A method to measure enforcement effort in shipping with incomplete information

Xichen Ji[1], Jan Brinkhuis[2], Sabine Knapp[3]

Econometric Institute Report 2014-12

Date: 4th July 2014

Abstract

Current methods used to evaluate enforcement standards of registries face some shortcomings like the inability to deal with small sample sizes, the use of biased samples and the lack of a satisfactory criterion to measure the effort to administer different fleet profiles. The effort to administer a fleet cannot be observed directly and we develop and apply a new, refined and less biased method accounting for the non-direct measurability of the '*enforcement effort*' and taking incentives into account. Application of this method to other entities than flags can be considered such as Recognized organizations or Document of Compliance Companies. To demonstrate the method, we apply it to a sample of three years of data and compare it to the method currently used, as well as one other method, with various combinations of weight factors. The results based on 99 flags demonstrate the change of the ranking of the flags, especially if compared to the method currently used. The method is flexible and transparent, and some straightforward adaptations of it, such as introducing some leniency for possible bad luck, lead to a refined formula that can solve some of the main shortcomings. In particular, it gives a measurement of the enforcement effort that appears to be as fair as the incomplete information allows. The method is not problem specific: it can be applied to all principal-agent problems where the effort of the agent into some task that the principal has set can only be observed by the principal partially, through undesirable events that are the result of chance and inadequate effort.

[1] CPB Netherlands Bureau for Economic Policy Analysis, PO Box 80510, 2508 GM, Den Haag, NL, tel: 070-3386000, email: X.Ji@cpb.nl;
[2] Econometric Institute, Erasmus University Rotterdam, PO Box 1738, 3000 DR, Rotterdam, NL, tel: 010-4081364, email: brinkhuis@ese.eur.nl
[3] Econometric Institute, Erasmus University Rotterdam, PO Box 1738, 3000 DR, Rotterdam, NL, email: knapp@ese.eur.nl;

Keywords:

Performance measurement, small sample sizes, inspections, detentions, incidents, measuring effort under incomplete information, accounting for bad luck, incentives.

Disclaimer for Ji and Knapp:

The views expressed in this article represent those of the authors and do not necessarily represent those of the Australian Maritime Safety Authority (AMSA) and the Netherlands Bureau for Economic Policy Analysis

Authors:

Xichen Ji, CPB Netherlands Bureau for Economic Policy Analysis, PO Box 80510, 2508 GM, Den Haag, NL, tel: 070-3386000, email: X.Ji@cpb.nl;

Jan Brinkhuis, Econometric Institute, Erasmus University Rotterdam, PO Box 1738, 3000 DR, Rotterdam, NL, tel: 010-4081364, email brinkhuis@ese.eur.nl

Sabine Knapp, Econometric Institute, Erasmus University Rotterdam, PO Box 1738, 3000 DR, Rotterdam, NL, email: knapp@ese.eur.nl

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. Enforcement of internationally agreed standards is the obligation of a registry or flag. It is not applied with equal force, and this creates opportunities for substandard shipping. 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[4] and the Tokyo MoU[5]) publish each year a list that rank flags according to their performance during inspections, the so called Black/Grey/White List (BGW-list), where black listed flags perform worst. The list has become the industry standard and is seen as a proxy to measure the effort (or quality) of enforcement of a registry, despite the shortcomings of this list, which will be discussed in the next paragraphs. The list is also used by the Paris MoU and the Tokyo MoU to target ships for inspection.

Other relevant developments dealing with performance measurement are taking place at the IMO. For example, Assembly Resolution A.1037(27) (IMO, 2013) asks for the development of performance indicators that can measure progress made towards its broad strategic directions including improving the safety of the industry. One of these directions deals with fostering global compliance, but this is not measured at the individual member state level. However, the IMO has established the Member State Audit scheme, where individual audits are performed, primarily based on qualitative measures.

Perepelkin et al. (2010) highlights some of the shortcomings of the current method to calculate the BGW-list and offers some solutions. There is the inability of the current method to deal with small sample sizes as well as with large numbers of inspections; these can also lead to complications since these can make the grey list so narrow that it becomes incomparably smaller than the black and white ones. The list is also characterized by the lack of a satisfactory criterion for the effort of a flag. Indeed, the criterion is defined in terms of the excess factor, the value of which depends on the BGW-list and for each of the three, black/grey/white, it is defined by a

[4] The Paris MoU covers the EU, parts of Canada and the Russian Federation
[5] The Tokyo MoU covers Asia, Australia, Chile and parts of the Russian Federation

different procedure (Perepelkin et al, 2010). Performance is currently only measured based on inspection outcomes such as detentions. Perepelkin et al. (2010) have considered, besides detention data, deficiencies and incident information. In principle, other factors might also be relevant, such as the age of the vessel, the sizes or the ship type, as these have an influence on the safety quality of ships (Bijwaard and Knapp, 2009). For ship types, the reason for this is that the major shipping markets have different characteristics.

Given this situation, this article builds on some aspects of the method developed by Perepelkin et al. (2010), and in particular it tries to address the lack of any common criterion that depicts the effort of a flag. We introduce a concept where we measure the effort of a flag in enforcing the internationally agreed standards. Effort is however not directly observable and we therefore propose an indirect measure that reflects differences of fleet profiles of a registry. 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 flag 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, we introduce `sympathy' into the measure, giving each flag the benefit of the doubt, but not more. Registries with smaller fleets get more sympathy, as desired.

In addition, the current method also suffers from the lack of coordination amongst PSC regimes to use combined data. From Knapp (2006), Knapp and Franses (2007) and Bijwaard and Knapp (2009), it is however evident that ships get inspected in more than one PSC regime each year and that the data from various regimes should be combined in order to decrease the sample bias (currently, each PSC regime only uses data of their own inspections) and improve targeting of ships. This suggests using all relevant data on a vessel. The combination of data also allows the evaluation of more registries. Incidents were considered by Perepelkin et al. (2010) but we extend this application to two degrees of seriousness – very serious and serious incidents.

The method that we propose is not restricted to the use of registries but could be extended to other organizations such as recognized organizations (RO) or to companies that are responsible for the safety management of a vessel (Document of Compliance Companies). In fact it can be applied to any situation where one party, called a principal, wants other parties, called agents, to put in an adequate effort into some tasks, but where the principal cannot directly observe the effort, but only certain undesirable events that must be ascribed to a mixture of chance and inadequate effort.

