



Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

A framework estimating the minimum sample size and margin of error for maritime quantitative risk analysis[☆]

Romanas Puisa^{a,*}, Jakub Montewka^b, Przemyslaw Krata^b

^a RINA, Glasgow, United Kingdom

^b Faculty of Mechanical Engineering and Ship Technology, Gdańsk University of Technology, Poland

ARTICLE INFO

Keywords:

Minimum sample size
Maritime risk and safety
Accident frequency
Passenger ships

ABSTRACT

The average accident frequency is essential for quantitative risk analysis and is conventionally estimated from accident statistics. This paper has systematically synthesised the knowledge on statistical errors and offered the missing instructions, a framework, for determining the minimum sample size and the margin of error (MOE) when calculating the average accident frequency from an accident database at hand. We have applied this framework to representative accident datasets in the maritime domain and presented the revealing results that can already be used in QRAs based on these datasets. The findings are useful to both QRA analysts and policy makers. Interestingly, the framework application has revealed that the determined minimum sample sizes would exceed the datasets available in existing maritime casualty databases by decades, requiring at least 10% MOE to be factored into pertinent QRAs. By the same token, the earlier notable QRAs (developed as part of formal safety assessments in support of rule making) had to consider the MOE of over 30%, given the sample sizes used, likely shifting the conclusions they arrived at. Other findings of the application have shown that the average accident frequencies for large passenger ships have remained constant over the past 40 years.

1. Introduction

The average frequency of accidents (\bar{x}), such as ship collisions or fires, is essential for quantitative risk analysis (QRA), not least for formal safety assessment (FSA) in support of rule making, and elementary trend analysis, e.g. [1–3]. It is analogous to the failure rate in reliability engineering and hence is used in predictive models (e.g., in event trees for accident consequence analysis) as well in cost–benefit analysis when benefits of safety interventions are compared against associated costs. In the operational and regulatory settings, the \bar{x} also serves as one of many lagging safety indicators to be monitored and guide safety management.

However important the \bar{x} may be, the treatment of associated *parameter uncertainty*, particularly when caused by data variability, deserves a more rigorous, evidence-based approach, [4]. It is not enough to merely acknowledge the use of incomplete historical data and then perform sensitivity–uncertainty analysis on data-ignorant assumptions e.g., [5], or recognise the fact that accidents are inherently rare—hence limited data—but carry on with applying statistical instruments that fundamentally require adequate sample sizes. One should not then expect the

conclusions to be robust when ignoring the statistical error associated with the sample size at hand. Equally, attempting to obtain as much data as one possibly can is unwise, because one could unwittingly expend more effort than actually required to achieve one's objectives. All this is a poor practice of QRA, likely leading to systematic errors along its application.

A good practice is to use statistical errors as part of analysis and decision making, as opposed to ignoring them. Specifically, it is about being aware of the required *minimum sample size* (MSS) for the variability in the given accident dataset, and factoring in the corresponding *margin of error* (MOE) when the actual sample size is below the MSS.

However, we have found no study that specifically determines the MSS for QRA in the maritime transportation context nor accounts for the effect of limited sample size on the obtained accidents frequencies and associated MOE. The wider problem per se has attracted some attention, particularly in light of the debate on the interpretation of an accident probability in the context of FSA, [6], the use of accident statistics in general [7,8], and application of other methods to estimate the accident probability, [9–11]. Examples include a prediction

[☆] Disclaimer. This paper represents the opinions of the authors, and is the product of professional research. It is not meant to represent the position or opinions of the authors' organisations or official position of any staff members. Any errors are the fault of the authors.

* Corresponding author.

E-mail address: rpuisa@gmail.com (R. Puisa).

<https://doi.org/10.1016/j.ress.2023.109221>

Received 22 October 2022; Received in revised form 1 March 2023; Accepted 3 March 2023

Available online 6 March 2023

0951-8320/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

of the fatal accident probability using worldwide shipping accident data [12], estimation of oil spill frequencies pertaining to oil tankers accidents [13], the issue of accident under-reporting [14] and a solution for it [15]. The majority of solutions though focus on achieving the highest prediction strength of the models adopted rather than striving for the scientific rigour in deriving quantities based on solid scientific foundations. In contrast, recommendations for the MSS outside the maritime domain are abundant, e.g. reliability and safety engineering [16], or transportation safety, [17]. Most of the recommendations followed from various analyses of data variability, as done in this study.

Hence, there is a clear practice gap when it comes to using statistical errors for average accident frequencies in the maritime domain. This is because the MSS and the MOE have not been readily available for specific datasets nor there has been a framework—a clear set of instructions—for the MSS/MOE determination. With this in mind, the paper seeks to facilitate the use of a good practice in the maritime QRA by offering a framework for the systematic determination of MSS and MOE, along with its subsequent application to specific datasets. The framework is a synthesis of the existing, textbook knowledge on statistical errors, which makes it robust and widely applicable. This constitutes the scope of this paper and the novelty it offers, ultimately providing answers to the following practical questions:

1. How to determine the MSS and MOE for an accident dataset at hand?
2. What is the MSS for specific maritime accident categories and ship types, and how does the MOE vary with the sample size (time exposure) for these datasets?

The application included the world fleet of two main categories of safety-critical ships, namely cruise ships over 10,000 Gross Tonnage (GT) and passenger Roll-on/Roll-off ferries (RoPax) above 1,000 GT and 4,000 GT. We have considered typical accident categories, namely collision, grounding (stranded), contact, and fire over the period from 1 January 1980 to 31 December 2020 (40 years). The obtained accident frequencies for this period were categorised into corresponding years and then normalised by the fleet size in each year. The resultant annual accident frequencies per ship constituted 40 data points (per each accident category and ship type, 12 datasets in total) which were then used in the analysis.

Note that throughout the paper, the term ‘accident’ refers to all events recorded under a given accident category regardless the consequences. This means that the term encompasses both incidents (hazardous events with no or minor consequences) and accidents (hazardous events with serious consequences). Also, the paper, however, does not aim to improve trend analysis (statistical comparison), for it is a different problem, and an curious reader should refer to corresponding literature, e.g. [18,19].

