data can be used which is easily accessible.

For reasons of conciseness, additional variables such as the age, size and ship type are not yet taken into account in our current analysis and further analysis would therefore be required to study how these ship and flag state characteristics should be best taken into account and how important these are with respect to measuring overall performance. Other areas of measurement such as results of flag state audits could also be incorporated in the future so that a similar method can be used to measure the performance of a member state where flag state performance would only be one component.

Criteria items 3 and 4 are more difficult to evaluate. Appendix 1 presents an indication of the variability of the results when applied to the same data and under various scenarios while Figure 2 presents the results graphically for the top 20 flags.

Figure 2: Top 20 Flags with respect to variability in ranks

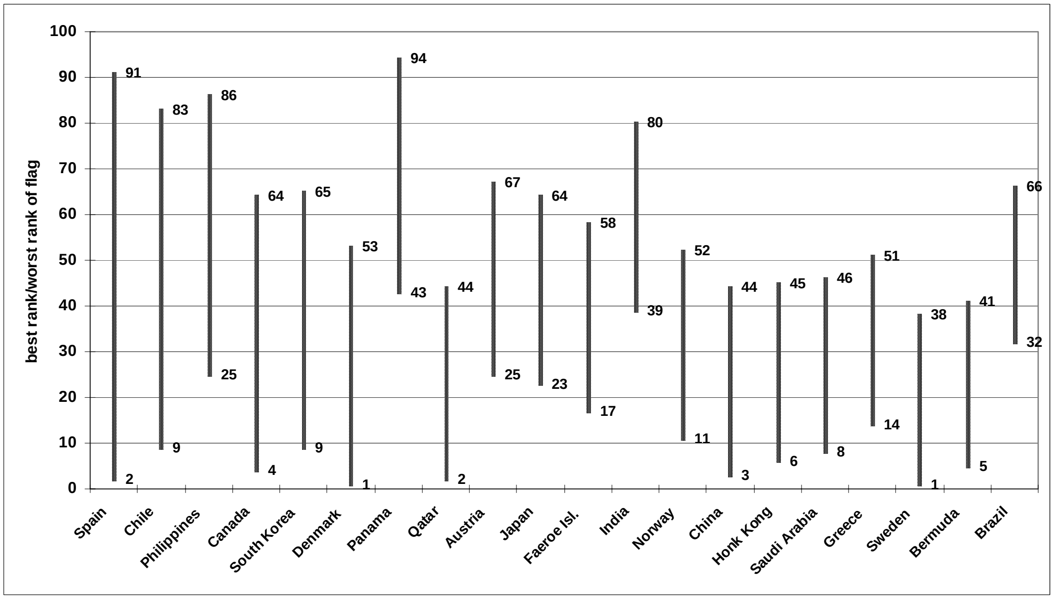


Figure 2 indicates the best rank and the worst rank by using the EF method and any of the scenarios of the proposed method. This differences are due to the incorporation of the deficiencies and the casualties (with various weight factors) and are therefore mostly the differences of the ranks of the EF method and the new method at k=2 or 1.5. Top difference in rank is with 89 ranks for Spain followed by 74 (Chile), 61 (Philippines) and 60 (Canada). The possible impact is the shift of some flags with respect to their ranks where the shift can go into both directions. Where shifts lead to a decrease in rank, it becomes politically very sensitive and vice versa, especially for flags where one can observe large variability. If the method is applied globally, then it should be applied to a combined dataset of all port state control regimes and casualty data by using the definitions of IMO. In this way, the underlying dataset is less biased and present a more complete picture then applying the method separately in each area.

Although the shift of the flags with respect to the classification of black, gray and white is not a straight forward comparison as indicated below, we will provide a high level interpretation. Most flags which are on the black list under the current method also appear on the list of the worst performing flags. Incorporation of deficiencies and casualties compared to the current method leads to a shift of some white listed flags under the current method to the grey or black area with different k factors. These flags are Panama, the Philippines, India, Turkey and Russia. In addition, some flags improve their ranking under the new method such as Denmark which is on the grey list under the current method (rank 53) and moved to rank 1 under the new method. Another example would be the Faeroe Islands, Austria, Estonia, Brazil, Lithuania, Latvia and Myanmar.