To begin with, we present the precise motivation for our method by following the development of the proposed new formula. We then apply two variants of the method, one of which takes into consideration serious accidents along with very serious accidents and detentions, and compare it to the current method and to the method developed by Perepelkin et al. (2010). The article closes with a discussion, and our conclusions and recommendations for the policy makers. Due to the political sensitivity of the topic at hand, we do not name registries when applying and comparing the methods since we are not interested in producing a ‘name and shame’ list similar to what is currently used as the Black/Grey/White list. We are interested in demonstrating the usefulness of the method and show how the ranks of the flags change if various methods are applied. To the best of our knowledge, there is no possible source of unfairness left in our most refined formula; if however an unfairness might be discovered, then we believe, based on our experience so far with the method, that it will be possible to devise a variant that removes this unfairness.

2. Proposed formula

2.1. A crude measure for the performance of a flag

In order for a method to be accepted, it is important that it is transparent to all parties involved how it works and that it is fair to all flags. Therefore, we explain carefully how the proposed method is based on a simple, clear and convincing idea. Then we show how a number of shortcomings of this idea can be overcome by means of straightforward adaptations and how this leads to the formula that we propose for measuring the effort of a flag. In order for a method to be moreover effective, it should take the incentives of all parties involved into consideration. Therefore, we address incentives as well.

The starting point is the desirability of optimal safety standards of the shipping industry. To this end, it is the task of each registry to enforce the internationally agreed standards. The reality is that the effort of enforcement is not the same for all flags and that it cannot be monitored directly. The challenge is to find a policy that gives registries some additional incentives to increase the effort to comply. The idea is to base such a policy on an indirect measurement or proxy for the effort. Such a measurement can enhance targeting for certain flags for inspections which is currently only based on detentions. The indirect measurement to be chosen is suggested by the following observation. 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. This simple idea is the base of the proposed method.

This idea turns out to be very flexible and we will see that by straightforward adaptations one can modify it in several ways, and as a result we will get a refined formula that does not have the shortcomings of the current method. To begin with, of course, one cannot just count the total number of all undesirable events for a flag, detentions and very serious accidents. In order to be fair, one has to take into consideration, for each flag, its total number of inspections, its fleet size and the fact that both types of undesirable events do not have equal weight. Finally, we take the varying market characteristics of fleet profiles into account by distinguishing between ship types.

This leads to 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, we get 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. Then we compare the effort of two flags, 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, we consider that the effort of F_2 is at least as good as that of F_1. We want to extend this idea in order to be able to compare the effort for each pair of flags. To this end, we introduce a weight factor c, to be chosen by policy makers. Then we 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, we take as a first attempt the following formula for measuring the performance of a flag F:

$$Q_F' = d_F + cz_F \qquad \text{`crude performance measure' (1)}$$

The lower this number is, the better the effort of the flag to enforce standards. This measure is a combination of inspections, detentions, very serious incidents and fleet size. We note that for a flag with an average number of detentions and very serious accidents, the performance measure is $1+c$. There are two other types of undesirable events that could be considered as well. There is the deficiency information from PSC inspections, and there is the number of serious accidents. Deficiency information is used in the Perepelkin et al. (2010) method. We do not use deficiency information, since we believe that detention information covers the main aspects. Moreover, using deficiency information would add complexity, since policy makers need to choose weight factors to distinguish between the level of importance of deficiencies and these weight factors are not easy to determine. There can be many types of deficiencies, which also need to be grouped. In most cases, it is the combination of deficiency types that will determine the level of non-compliance and if these deficiencies are of very serious nature, then this is reflected in the fact that the ship is detained, hence detention can be used to summarize seriousness of deficiencies. Furthermore, discovering some deficiencies can be seen as a desirable event in the sense that then these will be dealt with and safety quality will increase.

Besides detention, we use serious incidents along with very serious incidents already used by Perepelkin et al. (2010), We offer this variant of the method as the one above takes into account only extreme undesirable events: a relatively `mild' one, detentions, and one that should weigh much more, a very serious accident. Serious accidents are on the middle level. Our implementation does not show many differences in the outcomes, but the fact that it uses more data suggests a greater reliability. Therefore it seems slightly preferable to implement the method taking serious accidents into account as well. One could argue that at this moment, the population of serious incidents is incomplete since reporting to the IMO is biased. This however can change in the future and with better data population, the inclusion of serious incidents into the formula will account better for the effort since more observations are available compared to very serious incidents. Our proposed method only requires one weight factor for each incident category, to be determined by the policy makers: the weight factor for a very serious incident and the weight factor for a serious incident relative to a detention. As such we deal with only three levels of undesired outcomes and so the determination of the weight factors will be easier for policy makers compared to many different types of deficiencies. To keep formulas simple, we will not always display serious incidents in the formulas.

2.2. A finer measure for the performance of a flag taking ship types into account

As mentioned earlier, in order to better quantify the effort of a flag, we feel that it is best to distinguish between ship types. For the classification of ship types, we refer to Knapp (2006) and use five main ship types as follows: 1) general cargo, 2) dry bulk carrier, 3) container vessel, 4) tankers, 5) passenger vessels and 6) all other ship types. One reason for taking ship type into account is that ship type can be used as a proxy for the characteristics of the market a particular vessel trades in. These characteristics are determined by the nature of the trade flows, the legislative framework and the economic pressures (Bijwaard and Knapp, 2009, Heij and Knapp, 2014) and are relevant for measuring the performance of a flag. For example, tankers or some dry bulk carriers are subject to industry vetting inspections besides port state control inspections. Another example is that container vessels operate under regular liner trades and carry higher value cargo with better safety quality management. Other reasons for taking ship type into account are that ship types can also be used as a proxy for other factors such as age or size (refer to Table 1).

For example, some registries administer a fleet of older ships of high risk ship types such as general cargo vessels (Knapp, 2006, Knapp et al. 2013) which tend to engage in more regional trade compared to for instance large tankers, container vessels or dry bulk carriers. Monitoring a fleet of older ships engaged in local trade is more challenging than monitoring for instance a fleet of young tankers. Nevertheless, the same effort will lead on average (refer to Table 1) to a higher detention rate. Therefore, in order to be fair, the different ship types have to be taken into account.