The paper is organised as follows. Section 2 provides further details on the average accident frequency, required statistical properties for accident data, and introduces to the framework. Section 3 elaborates on the framework application details, specifically on the used datasets and outlines the application results in tabular and graphical forms. Section 4 offers a discussion on the study, whereas Section 5 highlights the caveats and limitations associated with it. Section 6 concludes the paper.

2. Definitions, prerequisites and framework

2.1. Definitions

The \bar{x} is often calculated as a statistical mean of event occurrences over a period of time (time exposure) normalised by the fleet size at risk. For example, if there have been ten collisions over the past five years in the fleet of 200 ships, the average collision frequency is $\bar{x} = 0.01$ per ship-year (i.e. a company with the fleet of 20 vessels should expect a collision every five years). The tacit assumption here is

that accidents occur at a constant average rate per year, i.e. the average frequency, or, at least, the rate that represents some central tendency (typical behaviour). In other words, the annual frequencies are assumed to be clustered around the mean value. However, this assumption is precarious and is completely at the mercy of the level of variability in the data. For instance, the following dataset (4, 1, 10)¹ represents real accident frequencies within three consecutive years. The average frequency would be 5 accidents per year (assuming the fleet size of one, for simplicity), which would be acceptably representative of the first year, but would grossly overestimate by factor of 5 and underestimate by factor of 2 the other years. This problem is conventionally mitigated by increasing the time exposure, i.e. by getting more sample years, as implied by the law of large numbers (LLN). Fig. 1 demonstrates the idea of how the sampling error (the difference between the sample mean and population mean or true mean) reduces with the sample size, depending of whether the underlying data distribution exhibits relatively low (Poisson) or high (Pareto) variability (measured by the variance). Hence, the higher variability is exhibited by a time series, the longer time exposure is required for the average frequency to be representative of the population mean (i.e., true frequency).

From the perspective of a data analyst, the annual variability in accident data is beyond control and it is driven by systematic and, to a lesser extent, random phenomena of global (at the industry level) and local (at the company level) nature. Examples of systematic factors are macroeconomic impacts such as changed fuel prices (could lead to slow or fast steaming and hence fewer or more collisions) and accident reporting, or rather under-reporting, culture [14,20]. Out-of-range environmental fluctuations that invalidate design assumptions could be an example of random factors; the blackout on cruise ship Viking Sky in 2019 is a case in point [21].

Hence, the increase in time exposure is the only means of improving the accuracy of the \bar{x} . The targeted accuracy level implies the determination of the minimum time exposure, referred to as the MSS, with a sample representing an annual accident frequency (count) within a given population — understood as a fleet of ships. The caveat, however, is that accident under-reporting, miss-reporting or other factors that affect the accuracy of annual accident frequencies introduce systematic errors into accident data. This means that even having data in excess of MSS does not guarantee that the true average frequency is captured.

The determination of MSS is straightforward as long as the data exhibit certain statistical properties such as: randomness, identical and thin-tailed distribution, independence, and stationarity. Section 2.2 elaborates on these statistical properties of accident data. Knowing MSS for various accident categories and ship types, one would also be able to determine the sampling error, expressed in terms of MOE, implicit in a dataset at hand or accident database that offers it, and adjust accordingly. If the database were found to contain a shorter time exposure than needed, the knowledge about the size of the gap would allow avoiding disproportional mitigation of the introduced parameter uncertainty. Thus for instance, small gaps that do not introduce high uncertainties may be ignored all together, whilst bigger ones would allow planning for a more rigorous analysis of the effect, e.g. by applying sensitivity analysis via Monte Carlo simulations [4].

2.2. Prerequisite statistical properties

Since the MSS calculation is based on the central limit theorem (CLT), there are inherent requirements that the data in question need to meet [22].

First, the data should be random, identically distributed and, ideally independent. That is, the time series should not contain non-random

¹ These correspond to collision frequencies within three consecutive years for RoPax ships above 4k GT; see Section 3.1.

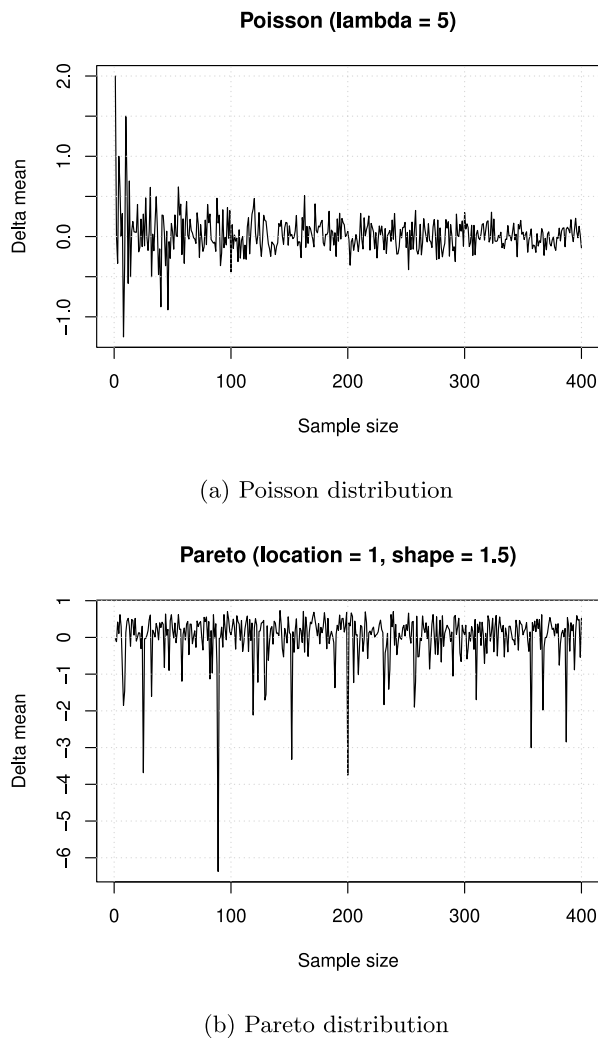


Fig. 1. Sample mean converges (a) or not converges (b) to the population mean with the sample size.

variability and be drawn from the same probability distribution. Second, the data should be stationary, i.e. drawn from a probability distribution whose variance, mean and auto-correlation structure do not change over time. In other words, the data are drawn from the same probability distribution over the entire time exposure. Third, and as indicated in Fig. 1, the underlying probability distribution should be not be fat-tailed or highly skewed, for the LLN is poorly applicable to such data [23]. The applicability of this and other requirements for the maritime accident data is discussed in the following section.