The added value of the proposed method can be summarized to be a method with a higher level of accuracy based on general theory of probability using standard statistical assumptions. It further allows the evaluation of flags with smaller sample sizes compared to the current method where the minimum sample size is recommended to be 90 (not the 30 currently used). The underlying combination of data is less biased and contains more relevant information to measure flag performance (e.g. inspection results from various regimes, the number and type of deficiencies or very serious casualties). Furthermore, regulators can determine the level of importance of deficiencies by using weights for deficiencies and casualties. Finally, the method can also be applied to recognized organizations (RO) once data and deficiency information becomes more readily available. Another area of application would be the DoC companies where sample sizes are very small due to the large amount of companies.

Acknowledgements

We would like to acknowledge the following port state control regimes for providing data: Paris MoU, Viña del Mar Agreement, Indian Ocean MoU, Caribbean MoU, the United States Coast Guard and AMSA. In addition, we would like to thank Lloyd's Register Fairplay for the casualty data. Furthermore, we would like to acknowledge the contribution of Vladimir Semenov and thank him for his valuable input.

References

[1] Assembly Resolution A.989(25), Strategic Plan for the Organization (for the six-year period 2008-2013), adopted 20th November 2007, IMO, London

[2] Assembly Resolution A.990(25), High Level Action Plan of the Organization and Priorities for the 2008-2009 biennium, adopted 29th November 2007, IMO, London

[3] Paris MoU Annual Report, 2006, http://www.parismou.org/

[4] Schader M, Schmid F, Two Rules of Thumb for the Approximation of the Binomial Distribution by the Normal Distribution, The American Statistician, 1989, Vol. 43, No. 1, pp. 23-24

[5] Wackerly D, Mendenhall W. and Scheaffer RL, Mathematical Statistics with Applications, 1995, 5th ed., P W S Publishers

[6] Knapp, S , The Econometrics of Maritime Safety – Recommendations to enhance safety at sea, 2006, Doctoral Thesis, Erasmus University, Rotterdam
[7] Based on data from Annual Reports for the years 2006 to 2006 from the following PSC regimes: Paris MoU, Viña del Mar Agreement, Indian Ocean MoU, Caribbean MoU, the United States Coast Guard and AMSA
[8] Knapp S, Franses PH, A global view on port state control - econometric analysis of the differences across port state control regimes, Maritime Policy and Management, 2007, 34(5), pages 453-483
[9] Bijwaard G and Knapp S, Analysis of Ship Life Cycles – The Impact of Economic Cycles and Ship Inspections, Marine Policy 2009, volume 33, pp. 350-369
[10] MSC/Circ. 953, MEPC/Circ. 372, Reports on Marine Casualties and Incidents, Revised harmonized reporting procedures, adopted 14th December 2000, IMO, London
[11] MSC Resolution MSC.255(84), Casualty Investigation Code, adopted 16 May 2008, IMO, London
[12] Ventzel, ES, Theory of Probability, Science, Moscow, 1969

Appendix 1: List of flag performance (LFP) - results of current method and new method for the years 2005-2007 for various scenarios of k