Table 1: Descriptive statistics ship types (2006 to 2008)

Ship type	Age Mean	GRT Mean	Deficiencies Mean	Detention rate	Incident rate (very serious)
General cargo	18.7	9,326	4.08	7.0%	0.0031
Dry bulk carriers	14.4	31,462	2.82	4.0%	0.0021
Container ships	9.7	33,885	1.79	1.8%	0.0020
Tankers	10.1	29,959	1.85	2.3%	0.0009
Passenger ships	18.9	33,626	3.00	2.4%	0.0020
Other ship types	20.8	4,315	3.96	8.1%	0.0099

Now we take the crude formula $Q_F' = d_F + cz_F$ for measuring the performance of a flag F and make the first improvement. We make one small change only: we take the different ship types into account. This gives the following improved formula for measuring the performance of a flag F:

$$Q_F = \frac{1}{N_F}\sum_{t\epsilon F}(\alpha_t * D_{t,F}) + \frac{c}{N_{ships\ F}}\sum_{t\epsilon F}(\beta_t * Z_{t,F}) \qquad \text{`finer performance measure' (2)}$$

where:

F: a flag,

Q_F: the performance measure of flag F;

N_F: the number of inspections of ships under flag F during the period under consideration;

$N_{ships\ F}$: the number of ships under flag F, averaged over the period under consideration;

c: a positive constant, to be chosen by policymakers, that gives the weight of a very serious casualty compared to a detention;

t : a type of ship, determined by age and tonnage group;

$t\epsilon F$: shorthand notation for `the types of ship that occur among the ships under flag F';

$D_{t,F}$: the number of detentions of ships of type t under flag F during the period under consideration;

$Z_{t,F}$: the number of very serious incidents of ships of type t under flag F during the period under consideration;

The coefficients α_t and β_t in the formula above are calculated with the following formulas:

$\alpha_t = \frac{N_t}{D_t}$ provided D_t is not zero, where D_t is the number of detentions of ships of type t of all flags during the period under consideration, and N_t is the number of inspections of ships of type t of all flags during the period under consideration; if $D_t = 0$, then we put $\alpha_t = 0$, for example (it does not matter what we put here, as in the summation α_t is multiplied with $D_{t,F}$, which is zero if $D_t = 0$).

$\beta_t = \frac{N_{ships_t}}{Z_t}$ if Z_t is not zero, where N_{ships_t} is the number of ships of type t of all flags during the period under consideration, and where Z_t is the number of very serious incidents of ships of type t for all flags, during the period under consideration; if $Z_t = 0$, then we put $\beta_t = 0$, for example (with a similar justification as given above for α_t). Including serious incidents as well, we obtain the following formula:

$$Q_F = \frac{1}{N_F}\sum_{t\epsilon F}(\alpha_t * D_{t,F}) + \frac{c}{N_{ships\ F}}\sum_{t\epsilon F}(\beta_t * Z_{t,F}) + \frac{d}{N_{ships\ F}}\sum_{t\epsilon F}(\Upsilon_t * S_{t,F}) \qquad (3)$$

where:

d: a positive constant, to be chosen by policymakers, that gives the weight of a serious casualty compared to a detention;

$\Upsilon_t = \frac{N_{ships_t}}{S}$ if S_t is not zero, where S_t is the number of serious incidents of ships of type t for all flags, during the period under consideration; if $S_t = 0$, then we put $\Upsilon_t = 0$, for example (with a similar justification as given above for α_t).

$S_{t,F}$: the number of serious incidents of ships of type t under flag F during the period under consideration.

2.3. Precise motivation for the finer performance measure

The reason for the chosen correction for detentions is as follows: without corrections for ship types, we would take for the contribution of the detentions to the measure of the performance of flag *F*, the ratio $\frac{D_F}{N_F}$, where D_F is the number of detentions of ships under flag *F*. This can be written as $\frac{1}{N_F}\sum_{t \epsilon F} D_{t,F}$. To make the numbers of detentions comparable between different types, it is reasonable to multiply, for each type *t*, the term $D_{t,F}$ by $\alpha_t = \frac{N_t}{D_t}$, the average number of inspections for one detention for ships of type *t*. This gives the contribution $\frac{1}{N_F}\sum_{t \epsilon F}(\alpha_t * D_{t,F})$ to the measure of the enforcement effort of flag *F*. In particular, this will make the contribution of the detentions of old ships smaller, as desired. The reason for the chosen correction for very serious (and serious) incidents is the same. In particular, for a flag that has an average number of detentions and very serious incidents, the enforcement effort will be 1+*c*. For the variant that takes serious incidents into account, this is 1+*c*+*d*.

2.4. Our final measure for the performance of a flag, taking the difference in variations in the observations for small and large fleets into account

The finer measure of the performance of a flag given above, Q_F, is not an adequate measure for the performance of flags with a small fleet. The numbers $D_{t,F}$ and $Z_{t,F}$ are subject to chance, they are stochasts, and for small fleet the effect of chance on the outcomes (for example, `bad luck'), and so on the outcome of the formula, can be unacceptably large. Ideally we would like to replace in the formula the stochasts by their expectations, but these are unknown. Therefore, we will give a systematic method to replace the numbers $D_{t,F}$ and $Z_{t,F}$ by numbers that are slightly smaller, just enough to make sure, within a certain precision, that these numbers are smaller than the expectations of these stochasts. This makes the outcome of the formula smaller, and so a flag is certain not to be given by bad luck a performance measure Q_F that is higher than deserved; to be more precise: it is very unlikely that this will happen. That is, the qualities of the flags are attributed with some sympathy, `the benefit of the doubt'. We will see that this systematic way accords more sympathy to flags with a small fleet than to flags with a larger fleet, as desired. Thus the shortcoming of the finer performance measure above will have been repaired.

2.5. Analysis of the variations in the observations

Next we provide an analysis for the variation in the observations and for this we would like to introduce the following additional notation: $N_{t,F}$, the number of inspections of ships of type t under flag F during the period under consideration ($\sum_{t\epsilon F}(N_{t,F}) = N_F$), and ${N_{ships}}_{t,F}$, the total number of ships of type t under flag F , averaged over the period under consideration ($\sum_{t\epsilon F}({N_{ships}}_{t,F}) = {N_{ships}}_F$). Assume that the number of detentions of a certain ship type under a certain flag follows a binomial distribution: $D_{t,F} \sim Bin\left(N_{t,F}, {p^d}_{t,F}\right)$, with ${p^d}_{t,F}$ the underlying probability of detention at one inspection of ship type t under the flag F. Moreover, we assume that the number of very serious incidents of ship type t under flag F also follows a binomial distribution $Z_{t,F} \sim Bin({N_{ships}}_{t,F}, {p^z}_{t,F})$, with ${p^z}_{t,F}$ the underlying probability of a very serious casualty for each ship of type t under flag F. The probabilities of detention or very serious casualties of one ship type differ among the flags because of the differences in management among the flags, the very effect we want to measure. When $N_{t,F}$ is large enough, the distribution of the stochast $D_{t,F}$ approximates the normal distribution with $E(D_{t,F}) = N_{t,F} * {p^d}_{t,F}$ and $Var(D_{t,F}) = N_{t,F} * {p^d}_{t,F} * (1 - {p^d}_{t,F})$.