2.2.1. Randomness, identical and independent distribution

Research shows that the accidents in question are not mere random technical failures, but the results of flawed socio-technical interactions which are systematic and locally rational [24,25]. For instance, in case of onboard fires, pre-ignition events take time to develop. Conditions (e.g., wrong design assumptions, presence of design limitations) would lead to events or other conditions (e.g., ill-informed management, training, and operational procedures), which in turn lead to other events and conditions and so on [26,27]. The metaphors like “incubation period” and “drifting into failure” are used to explain the dormant, latent conditions in a system that, with time, insidiously degrade the system to the point when an accident becomes imminent [28]. This degradation or drift is systematic (not random) and fuelled by natural phenomena

of adaptation to new circumstances (endogenous and exogenous) and optimisation of resources [29]. It may also be seen as the inexorable manifestation of entropy.

The assumption of randomness is a simplification necessary for facile application of the mathematical instruments, particularly in QRA. This assumption can be justified on the conditions of full awareness of the common pitfalls and good practices are followed to avoid them. If this is done with adequate scientific rigour, it can indeed be useful. Alas, poor examples also exist [30]. Additionally, statistical tests should confirm the absence of non-random variation, as explained further.

If the process shows only random variation, the data points (time series) will be well distributed around the median [31]. This can be determined visually by looking for non-random variations. Specifically, if the process centre (mean) is shifting, one may observe unusually long runs of consecutive data points on the same side of the median or that the run chart crosses the median unusually few times. The length of the longest run and the number of crossings in a random process are predictable within limits and depend on the total number of data points in the run chart [32].

A *systematic shift* signal is present if any run of consecutive data points on the same side of the median is longer than the prediction limit (PL1) defined as:

$$PL1 = \lfloor \log_2(n) + 3 \rfloor \quad (1)$$

Data points that fall on the median do not count, they do neither break nor contribute to the run [33].

A *crossings signal* is present if the number of times the graph crosses the median is smaller than the prediction limit (PL2), [34].

$$PL2 = Q_{binom}(0.05, n - 1, 0.5) \quad (2)$$

Note that the shift and the crossings signals have the false positive signal rate of around 5% and hence have been proven useful in practice. The shift and crossings signals are two sides of the same coin and will often signal together, and hence any one of them is diagnostic of non-random variation [35]. However, we adopted a more conservative approach and required both signals to be present to confirm the non-random variation.

Another requirement is that the data are identically and independently distributed (i.i.d). That is, the data should be drawn from the same probability distribution with data points being mutually independent. This is another assumption that simplifies the underlying mathematics of statistical analysis, although it may, in some cases, not be realistic [36]. However, if we grouped accidents by ship type, the accidents within such groups become germane as if they were sampled from the same probability distribution. The assumption of independence can strongly hold for accidents caused by ships belonging to different operators—hence different safety management systems—who are also spread geographically and hence regulated by different authorities. As the data used in this paper belong to the world fleet (Section 3.1), this is exactly the case. Section 2.2.3 addresses these assumptions further.

2.2.2. Stationarity

A stationary process has the property that the mean, variance and auto-correlation structure do not change over time [37]. More generally, the location, scale and other parameters of an underlying probability distribution generating the events (e.g., number of collisions per year) are time-invariant.

Since a time series can be random but not stationary (e.g., a random walk), we test for stationarity using the Ljung–Box test for independence [38]. The test checks for significant evidence of nonzero correlations at a given time lag, which in this case is one, meaning that the correlation is calculated between accident frequencies that are one year apart. The null hypothesis of independence in a given time series is assumed and it is rejected if the p -value is less than 5% [39]. Thus, the

rejection of the null hypothesis means that the time series represents a non-stationary signal [40].

We note that assuming stationarity in the case of an inconclusive result would be more prudent than assuming otherwise. Serinaldi and Kilsby [41] argue that when the model structure and physical dynamics are uncertain, stationary models should be retained because they are simpler, more theoretically coherent, and more reliable for practical applications.

2.2.3. Type of underlying probability distribution

The average accident frequency per ship-year is often modelled with the use of a Poisson distribution in the context of risk analysis [42]. Conversely, as a stable average is rarely observed, as exemplified in Section 2.1, the annual accident frequency might, for instance, be biased towards lower values when nothing much happens most of the time and there are a few years with high accident frequencies. This is characteristic for some fat-tailed and other skewed distributions such as Pareto or Log-normal distribution, as was assumed in [2].

The challenge is, however, to fit any distribution to rather limited accident data (40 sample-years, see Section 3.1) or perform corresponding statistical tests to reveal the distribution type. Therefore, a more qualitative, first-principles approach is required instead. As argued below, a *negative binomial* can be assumed to model accident frequencies. This distribution can be skewed towards lower values, hence capturing the bias discussed above, and it has the stable mean and variance, i.e. the LLN applies to it.

The assumption of negative binomial distribution can be justified by considering the three classical models for accident frequencies or counts: *pure chance*, *prone* and *true contagion* [43,44]. All three models assume that ships subject to accidents operate under conditions of equal safety risk. If accidents are assumed to occur purely at random in a homogeneous fleet of ships, the frequency model leads to a Poisson distribution. The prone model assumes that not all ships are equally accident prone. Instead, ships are divided into sub-fleets with different susceptibilities to accidents. The overall distribution of accidents is then a composite of several Poisson sub-distributions. The resulting distribution is a negative binomial distribution.

Based on the true contagion model, it is assumed that each ship starts its operation with the same probability of having an accident. Then, if a ship happens to suffer an accident, the probability that it will suffer further accidents increases or decreases. This frequency model also results in a negative binomial distribution.