	Total	Total	Total				no Z	k=0.5	k=1.0	k=1.5	k=2.0	EF	Q	Q	Q	Q	Q	best	worst
Flag	N	D	Def	Z	EF(90)	EF(30)	Q	Q	Q	Q	Q	no Z	no Z	k=0.5	k=1.0	k=1.5	k=2.0	rank	rank
Indonesia	591	111	5424	16	4.068	4.068	0.282	1.050	1.820	2.590	3.360	95	93	105	106	107	107	93	107
Bangladesh	45	16	448	3	BSS	6.661	0.785	1.380	1.970	2.560	3.150	103	105	107	107	106	106	103	107
North Korea	1291	370	13625	16	7.533	7.533	0.371	0.974	1.580	2.180	2.780	105	99	103	104	105	105	99	105
Jordan	43	17	428	2	BSS	7.738	0.824	1.230	1.640	2.050	2.460	106	106	106	105	104	104	104	106
Honduras	225	46	1307	4	4.061	4.061	0.276	0.595	0.915	1.230	1.550	94	91	98	102	102	103	91	103
Libya	67	25	803	1	BSS	7.863	0.824	0.992	1.160	1.330	1.490	107	107	104	103	103	102	102	107
Comoros	1049	221	9476	6	5.009	5.009	0.278	0.504	0.730	0.956	1.180	98	92	97	100	100	101	92	101
Bolivia	88	24	856	1	BSS	5.244	0.558	0.696	0.833	0.970	1.110	100	101	102	101	101	100	100	102
Sierra Leone	560	125	5501	3	5.160	5.160	0.331	0.489	0.647	0.804	0.962	99	97	96	98	99	99	96	99
Maldives Islands	40	5	209	1	BSS	0.849	0.290	0.447	0.605	0.763	0.920	73	94	94	96	98	98	73	98
Cambodia	6789	908	54122	14	2.901	2.901	0.152	0.326	0.499	0.673	0.846	89	74	89	93	96	97	74	97
St. Kitts and Nevis	380	79	3539	2	4.486	4.486	0.327	0.451	0.575	0.700	0.824	96	96	95	94	97	96	94	97
Cook Islands	145	14	709	2	0.847	0.847	0.156	0.306	0.457	0.607	0.757	72	75	87	91	93	95	72	95
Panama	39677	1977	132138	52	-0.612	-0.612	0.050	0.220	0.391	0.561	0.732	43	52	77	86	90	94	43	94
Mongolia	693	107	5769	3	3.084	3.084	0.223	0.346	0.469	0.593	0.716	90	86	92	92	91	93	86	93
St. Vincent & the Gren.	4868	489	25087	11	1.784	1.784	0.108	0.251	0.393	0.535	0.678	83	69	79	87	88	92	69	92
Spain	49	0	86	2	BSS	0.001	-0.042	0.134	0.310	0.486	0.662	52	2	61	77	86	91	2	91
Sri Lanka	41	13	334	0	BSS	5.395	0.648	0.648	0.648	0.648	0.648	101	104	101	99	95	90	90	104
Slovakia	494	89	3043	2	3.752	3.752	0.230	0.334	0.437	0.541	0.644	92	87	91	89	89	89	87	92
United Arab Emirates	79	10	270	1	BSS	1.079	0.169	0.282	0.395	0.508	0.621	76	79	83	88	87	88	76	88
Moldavia	62	19	560	0	BSS	5.764	0.618	0.618	0.618	0.618	0.618	102	103	100	97	94	87	87	103
Philippines	609	12	1547	8	-1.272	-1.272	0.023	0.168	0.313	0.459	0.604	25	34	70	79	85	86	25	86
Kiribati	71	17	752	0	BSS	4.024	0.599	0.599	0.599	0.599	0.599	93	102	99	95	92	85	85	102
Dominica	375	40	2338	2	1.404	1.404	0.159	0.255	0.351	0.447	0.544	79	77	80	85	84	84	77	85
Chile	40	1	101	1	BSS	0.215	-0.003	0.124	0.251	0.378	0.505	56	9	59	73	78	83	9	83
Azerbaijan	168	19	1020	1	1.217	1.217	0.196	0.270	0.344	0.418	0.492	77	84	82	82	82	82	77	84
Lebanon	256	35	1784	1	2.116	2.116	0.221	0.285	0.349	0.414	0.479	85	85	84	84	81	81	81	85
India	681	23	1762	5	-0.739	-0.739	0.035	0.143	0.251	0.358	0.466	39	47	64	72	76	80	39	80
Georgia	1945	383	16034	2	4.742	4.742	0.239	0.293	0.347	0.401	0.455	97	88	85	83	80	79	79	97
Thailand	1590	105	7513	5	0.317	0.317	0.077	0.171	0.265	0.359	0.453	60	63	72	75	77	78	60	78