The same holds for the distribution of stochast $Z_{t,F}$: when ${N_{ships}}_{t,F}$ is large enough, the distribution of the stochast $Z_{t,F}$ approximates a normal distribution, with $E(Z_{t,F}) = {N_{ships}}_{t,F} * {p^z}_{t,F}$ and $Var(Z_{t,F}) = {N_{ships}}_{t,F} * {p^z}_{t,F} * (1 - {p^z}_{t,F})$. It will be convenient to write $({p^d}_{t,F})' = \frac{D_{t,F}}{N_{t,F}}$ and to view this as an observation of a normally distributed stochast with mean ${p^d}_{t,F}$ and variation $\frac{{p^d}_{t,F}(1-{p^d}_{t,F})}{N_{t,F}}$. Similarly for $\left({p^z}_{t,F}\right)' = \frac{Z_{t,F}}{{N_{ships}}_{t,F}}$.

Ideally we would like to take as a measure for the performance of a flag:

$$Q_F{}^* = \frac{1}{N_F}\sum_{t\epsilon F}(\alpha_t * N_{t,F} * {p^d}_{t,F}) + \frac{C}{{N_{ships}}_F}\sum_{t\epsilon F}(\beta_t * {N_{ships}}_{t,F} * {p^z}_{t,F}) \qquad (4)$$

This formula has been obtained from the one for Q_F by replacing $D_{t,F} = N_{t,F} * ({p^d}_{t,F})'$ by $N_{t,F} * {p^d}_{t,F}$, and by replacing $Z_{t,F} = {N_{ships}}_{t,F} * ({p^z}_{t,F})'$ by ${N_{ships}}_{t,F} * {p^z}_{t,F}$. Unfortunately, we cannot observe the underlying probabilities. Therefore, we take, instead systematic `lower bounds' for these probabilities. For this we use the method from Perepelkin et al. (2010).

Now we show how to get in a systematic way these `lower bounds'. Let p' be the observed value of the stochast. Thus, p stands for ${p^d}_{t,F}$ respectively ${p^z}_{t,F}$, and p' stands for $\frac{D_{t,F}}{N_{t,F}}$ respectively $\frac{Z_{t,F}}{{N_{ships}}_{t,F}}$. The standard deviation is $\sigma = \sqrt{\frac{p(1-p)}{N}}$. Let us fix a confidence level a (for example, $a = 0.95$ is a popular choice) and define for each p the confidence interval $(p - c_a, p + c_a)$ such that a value p' of the stochast is contained in this interval with probability a:

$$\Pr[|p' - p| < c_a] = a.$$

The meaning of this interval is as follows. *The values of p that satisfy this condition are the values of p for which the hypothesis that p' lies in the confidence interval of the normal distribution with mean p and standard deviation σ cannot be rejected with confidence level a* . To put it more simply, but not entirely precisely: we are `sure' that the true value of the underlying probability $p$ does not differ more than $c_a$ from the observed value $p'$ (here `sure' means roughly that the probability that this statement is incorrect for given p equals 1- a). We take $L(p)$, the smallest value of p that satisfies this condition as our `lower bound' of p. The reason for this terminology is that for given p we are confident that $L(p)$is smaller than p, given the prescribed confidence level a. Because of the fact that p' is normally distributed, it holds that:

$$\Pr[|p' - p| < c_a] = 2 * \Phi_{\frac{c_a}{\sigma}} - 1,$$

where Φ is the cumulative function of the standard normal distribution and $\Phi_{\frac{c_a}{\sigma}}$ gives the cumulative probability for $Z \leq \frac{c_a}{\sigma}$ and $Z \sim Norm(0, 1)$.

It follows that:

$$|p' - p| < \Phi^{-1}(\tfrac{1+a}{2})\, \sigma.$$

Then, we have:

$|p' - p| < t_a\sqrt{\frac{p(1-p)}{N}}$ with $t_a = \Phi^{-1}(\frac{1+a}{2})$

$(p' - p)^2 < {t_a}^2 * \frac{p(1-p)}{N}$

${p'}^2 - 2p'p + p^2 < \frac{{t_a}^2}{N}p - \frac{{t_a}^2}{N}p^2$

$\left(1 + \frac{{t_a}^2}{N}\right)p^2 + \left(-2p' - \frac{{t_a}^2}{N}\right)p + {p'}^2 < 0.$

This is an inequality in a quadratic polynomial in p. As $1 + \frac{{t_a}^2}{N}$, the coefficient of p^2, is positive, the solutions form an open interval with endpoints the roots of the quadratic equation

$\left(1 + \frac{{t_a}^2}{N}\right)p^2 + \left(-2p' - \frac{{t_a}^2}{N}\right)p + {p'}^2 = 0.$

Then, the abc-formule can be used to get the `lower bound' for p:

$$L(p) = \frac{-\left(-2p' - \frac{{t_a}^2}{N}\right) - \sqrt{(2p' + \frac{{t_a}^2}{N})^2 - 4\left(1 + \frac{{t_a}^2}{N}\right)\left({p'}^2\right)}}{2(1 + \frac{{t_a}^2}{N})}.$$

The other root is the ` upper bound', denoted by $U(p)$. For flags with small fleet, the variation in the p' is larger. Indeed, a good measure for this variation is the difference $U(p) - L(p)$. As the denominator in the expression $L(p)$ (and $U(p)$) is approximately 2, we see that $(U(p) - L(p))^2$ is approximately $(2p' + \frac{{t_a}^2}{N})^2 - 4\left(1 + \frac{{t_a}^2}{N}\right)\left({p'}^2\right)$. This can be rewritten as $\frac{{t_a}^4}{N^2} + \frac{2p' t_a}{N}(t_a - p')$. As $t_a \approx 2$ (usually) and $p' \in (0,1)$, we see that this is decreasing in N. Thus, for flags with small fleet, the variation is large.

Now we apply this formula to the probabilities of detention and of very serious incidents. After simplification, we get the promised `lower bounds' of probabilities of detention and very serious incidents for different ship type and different flags:

$$L\left({p^d}_{t,F}\right) = \frac{\frac{D_{t,F}}{N_{t,F}} + \frac{1}{2} * \frac{{t_a}^2}{N_{t,F}} - t_a\sqrt{\left(\frac{D_{t,F}\left(N_{t,F} - D_{t,F}\right)}{{N_{t,F}}^3}\right) - \frac{1}{4} * \frac{{t_a}^2}{{N_{t,F}}^2}}}{1 + \frac{{t_a}^2}{N_{t,F}}}$$

and

$$L\left({p^z}_{t,F}\right) = \frac{\frac{Z_{t,F}}{{N_{ships}}_{t,F}} + \frac{1}{2} * \frac{{t_a}^2}{{N_{ships}}_{t,F}} - t_a \sqrt{\left(\frac{Z_{t,F}\left({N_{ships}}_{t,F} - Z_{t,F}\right)}{{{N_{ships}}_{t,F}}^3}\right) - \frac{1}{4} * \frac{{t_a}^2}{{{N_{ships}}_{t,F}}^2}}}{1 + \frac{{t_a}^2}{{N_{ships}}_{t,F}}}.$$