It should be noted that the parameters of a negative binomial distribution (i.e., the number of failures before a given number of successes and the probability of failure in each experiment) are determined directly from the mean and variance.

2.3. The framework

This section synthesises the knowledge on the statistical error into a framework for the MSS and MOE determination.

The MSS is conventionally calculated with the help of the CLT [45]. The basic idea behind the CLT is that the sample means (or any other linear combination of samples) are normally distributed regardless the underlying distribution from which the samples are drawn. Then the properties of a normal distribution allow estimating the confidence interval (CI) for the population mean:

$$CI = \left(\bar{x} - \frac{Z\sigma}{\sqrt{n}}, \bar{x} + \frac{Z\sigma}{\sqrt{n}} \right) \quad (3)$$

where \bar{x} is the sample mean (i.e. estimated population mean), n is the sample size (number of sample years) with a sample representing an annual accident frequency within a given fleet of ships, Z is a standard Z-score for the desired level of confidence (e.g., $Z = 1.96$ for 95% confidence interval), and σ would normally be the population standard

deviation if it is known. In practice, the unknown σ is often approximated by the sample standard deviation or by using *bootstrapping*, as explained further in Section 2.3.1.

The Eq. (3) shows how the sample mean becomes more accurate as the sample size, n , increases. Hence, if we wish to have a confidence interval that is W units in width ($W/2$ on each side of the sample mean), the above equation is solved for n to obtain the MSS:

$$MSS \geq \left(\frac{Z\sigma}{E} \right)^2 \quad (4)$$

where $E = W/2$, i.e. is the tolerable error (target precision), which is a half width of confidence interval shown in Eq. (3), i.e. it is the radius around the sample mean value at a certain confidence level. The E also corresponds to the MOE, thus, for a given n , the maximum MOE is:

$$MOE = \pm \frac{Z\sigma}{\sqrt{n}} \quad (5)$$

and in the percentage form from the sample mean value, as reported in the result Section 3.2:

$$MOE_{\%} = \pm \frac{Z\sigma}{\bar{x}\sqrt{n}} \quad (6)$$

In many fields of research, a typical MOE acceptable for a study is 5% [46]. It would correspond to the error of ± 5 accidents per year within the fleet of 100 ships. However, the accident data samples are inherently limited and hence a higher MOE should be allowed. We, therefore, assumed the MOE to be 10% (± 10 accident per year within the fleet of 100 ships), leading to $E = 0.1\bar{x}$ in Eq. (4).

The use of MOE in QRA is as follows. If, for instance, an event tree is used to model risk in terms of potential loss of life (PLL) or expected number of fatalities the risk model takes form of a linear model such as:

$$PLL = E(N) = F \cdot P \cdot N \quad (7)$$

where F is the accident frequency per ship-year, P is a product of the probabilities within the event tree, and N is the maximum number of fatalities from the initial event. Knowledge of MOE in conjunction with the F used allows the variation interval to be defined for F itself or directly for the resulting PLL such as:

$$PLL_{updated} = PLL \cdot (1 \pm MOE_{\%}) \quad (8)$$

The conservative value should then be assumed in the cost-benefit analysis. The flowchart demonstrating the main steps of the introduced framework is depicted in 2.

2.3.1. Bootstrapping

As indicated above, the calculation of the MSS and MOE requires the presence of the population standard deviation (σ). In practice, the unknown σ is either approximated by the sample standard deviation, σ_S , or determined by bootstrapping. The latter is a method which applies random resampling with replacement on the given data samples, thereby mimicking the original sampling process. Given that σ_S may significantly deviate from σ , the bootstrapping allows to statistically capture this uncertainty through estimates of an empirical distribution function for σ .

We generate an empirical distribution function of the population variance, σ^2 , and calculate distribution quantiles at 50% (median) and 95% probabilities. The two quantiles correspond to the most typical and high (conservative) population variances.

Fig. 3 shows an example distribution of the population variance generated by the bootstrapping for the RoPax (>4k GT) fires' dataset. The two quantiles are then used as input for the MSS calculations. The number of random runs was 1,000 and they were executed by using R package *boot* [47].

The application of the bootstrapping has the following caveats. The technique assumes that the input data samples are independent and identically distributed. As discussed in Section 2.2.1, these assumptions

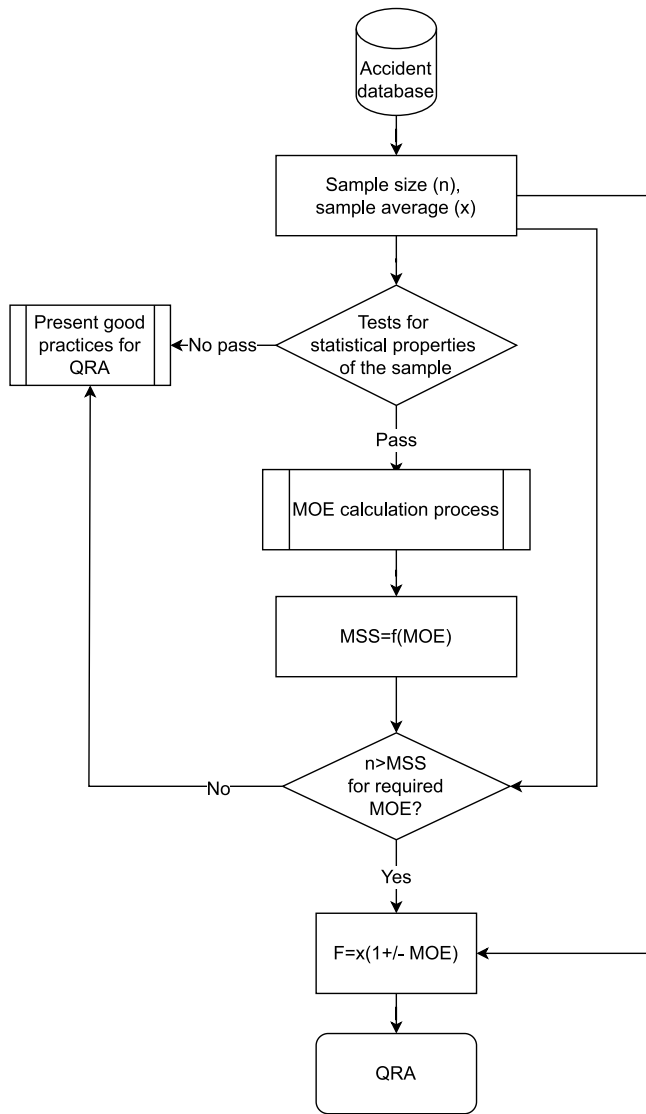


Fig. 2. A framework evaluating MOE and resulting MSS.