	Total	Total	Total				no Z	k=0.5	k=1.0	k=1.5	k=2.0	EF	Q	Q	Q	Q	Q	best	worst
Flag	N	D	Def	Z	EF(90)	EF(30)	Q	Q	Q	Q	Q	no Z	no Z	k=0.5	k=1.0	k=1.5	k=2.0	rank	rank
Albania	304	94	2636	0	7.530	7.530	0.442	0.442	0.442	0.442	0.442	104	100	93	90	83	77	77	104
Turkey	3975	244	15795	8	-0.064	-0.064	0.065	0.156	0.248	0.339	0.431	50	60	66	71	75	76	50	76
Syria	782	135	7836	1	3.699	3.699	0.258	0.299	0.339	0.380	0.420	91	89	86	81	79	75	75	91
Taiwan, China	438	33	1642	2	0.626	0.626	0.090	0.167	0.244	0.320	0.397	70	66	69	70	73	74	66	74
Russia	4575	256	14451	8	-0.271	-0.271	0.056	0.138	0.220	0.302	0.384	47	54	62	68	69	73	47	73
Belize	3041	298	18558	4	1.645	1.645	0.113	0.178	0.243	0.308	0.372	81	70	75	69	70	72	69	81
Tuvalu	616	82	4685	1	2.383	2.383	0.195	0.236	0.277	0.318	0.359	88	83	78	76	72	71	71	88
Pakistan	65	10	436	0	BSS	1.615	0.334	0.334	0.334	0.334	0.334	80	98	90	80	74	70	70	98
Jamaica	97	16	668	0	2.269	2.269	0.313	0.313	0.313	0.313	0.313	86	95	88	78	71	69	69	95
Papua New Guinea	45	2	405	0	BSS	0.327	0.264	0.264	0.264	0.264	0.264	62	90	81	74	68	68	62	90
Vietnam	1188	134	7819	1	1.940	1.940	0.145	0.172	0.200	0.227	0.255	84	72	73	67	67	67	67	84
Unknown Flag	7008	603	29559	4	1.355	1.355	0.089	0.130	0.170	0.210	0.250	78	65	60	64	66	66	60	78
South Korea	3884	63	15536	9	-1.673	-1.673	0.019	0.074	0.130	0.186	0.241	9	28	53	58	65	65	9	65
Canada	174	1	95	2	-1.386	-1.386	-0.021	0.043	0.107	0.171	0.235	19	4	37	55	61	64	4	64
Japan	624	12	1427	3	-1.299	-1.299	0.021	0.074	0.127	0.181	0.234	23	30	52	57	64	63	23	64
Malta	9514	432	31061	6	-0.698	-0.698	0.046	0.084	0.123	0.161	0.200	40	50	56	56	58	62	40	62
Algeria	164	16	1031	0	0.885	0.885	0.180	0.180	0.180	0.180	0.180	74	82	76	66	63	61	61	82
Tonga	87	8	419	0	BSS	0.716	0.175	0.175	0.175	0.175	0.175	71	81	74	65	62	60	60	81
Israel	210	5	315	1	-0.638	-0.638	0.023	0.060	0.097	0.133	0.170	42	35	46	51	54	59	35	59
Poland	116	12	548	0	0.886	0.886	0.170	0.170	0.170	0.170	0.170	75	80	71	63	60	58	58	80
Egypt	293	35	1571	0	1.677	1.677	0.166	0.166	0.166	0.166	0.166	82	78	68	62	59	57	57	82
Cyprus	6640	216	16528	5	-1.123	-1.123	0.032	0.064	0.097	0.130	0.163	29	43	49	52	53	56	29	56
Morocco	169	12	1131	0	0.514	0.514	0.156	0.156	0.156	0.156	0.156	68	76	67	61	57	55	55	76
Ukraine	986	125	5143	0	2.331	2.331	0.149	0.149	0.149	0.149	0.149	87	73	65	60	56	54	54	87
Tunisia	57	3	333	0	BSS	0.365	0.139	0.139	0.139	0.139	0.139	65	71	63	59	55	53	53	71
Norway	203	3	349	1	-1.007	-1.007	-0.001	0.033	0.067	0.101	0.135	33	11	28	46	50	52	11	52
Greece	4348	88	6448	4	-1.535	-1.535	0.018	0.044	0.070	0.095	0.121	14	26	38	47	48	51	14	51
United States of America	339	6	742	1	-1.155	-1.155	0.022	0.047	0.071	0.096	0.120	28	32	41	48	49	50	28	50
Kuwait	111	9	342	0	0.625	0.625	0.106	0.106	0.106	0.106	0.106	69	68	58	54	52	49	49	69
Liberia	8797	227	19097	4	-1.378	-1.378	0.025	0.045	0.065	0.086	0.106	21	37	39	44	47	48	21	48
Ethiopia	44	2	207	0	BSS	0.336	0.104	0.104	0.104	0.104	0.104	63	67	57	53	51	47	47	67