Now we can define the corrected measure for the performance measure of a flag. We replace the observed value of each stochast $\left({p^d}_{t,F}\right)'$ (and $\left({p^z}_{t,F}\right)'$) by the `lower bound' $L\left({p^d}_{t,F}\right)$ (and $L\left({p^z}_{t,F}\right)$) for the mean value of the stochast, as defined above.

$$L(Q_F) = \frac{1}{N_F} \sum_{t \epsilon F} (\alpha_t * N_{t,F} * \ L({p^d}_{t,F})) + \ \frac{c}{{N_{ships}}_F} \sum_{t \epsilon F} (\beta_t * {N_{ships}}_{t,F} * \ L({p^z}_{t,F})). \qquad (5)$$

We take this `lower bound' for Q_F as our final performance measure for a flag F. Now both shortcomings of the crude performance measure $Q_F{}' = d_F + c z_F$ have been overcome: this measure takes into account ship types and it takes into account the variations in the observations. We do not display the formula for the variant of the method where also the serious incidents are taken into account, as the changes are straightforward.

In applying the proposed method, we have replaced the terminology of Black/Grey/White by the following groupings of the performance measure of a flag viewed as proxy for its effort to administer its fleet.

- *the worst quartile according to Q-ranking is called high risk (proxy to low effort); our reasons: this corresponds according to experts reasonably well to substandard performance of a flag;*
- *the second worst quartile is called medium risk (proxy to medium effort);*
- *the best two quartiles are called low risk (proxy to high effort).*

We propose that the values are determined every three years based on data of a three year period in order to give a flag administration the opportunity to demonstrate improvement. The next section will apply this method and compare the results to the currently used method as well as to the method developed by Perepelkin et al. (2010). In the future, a five year time frame could also be considered compared to three years, in order to use more data and assess more flags.

3. Application of methods and discussion of results

We apply our method using inspection (detention) data, incident data and data on the world fleet for the time period 2006 to 2008. We base our analysis on data from Perepelkin et al. (2010) with additional data to extend the time period. We end up using roughly 183 thousand inspections and 8,646 detentions from various PSC regimes or individual countries (Paris MoU, USCG, Indian Ocean MoU, Vina del Mar Agreement, AMSA) although inspection data are not available for each of these regimes for the entire time period.

We further use incident data based on Knapp (2013) from four sources (IHS-Fairplay, Lloyds Maritime Intelligence Unit, International Maritime Organization and the Australian Maritime Safety Authority). For the classification of seriousness, we use the IMO definitions (IMO, 2000) and consider very serious (524 observations) and serious (3,883 observations) incidents. The incident data needed to be manually reclassified in order to ensure compatibility of the four sources. Fleet data was not entirely available for the entire time period by major ship types so we ended up estimating the number of ships for each ship type used (general cargo, dry bulk, container, tanker, passenger vessels and other ship types) and for each flag, based on data from IHS-Fairplay. The following input data are needed to apply the proposed method:

- *Total number of inspections by ship type and flag*
- *Total number of detentions by ship type and flag*
- *Total number of very serious and total number of serious incidents by ship type and flag*
- *Total number of vessels in service by ship type for each flag*
- *A weight factor for c for very serious incidents and a weight factor d for serious incident, relative to detention, to be determined by policy makers*

As mentioned earlier, we will not name individual flags due to the political nature of the subject matter. Our interest lies in applying the proposed method and in demonstrating how the ranks of the flags change by applying the different methods rather than producing a list to 'name and shame' of registries. More detailed results can be made available from the authors upon request.

We compare three methods –the current excess factor method (EF) used by the Paris MoU, the method proposed by Perepelkin et al. (2010) and our proposed method with two variations (very serious incidents only, and both incident categories). In order to be able to make comparisons across the methods, we look at registries that can be compared across all methods and we need to choose a minimum sample size of 30 inspections, which is what is currently used for the EF method. From a total of 132 flags, we can evaluate 99 flags across all three methods. There are many different variations one could compare but for the sake of demonstration, we consider the following four combinations:

- *Current excess factor method where only detentions are considered*
- *Method based on Perepelkin et al.2010 where detentions and very serious incidents are considered with weight factors c=4 and c=5 respectively*
- *Our proposed method with information by ship type for detentions and very serious incidents with weight factors c=4 and c=5 respectively*
- *Our proposed method with information by ship type for detentions, very serious and serious incidents with weight factors c=4 and d=2 and for c=5 and d= 3 for very serious/serious incidents respectively*

We apply the combinations above and rank the flags from best to worst (1 meaning best and 99 meaning worst rank or least effort). As first comparison, we compare the EF method with each of the combinations and different weight factors *c (very serious incidents) and d (serious incidents)* to see how the ranks change and a series of graphs are presented in Appendix A for easy comparison. We can observe some agreement compared to the method in current use but also large changes in rank for a number of flags. The change in ranks reflects the effect of incorporating incidents (either serious or very serious) compared to detentions or deficiencies only. We identify at least 20 flags with change of up to 50 ranks.

The change of ranks of our proposed method compared to the EF method can further be explained by the incorporation of the ship types, which accounts for the differences in the fleet profiles. We provide more `sympathy' for maritime administrations that register for instance older general cargo ships or have varying fleet profiles in general. To provide more insight into this, Table 2 provides the values for the three parameters – *alpha, beta and gamma* for each ship data based on the sample data used to apply our method (given in equation 3 earlier).

The parameters are weight factors for detentions (α_t), very serious incidents (β_t) and serious incidents (γ_t) for a certain ship type *t*. The larger the weight factor α_t, the smaller is the probability for the ship type to be detained. This also holds for serious and very serious incidents. Due to the fact that very serious incidents are rare events, values of β_t are in average much larger than the other parameters. For small weight factor values, the proposed method will provide more sympathy to the registry since it will be more challenging to administer an older fleet (eg. general cargo) trading in coastal areas compared to vessels that trade internationally on either set routes (e.g. container trade, dry bulk trades) or are inspected more often (e.g. tankers).

Table 2: Weight factors for leniency by ship type

Ship type	α_t	β_t	γ_t
general cargo	14.45	328.78	54.14
dry bulk carriers	24.87	474.47	48.98
container ships	56.79	602.54	42.36
tankers	44.21	1052.95	89.80
passenger ships	44.12	483.46	40.51
other ship types	20.25	95.90	21.12

Interesting to observe from Appendix A is that we can find not much difference in the effect of the chosen weight factors for either very serious incidents ($c=4$ *or* $c=5$) or serious incidents ($d=2$, $d=3$). We only investigated a change of one unit for each incident type but perhaps larger increases for both weights compared to detention might be more appropriate. This could be a further topic for research in the future. For now, we suggest that policy makers should provide these weight factors based on expert knowledge. We will not investigate this further here but choose two weight factors, namely $c=4$ for very serious and $d=2$ for serious incidents and compare the ranks of these two combinations with the current excess factor method and with each other.