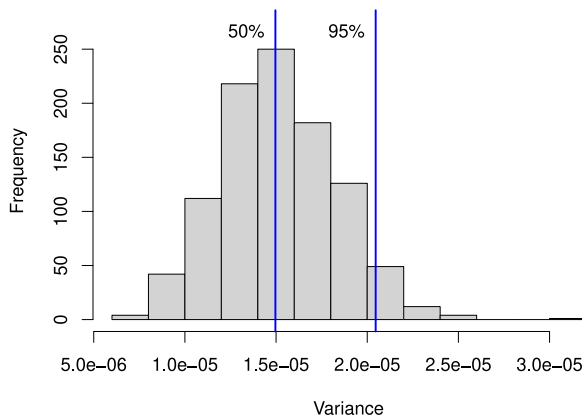


Fig. 3. Variance quantiles at 50% (median) and 95% probabilities for the RoPax (>4k GT) fires' dataset.

can be accepted for the data in question. If the underlying population lacks a finite variance, which can be the case with some heavy tailed distributions, the technique will not work reliably [48]. Section 2.2.3

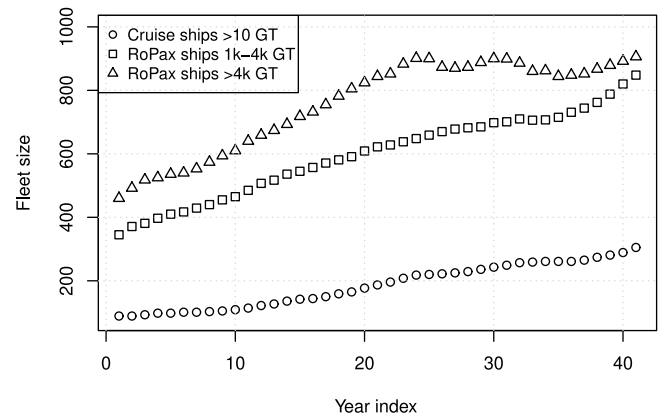


Fig. 4. Fleet size evolution.

Table 1

Number of accidents between 01/01/1980 to 31/12/2020, according to Sea-Web database.

	Collision	Contact	Grounding	Fire
Cruise ships (>10k GT)	78	70	61	109
RoPax ships (1k-4k GT)	157	170	135	106
RoPax ships (>4k GT)	318	420	154	280

argues that the underlying distribution is a negative binomial and hence it has a finite variance.

3. Application

This section describes the datasets used to apply the framework to, along with the application results.

3.1. Data

To obtain maritime time series of accident frequencies we used the Sea-web² causality database. The database allows performing multi-criteria searches such as causality type, geographic location, ship type, ship size and other ship particulars, etc. For the sake of this paper, we searched for causalities with cruise and RoPax ships worldwide within the time period 01/01/1980 to 31/12/2020 (40 years).³ We stopped at the 40 year time exposure because longer time exposures hardly added data to the datasets; possibly due to general unavailability of older records and higher under-reporting rates in the past. Hence, the 40 year period may be considered somewhat arbitrary.

We limited ship sizes in accordance with earlier FSA studies for these two ship types [49,50]. Specifically, cruise ship sizes were limited to be above 10,000 GT, whereas for RoPax ships were split into two size categories: 1,000–4,000 GT and above 4,000 GT. Table 1 shows the number of events obtained from the database. Fig. 4 shows the annual data on the fleet sizes obtained. The fleet sizes have been growing for all three ship categories; Table 2 indicates the approximate fleet sizes at the end of the period. The annual accident frequencies normalised by fleet size are reported in Section 3.2.

Note, the accident data obtained from the Sea-Web database contain multiple records per year, with detailed time (hours, day, month, year) and other attributes such as accident locations, fatalities involved, etc. For instance, there was one registered fire accident in 1980 on cruise

² <https://maritime.ihs.com>

³ Selected references to cruise ships: “Passenger (Cruise) Ship, Passenger /Cruise, Passenger Ship”, and RoPax: “Passenger/Ro-Ro Cargo, Passenger/Ro-Ro Cargo Ship, Passenger/Ro-Ro Ship (Vehicles), Passenger/Ro-Ro Ship (Vehicles/Rail)”