	Total	Total	Total				no Z	k=0.5	k=1.0	k=1.5	k=2.0	EF	Q	Q	Q	Q	Q	best	worst
Flag	N	D	Def	Z	EF(90)	EF(30)	Q	Q	Q	Q	Q	no Z	no Z	k=0.5	k=1.0	k=1.5	k=2.0	rank	rank
Italy	2214	52	4159	2	-1.354	-1.354	0.021	0.041	0.061	0.080	0.100	22	31	35	40	46	46	22	46
Honk Kong, China	7369	109	15755	5	-1.760	-1.760	0.014	0.035	0.057	0.078	0.099	6	20	31	35	45	45	6	45
China	3477	38	8377	4	-1.859	-1.859	0.011	0.033	0.054	0.076	0.098	3	17	27	33	40	44	3	44
Vanuatu	462	8	740	1	-1.292	-1.292	0.016	0.036	0.057	0.077	0.097	24	23	32	36	42	43	23	43
Norway (NIS)	4180	84	7293	3	-1.537	-1.537	0.018	0.038	0.058	0.077	0.097	13	27	33	37	43	42	13	43
Bermuda	803	7	762	2	-1.790	-1.790	0.006	0.029	0.051	0.073	0.096	5	14	23	32	39	41	5	41
Bulgaria	586	37	2439	0	0.311	0.311	0.078	0.078	0.078	0.078	0.078	59	64	55	50	44	40	40	64
Romania	76	4	313	0	BSS	0.341	0.076	0.076	0.076	0.076	0.076	64	62	54	49	41	39	39	64
Sweden	1204	9	1649	2	-1.897	-1.897	0.006	0.023	0.039	0.056	0.073	1	15	15	26	32	38	1	38
Latvia	174	10	442	0	0.319	0.319	0.066	0.066	0.066	0.066	0.066	61	61	51	45	38	37	37	61
Myanmar	174	7	584	0	0.071	0.071	0.065	0.065	0.065	0.065	0.065	55	59	50	43	37	36	36	59
Lithuania	345	20	918	0	0.250	0.250	0.062	0.062	0.062	0.062	0.062	57	58	48	42	36	35	35	58
Isle of Man	1685	36	2219	1	-1.401	-1.401	0.017	0.028	0.039	0.050	0.061	18	25	22	25	28	34	18	34
Bahamas	9116	217	17043	2	-1.450	-1.450	0.022	0.032	0.041	0.051	0.061	16	33	26	27	29	33	16	33
Brazil	72	4	250	0	BSS	0.372	0.061	0.061	0.061	0.061	0.061	66	57	47	41	35	32	32	66
Malaysia	1006	53	3689	0	-0.133	-0.133	0.060	0.060	0.060	0.060	0.060	49	56	45	39	34	31	31	56
Estonia	139	5	419	0	0.066	0.066	0.059	0.059	0.059	0.059	0.059	54	55	44	38	33	30	30	55
Antigua & Barbuda	7625	262	15880	1	-1.067	-1.067	0.033	0.039	0.045	0.051	0.058	30	44	34	29	30	29	29	44
Singapore	5126	80	10053	2	-1.712	-1.712	0.014	0.025	0.035	0.046	0.056	8	21	20	24	26	28	8	28
Netherlands Antilles	1129	63	3049	0	-0.047	-0.047	0.056	0.056	0.056	0.056	0.056	51	53	43	34	31	27	27	53
Denmark (DIS)	1856	33	2915	1	-1.550	-1.550	0.015	0.025	0.034	0.044	0.053	12	22	17	22	24	26	12	26
Austria	56	3	179	0	BSS	0.374	0.049	0.049	0.049	0.049	0.049	67	51	42	31	27	25	25	67
Iran	549	19	2077	0	-0.641	-0.641	0.046	0.046	0.046	0.046	0.046	41	49	40	30	25	24	24	49
Croatia	396	15	1011	0	-0.389	-0.389	0.042	0.042	0.042	0.042	0.042	45	48	36	28	23	23	23	48
Cayman Islands	1040	38	2303	0	-0.740	-0.740	0.035	0.035	0.035	0.035	0.035	38	46	30	23	22	22	22	46
Madeira	620	20	1581	0	-0.770	-0.770	0.034	0.034	0.034	0.034	0.034	36	45	29	21	21	21	21	45
United Kingdom	2643	27	3769	1	-1.867	-1.867	0.008	0.015	0.021	0.027	0.033	2	16	12	14	17	20	2	20
Irish Republic	190	6	298	0	-0.233	-0.233	0.030	0.030	0.030	0.030	0.030	48	42	25	20	20	19	19	48
Barbados	545	13	1543	0	-1.061	-1.061	0.029	0.029	0.029	0.029	0.029	31	41	24	19	19	18	18	41
Faeroe Islands	71	3	201	0	BSS	0.256	0.028	0.028	0.028	0.028	0.028	58	40	21	18	18	17	17	58
Gibraltar	1124	32	1891	0	-1.058	-1.058	0.025	0.025	0.025	0.025	0.025	32	38	18	17	16	16	16	38