The top-left picture of Figure 1 for instance, gives the points (*x_f*, *y_f*) for all flags *f*, where *x_f* is the rank according to the excess factor and *y_f* is the rank according to the method of Perepelkin with weight factor *c=f* (very serious incidents only). The picture below presents the same but compared to our proposed method with weight factor $c=4$ (very serious incidents) and $d=2$ (serious incidents).

The proposed method shows more variability compared to the EF method than the method of Perepelkin et al. (2010), which uses only very serious incidents and which does not make any distinction between ship types. One can also notice that there is less change in the new method compared to Perepelkin et al. (2010).

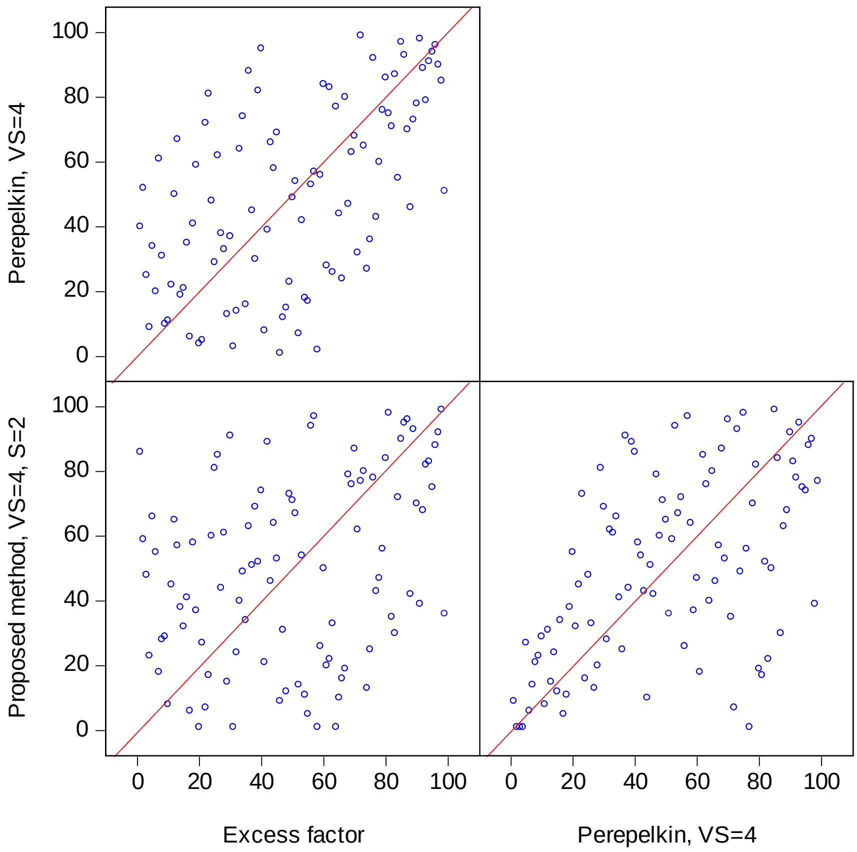


Figure 1: Comparison of ranks with various methods, VS=very serious, S=serious incidents

Next, we are interested to see how the risk categories change and we classify the flags according to their risk levels as mentioned earlier, where the worst quartile of Q is classified as *high risk* (proxy to low effort), the second worst quartile is *medium risk* (proxy to medium effort) and the remaining two quartiles are *low risk* flags (proxy to high effort). Appendix B provides selected results for the different combinations and weight factors with the changes of the risk categories for all flags that were evaluated. More detailed results can be obtained from the authors upon request. Not surprisingly by now, given Figure 1, one can observe many shifts from the high risk category based on the EF method to the medium or low risk category based on either Perepelkin et al. (2010) or our proposed method. When including serious incidents in our proposed method, one can further observe shifts of some flags from the low risk category to the medium risk category and vice versa, compared to the method by Perepelkin et al. (2010).

Some of the shifts from high risk or medium to low risk can be explained by a combination of ship types – so maritime administrations with general cargo vessels were given more sympathy. Shifts from the low risk category to the medium risk category can be explained by very serious incidents on tankers and passenger vessels, and therefore receive less sympathy since in general it will take less effort to administer these ship types. Other shifts are simply due to the addition of serious incidents. Nine flags do not change across the methods and remain in the ‘high risk’ category irrespective of the method applied. These flags perform consistently worse even though their fleet profile is strongly characterized by a high proportion of general cargo vessels or in one case a combination of general cargo vessels and tankers. However, even if sympathy is given, these flags show the least effort in enforcing international standards.

5. Conclusions and recommendations

We propose an updated method to measure the effort of a flag to administer its fleet given that some registries have varying fleet profiles or take vessels into their registry that might be more challenging to administer (e.g. older general cargo vessels, smaller fleet profiles, etc.). We then apply this method and compare the changes of the ranks across three methods and demonstrate how the ranks of some flags change considerably. The advantages of the new method over the other methods applied here are as follows. It is based on one extremely simple and convincing idea: counting undesirable events and using this as a proxy for the effort of a flag, which cannot be measured directly. There is no need for policy makers to determine the many weight factors for various types of deficiencies as only two weight factors are needed. In this way, policy makers have the possibility to fine-tune the method by choosing one (or two) weight factors. More information is used compared to the method currently used and so the sample bias is smaller and more accurately reflects reality.

Distinction of ship types can be taken into account, and this is strongly desirable because of the impact of different market characteristics on safety. Some sympathy is given to registries with more challenging fleet profiles and/or small fleet; the latter are more susceptible to bad luck. For more challenging fleet types, this is done by distinguishing ship types. For small fleet, this is done as follows. Some sympathy is given to each flag to account for possible bad luck; this is done in a systematic way and the sympathy given is just enough to account for possible bad luck of flags; this sympathy is greater for small flags, as is desired: these flags are more susceptible to bad luck. Although not demonstrated here, the proposed method can also evaluate flags with smaller sample sizes (below 30).

The proposed method of counting undesirable events is very flexible and easy to implement practically. The flexibility is demonstrated by giving straightforward adaptations from the basic counting idea that take into account all issues that arise in practice and that lead to the shortcomings of the current excess factor method, partly addressed by the method developed by Perepelkin et al. (2010).