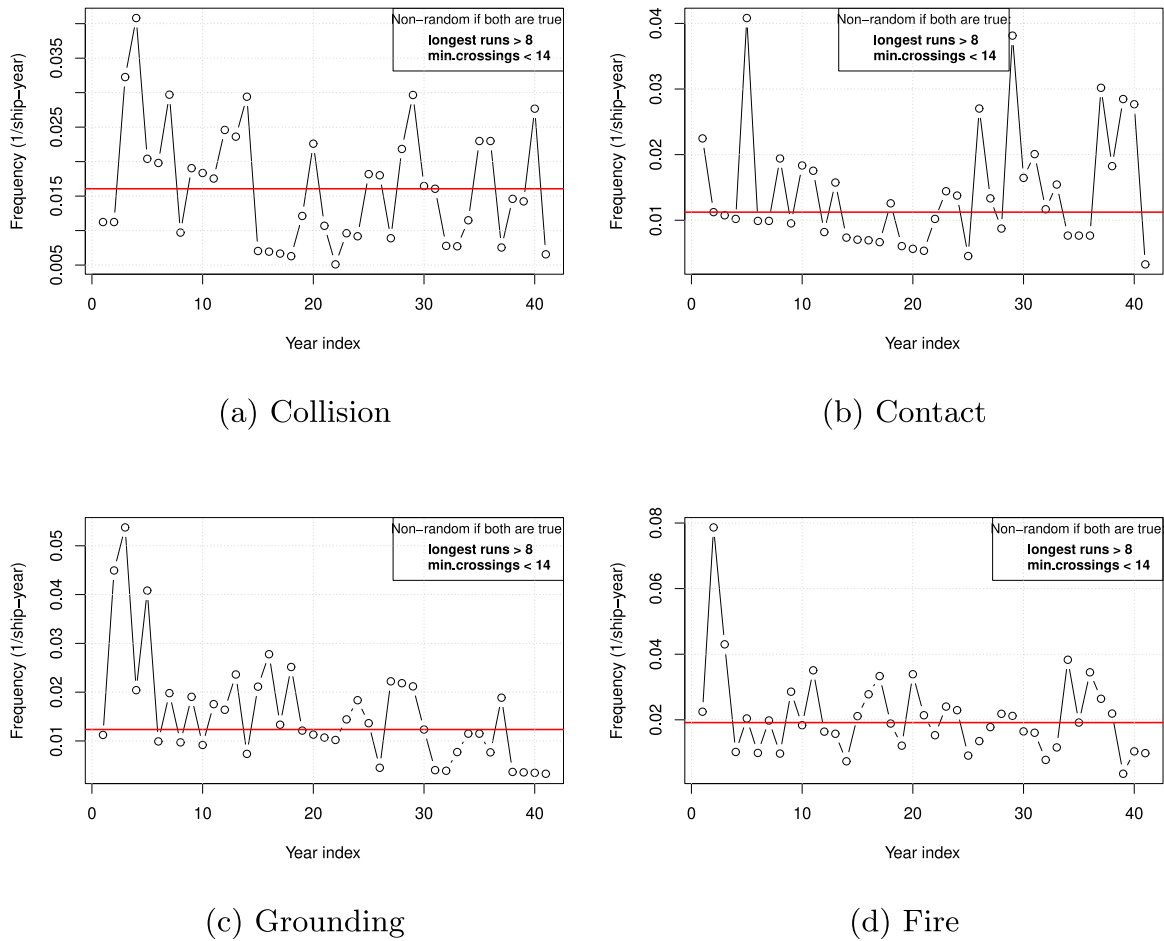


Fig. 5. Run charts for cruise (>10k GT) accidents. The horizontal red line corresponds to the median. Refer to Section 2.2 for explanation of the terms.

Table 2
Approximate fleet sizes on 31/12/2020, according to Sea-Web database.

	Fleet size
Cruise ships (>10k GT)	305
RoPax ships (1k–4k GT)	848
RoPax ships (>4k GT)	906

ships >10k GT, but already 6 fires in 1981, 3 fires in 1982, again 1 fire in 1984, and even 9 accidents in 2013. Hence, in principle one could work with shorter time periods (e.g., month) and calculate accident frequencies per month, per day, per hour etc. However, it is customary to work with the annual frequencies in the maritime domain [42]. Hence, we limited our study to this time scale.

3.2. Results

This section outlines the application results of the MSS/MOE framework presented in Section 2.3. The focus here is on the numerical results rather than the framework application aspects. The latter are discussed in Section 4.

3.2.1. Non-random variation

Figs. 5 to 7 display run charts of annual accident frequencies per ship-year used to visually test for non-random variation. We used the effective fleet size in each year. We specifically looked at the plot legends that contain the data specific criteria and then compared them against the plotted data points. Thus for instance, Fig. 5(a) shows that the longest runs (the number of consecutive data points on the same

side of the median, excluding the points of the median itself) have to be above 8 points for a non-random (systematic) variation to be present, as per Eq. (1). Additionally, the minimum number of crossings (the number of times the graph crosses the median) has to be below 14 if a non-random variation is present, see Eq. (2).

Looking at the plot, there are maximum 6 longest runs (less than required) and 14 crossings (more than required). Hence, the time series for cruise collisions in Fig. 5(a) does not exhibit non-random variation. Table 3 summarises the test results for all 12 datasets, indicating if the non-random variation is present or absent. The results show that 2 out of 12 datasets exhibit non-random variation.

3.2.2. Stationarity

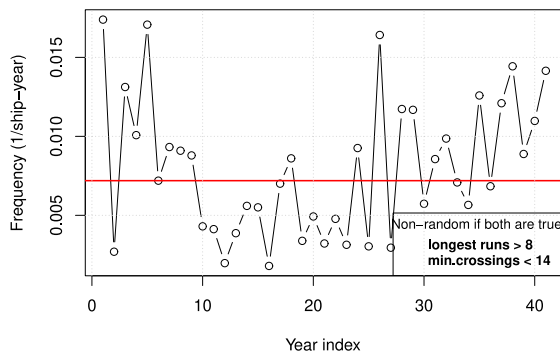
The results of the Ljung–Box tests for stationarity of the data are also given in Table 3. A stationary signal has a p -value greater than 5%. The results show that only the cruise ship grounding data set is non-stationary. Based on the tests for nonrandom variation and stationarity, Table 3 then indicates whether the further statistical analysis, i.e., the calculations of MSS and MOE, was applicable (A) or not (NA).

We note that only 3 of 12 records failed the tests: grounding accidents for cruise ships and contact accidents for both types of RoPax ships. This statistical property supports the visual observation that the trend for grounding accidents for cruise ships is decreasing (Fig. 5(c)), while the rate of contact accidents for RoPax ships appears to be increasing (see Figs. 6(b) and 7(b)). Note that in the absence of this statistical property, any visual trend analysis can be misleading, as in the case of fire accidents for RoPax (1k–4k GT) in Fig. 6(d). It looks as if the frequency of fires has decreased overall, but this contradicts what the statistical analysis says: the time series is random and stationary.