	Total	Total	Total				no Z	k=0.5	k=1.0	k=1.5	k=2.0	EF	Q	Q	Q	Q	Q	best	worst
Flag	**N**	**D**	**Def**	**Z**	**EF(90)**	**EF(30)**	**Q**	**Q**	**Q**	**Q**	**Q**	**no Z**	**no Z**	**k=0.5**	**k=1.0**	**k=1.5**	**k=2.0**	**rank**	**rank**
Spain (CSR)	306	8	504	0	-0.746	-0.746	0.025	0.025	0.025	0.025	0.025	37	39	19	16	15	15	15	39
Australia	159	2	670	0	-0.942	-0.942	0.024	0.024	0.024	0.024	0.024	35	36	16	15	14	14	14	36
Netherlands	4255	94	6352	0	-1.466	-1.466	0.019	0.019	0.019	0.019	0.019	15	29	14	13	13	13	13	29
Marshall Islands	4325	82	7948	0	-1.581	-1.581	0.017	0.017	0.017	0.017	0.017	11	24	13	12	12	12	11	24
Finland	558	6	1097	0	-1.632	-1.632	0.012	0.012	0.012	0.012	0.012	10	19	11	11	11	11	10	19
Germany	7973	110	10181	0	-1.798	-1.798	0.012	0.012	0.012	0.012	0.012	4	18	10	10	10	10	4	18
France	144	1	448	0	-1.162	-1.162	0.005	0.005	0.005	0.005	0.005	27	13	9	9	9	9	9	27
Saudi Arabia	133	3	274	0	-0.338	-0.338	0.003	0.003	0.003	0.003	0.003	46	12	8	8	8	8	8	46
Belgium	322	4	424	0	-1.381	-1.381	-0.002	-0.002	-0.002	-0.002	-0.002	20	10	7	7	7	7	7	20
Switzerland	193	2	310	0	-1.207	-1.207	-0.006	-0.006	-0.006	-0.006	-0.006	26	8	6	6	6	6	6	26
French Antarctic Territory	233	2	273	0	-1.416	-1.416	-0.009	-0.009	-0.009	-0.009	-0.009	17	7	5	5	5	5	5	17
Luxembourg	160	2	210	0	-0.950	-0.950	-0.011	-0.011	-0.011	-0.011	-0.011	34	6	4	4	4	4	4	34
France (FIS)	184	0	225	0	-1.739	-1.739	-0.016	-0.016	-0.016	-0.016	-0.016	7	5	3	3	3	3	3	7
Qatar	92	1	113	0	-0.501	-0.501	-0.026	-0.026	-0.026	-0.026	-0.026	44	3	2	2	2	2	2	44
Denmark	44	0	79	0	BSS	0.031	-0.044	-0.044	-0.044	-0.044	-0.044	53	1	1	1	1	1	1	53

Note: Flags below sample size are not listed and are as follows: Seychelles, Ghana, Fiji, Mauritius, Mexico, Bahrain, Cuba, New Zealand, Nicaragua, Nigeria, Venezuela, Brunei, Namibia, Iceland, Argentina, Guyana, Cape Verde, Eritrea, Portugal, Samoa, Colombia, Sao Tome & Principe, Wallis & Futuna, Peru, Angola, Tanzania, Trinidad & Tobago, Sudan, Somalia, South Africa, Ecuador, and Kazakhstan.