We recommend to policy makers to use the following undesirable events as proxy to determine the effort in enforcing international standards: detentions, serious accidents and very serious incidents. With respect to the use of serious incidents, the method will become more precise once better data is being populated by the IMO via the Global Integrated Ship Information System (GISIS). We also feel that a change in terminology from the current division of Black/Grey/White into High/Medium/Low Risk, where the first two groups are the worst two quartiles, is more appropriate.

Furthermore, we recommend to fix the boundaries between these three groups for three years, based on data from the last three years (possibly use five years' worth of data in the future), and to give flags the opportunity to move upward, especially from High Risk to Medium Risk, in order to reduce occurrence of substandard ships and improve overall safety.

Finally, we recommend that the method is implemented in such a way that no name-and-shame effects can arise for flags from changes in ranks compared to the method currently in use.

Acknowledgements

We would like to thank our data providers for the provision of incident and fleet data, in particular IHS-Fairplay, LMIU and the International Maritime Organization.

References

Bijwaard G and Knapp S (2009), Analysis of Ship Life Cycles – The Impact of Economic Cycles and Ship Inspections, Marine Policy, volume 33, pp. 350-369

Heij C, Knapp S (2014), Dynamics in the dry bulk market: Economic activity, trade flows, and safety in shipping, Journal of Transport Economics and Policy, Volume 48, pages 499-514

International Maritime Organization (2000), MSC/Circ. 953, MEPC/Circ. 372, Reports on Marine Casualties and Incidents, Revised harmonized reporting procedures, adopted 14th December 2000, London

International Maritime Organization (2013),Assembly Resolution A.1037(27), Strategic Plan for the Organization (for the six-year period 2012-2017), adopted 22nd November 2011, London

Knapp, S (2006) , The Econometrics of Maritime Safety – Recommendations to enhance safety at sea, Doctoral Thesis, Erasmus University, Rotterdam

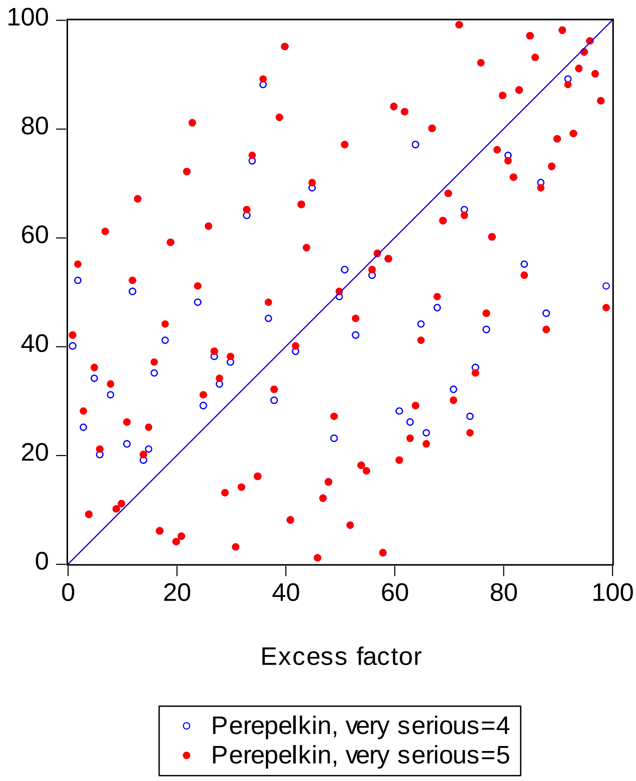
Knapp S, Franses PH (2007) A global view on port state control - econometric analysis of the differences across port state control regimes, Maritime Policy and Management, 34(5), pages 453-483

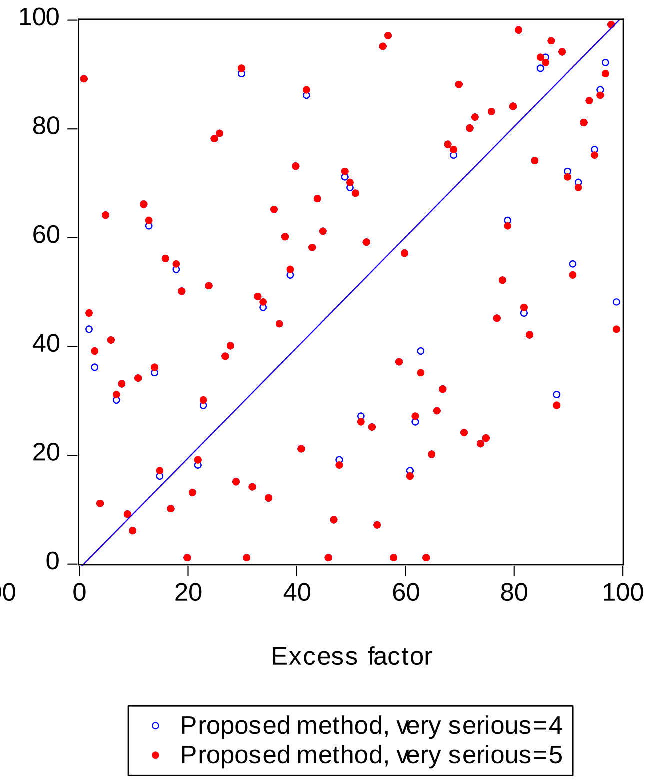
Knapp S (2013), An integrated integrated risk estimation methodology: Ship specific incident type risk, EI report 2013-11, http://repub.eur.nl/res/pub/39596/

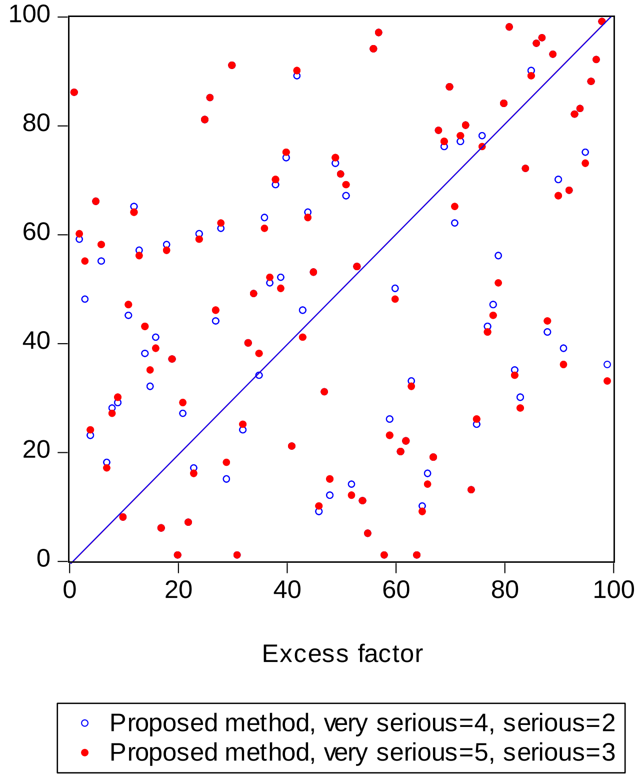
Knapp A, Heij C, Henderson R, Kleverlaan E (2013), Ship incident risk in the areas of Tubbataha and Banc d'Arguin: A case for designation as Particular Sensitive Sea Area?