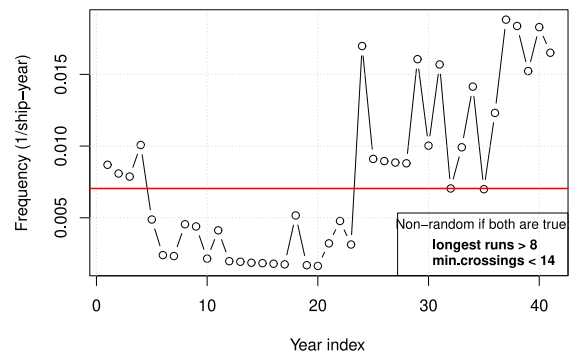
Table 3

Summary of statistical test results for non-random variation (present or absent), stationarity (p -value > 5%) and combined (further statistical analysis is applicable -A- or not applicable -NA).

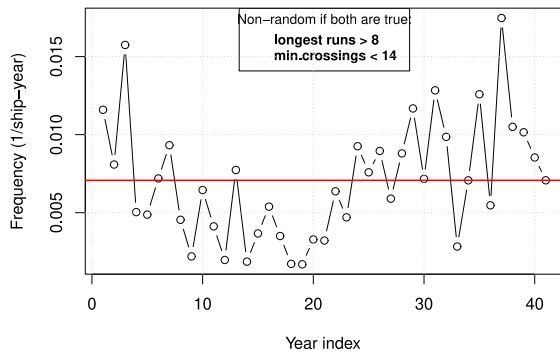
	Collision	Contact	Grounding	Fire
Cruise ships (>10k GT)	Absent	Absent	Absent	Absent
Non-random variation	5.9	99.6	1.1	15.6
Stationarity (% p -value)	A	A	NA	A
Verdict				
RoPax ships (1k–4k GT)	Absent	Present	Absent	Absent
Non-random variation	84.0	20.3	5.6	10.1
Stationarity (% p -value)	A	NA	A	A
Verdict				
RoPax ships (>4k GT)	Absent	Present	Absent	Absent
Non-random variation	62.8	46.3	21.8	15.5
Stationarity (% p -value)	A	NA	A	A
Verdict				



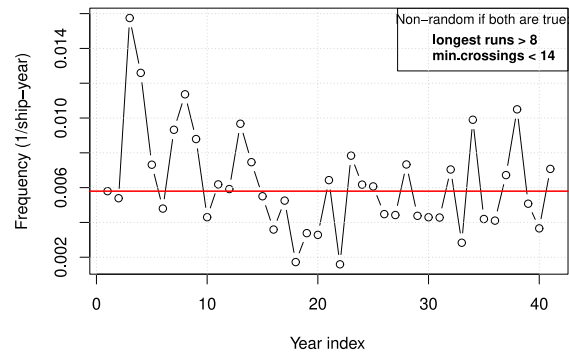
(a) Collision



(b) Contact



(c) Grounding



(d) Fire

Fig. 6. Run charts for RoPax (1k–4k GT) accidents. The horizontal red line corresponds to the median. Refer to Section 2.2 for explanation of the terms.

The remaining 9 data sets—the majority—are stationary and randomly distributed around the median values of the last 40 years. Thus, there are arguably no statistically significant global or industry-wide effects on the corresponding accident rates that would involve nonrandom variation or nonstationary behaviour.

This finding is bad news for maritime safety efforts. However, it is good news for analysts because a larger pool of data is potentially available for statistical analysis beyond what has been presented in this paper. For example, one might attempt to fit negative binomial distributions to the data and then compare different accident or vessel categories in terms of the statistical characteristics of the fitted distributions. This could be more informative than simply comparing average accident rates, such as was done in [3].

Then, the test results were used to determine which datasets could be used to calculate MSS and MOE, as shown in the following section.

3.2.3. MOE and MSS

Fig. 8 shows the log–log plots of margin of error (MOE) at 95% confidence level for the valid datasets. These plots can be used to determine the MOE for a given sample size, and hence the maximum distance between the sample and population mean—the sampling error. The median (typical) and conservative (high variability) MSS for the datasets are presented in Table 4.

The recommended MSS in Table 4 indicate that if we had only 40 years-long time exposure—as obtained for the current study—that would not be enough to calculate the average accident frequencies