Perepelkin M, Knapp S, Perepelkin G and de Pooter M (2010), A method to measure flag performance for the shipping industry, Marine Policy, Volume 34(3), pages 395-405

Appendix A – Change of ranks for each method and weight factor







Appendix B – Selected results by flag for each method

Flag	Current method excess factor no weights	Perepelkin vs, c=4	New method vs, c=4	New method vs, c=4 s, d=2
Flag 1	medium risk	low risk	low risk	low risk
Flag 2	low risk	low risk	low risk	low risk
Flag 3	low risk	low risk	low risk	low risk
Flag 4	medium risk	high risk	low risk	low risk
Flag 5	medium risk	low risk	low risk	low risk
Flag 6	low risk	low risk	low risk	low risk
Flag 7	low risk	medium risk	low risk	low risk
Flag 8	low risk	low risk	low risk	low risk
Flag 9	low risk	low risk	low risk	low risk
Flag 10	medium risk	low risk	low risk	low risk
Flag 11	medium risk	low risk	low risk	low risk
Flag 12	low risk	low risk	low risk	low risk
Flag 13	medium risk	low risk	low risk	low risk
Flag 14	medium risk	low risk	low risk	low risk
Flag 15	low risk	low risk	low risk	low risk
Flag 16	medium risk	low risk	low risk	low risk
Flag 17	low risk	high risk	low risk	low risk
Flag 18	low risk	medium risk	low risk	low risk
Flag 19	medium risk	high risk	low risk	low risk
Flag 20	medium risk	low risk	low risk	low risk
Flag 21	low risk	low risk	low risk	low risk
Flag 22	medium risk	high risk	low risk	low risk
Flag 23	low risk	low risk	low risk	low risk
Flag 24	low risk	low risk	low risk	low risk
Flag 25	high risk	low risk	low risk	low risk
Flag 26	medium risk	medium risk	low risk	low risk
Flag 27	low risk	low risk	low risk	low risk
Flag 28	low risk	low risk	low risk	low risk
Flag 29	low risk	low risk	low risk	low risk
Flag 30	high risk	high risk	low risk	low risk
Flag 31	low risk	low risk	low risk	low risk
Flag 32	low risk	low risk	low risk	low risk
Flag 33	medium risk	low risk	low risk	low risk
Flag 34	low risk	low risk	low risk	low risk
Flag 35	high risk	medium risk	low risk	low risk
Flag 36	high risk	medium risk	low risk	low risk
Flag 37	low risk	medium risk	low risk	low risk
Flag 38	low risk	low risk	low risk	low risk
Flag 39	high risk	high risk	medium risk	low risk
Flag 40	low risk	medium risk	low risk	low risk
Flag 41	low risk	low risk	medium risk	low risk
Flag 42	high risk	low risk	low risk	low risk
Flag 43	high risk	low risk	low risk	low risk
Flag 44	low risk	low risk	low risk	low risk
Flag 45	low risk	low risk	low risk	low risk
Flag 46	low risk	medium risk	medium risk	low risk
Flag 47	high risk	medium risk	medium risk	low risk
Flag 48	low risk	low risk	low risk	low risk
Flag 49	low risk	medium risk	low risk	low risk
Flag 50	medium risk	high risk	medium risk	low risk

Flag	Current method excess factor no weights	Perepelkin vs, c=4	New method vs, c=4	New method vs, c=4 s, d=2
Flag 51	low risk	low risk	low risk	medium risk
Flag 52	low risk	high risk	medium risk	medium risk
Flag 53	low risk	medium risk	medium risk	medium risk
Flag 54	medium risk	low risk	medium risk	medium risk
Flag 55	low risk	low risk	low risk	medium risk
Flag 56	high risk	high risk	medium risk	medium risk
Flag 57	low risk	medium risk	medium risk	medium risk
Flag 58	low risk	low risk	medium risk	medium risk
Flag 59	low risk	medium risk	low risk	medium risk
Flag 60	low risk	low risk	medium risk	medium risk
Flag 61	low risk	low risk	low risk	medium risk
Flag 62	medium risk	low risk	low risk	medium risk
Flag 63	low risk	high risk	medium risk	medium risk
Flag 64	low risk	medium risk	medium risk	medium risk
Flag 65	low risk	low risk	medium risk	medium risk
Flag 66	low risk	low risk	medium risk	medium risk
Flag 67	medium risk	medium risk	medium risk	medium risk
Flag 68	high risk	high risk	medium risk	medium risk
Flag 69	low risk	low risk	medium risk	medium risk
Flag 70	high risk	high risk	medium risk	medium risk
Flag 71	low risk	low risk	medium risk	medium risk
Flag 72	high risk	medium risk	medium risk	medium risk
Flag 73	low risk	low risk	medium risk	medium risk
Flag 74	low risk	high risk	medium risk	medium risk
Flag 75	high risk	high risk	high risk	high risk
Flag 76	medium risk	medium risk	high risk	high risk
Flag 77	medium risk	high risk	high risk	high risk
Flag 78	high risk	high risk	high risk	high risk
Flag 79	medium risk	low risk	high risk	high risk
Flag 80	medium risk	medium risk	high risk	high risk
Flag 81	low risk	low risk	high risk	high risk
Flag 82	high risk	high risk	high risk	high risk
Flag 83	high risk	high risk	high risk	high risk
Flag 84	high risk	high risk	high risk	high risk
Flag 85	low risk	medium risk	high risk	high risk
Flag 86	low risk	low risk	high risk	high risk
Flag 87	medium risk	medium risk	high risk	high risk
Flag 88	high risk	high risk	high risk	high risk
Flag 89	low risk	low risk	high risk	high risk
Flag 90	high risk	high risk	high risk	high risk
Flag 91	low risk	low risk	high risk	high risk
Flag 92	high risk	high risk	high risk	high risk
Flag 93	high risk	medium risk	high risk	high risk
Flag 94	medium risk	medium risk	high risk	high risk
Flag 95	high risk	high risk	high risk	high risk
Flag 96	high risk	medium risk	high risk	high risk
Flag 97	medium risk	medium risk	high risk	high risk
Flag 98	high risk	high risk	high risk	high risk
Flag 99	high risk	high risk	high risk	high risk

Note: vs=very serious, s=serious