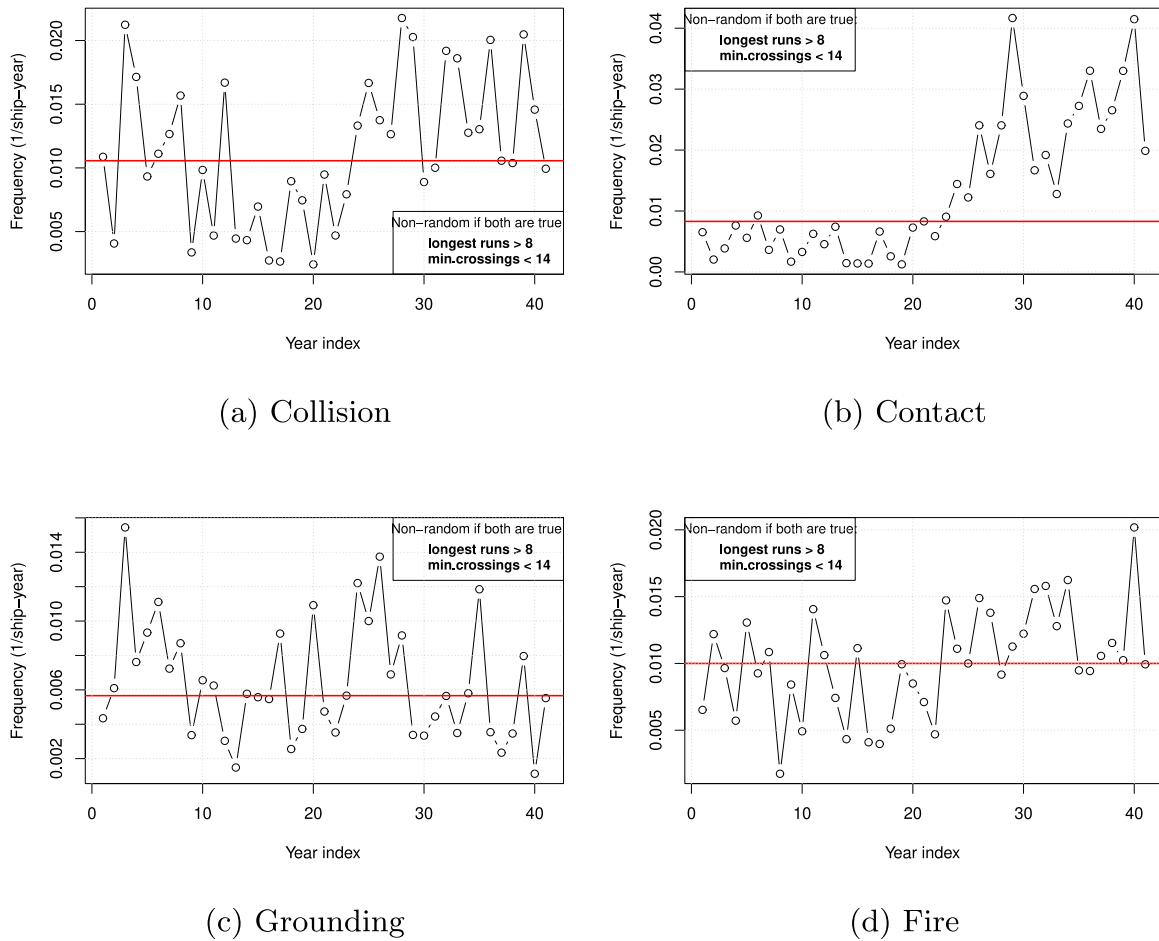


Fig. 7. Run charts for RoPax (>4k GT) accidents. The horizontal red line corresponds to the median. Refer to Section 2.2 for explanation of the terms.

Table 4
Minimum sample sizes (number of sample years) for median (typical) and high variability scenarios, at 95% confidence level and the tolerable error (or MOE) of $\pm 10\%$.

	Collision	Contact	Grounding	Fire
Cruise ships (>10k GT)	(102, 144)	(151, 222)	NA	(139, 277)
RoPax ships (1k–4k GT)	(109, 144)	NA	(110, 155)	(80, 127)
RoPax ships (>4k GT)	(93, 122)	NA	(108, 146)	(57, 79)

(rates) with the target precision of $\pm 10\%$. Since longer time exposures are not readily available, Fig. 8 shows the MOE that one could expect for a dataset at hand. With the current time exposure, one should factor in some $\pm(20\%–30\%)$ MOE when analysing the cruise ship accidents in QRA studies (see Section 2.3).

4. Discussion

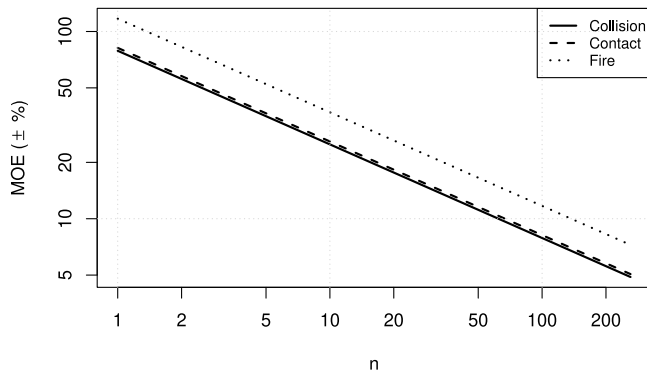
This study began with the premise that one must follow a good QRA practice that involves the systematic use of statistical errors in analysis and subsequently decision making. In the context of maritime risk analysis this practice has not been universally followed when it comes to the average frequency of accidents per ship-year. This is because the MSS and the MOE have not been readily available for specific datasets (characterised in terms of accident category, ship type and ship size) nor there has been a framework—a clear set of instructions—for the MSS/MOE determination.

We argue that the unavailability of MSS/MOE could have shifted conclusions in the following earlier QRA/FSA studies where the statistical errors were ignored. Thus, an FSA for cruise ships above 10k

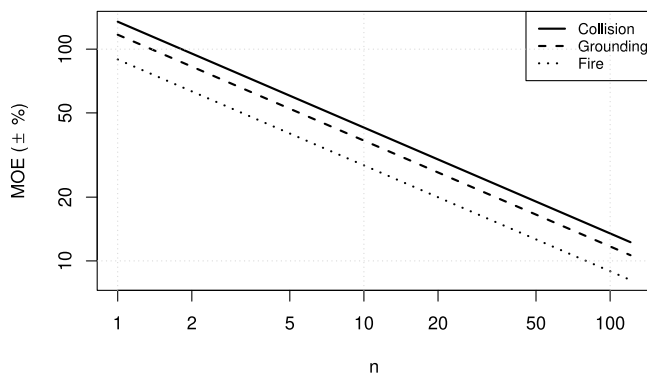
GT used data over 15 years (1990–2004) for collision, contact, grounding and fire accidents [51]. This would require the MOE of some $\pm(30\%–45\%)$, depending on the accident category. Had this MOE been taken into account, the risk value could have been reduced by 55%, potentially invalidating the cost-effectiveness of the risk control options (e.g., RCO 1+3 and RCO 1+2+3) the assessment recommended. In another FSA study on RoPax ships well above 1k GT, the used time exposure was just over 10 years (1990 – 2004) [52]. This would lead to the MOE of $\pm(25\%–50\%)$, with analogous consequences to the conclusions of cost-benefit analysis. Hence, both FSA studies would have benefited from the results of this paper. Another statistical analysis of so-called safety level by Eliopoulou et al. [3] looked at accident records with RoPax, cruise and other ships over the period of 12 years (2000–2012). The study derived trends for collision, contact, grounding, fire and other events, and calculated average accident frequencies. The latter were used to compare the accident categories or ship types without considering the error in the values. We argue that the comparison conclusions would have been shifted, should the expected MOE of at least 30% have been considered.

It should be noted that the fact that the three datasets which did not pass the statistical tests (see Table 3) does corroborate with the visual analysis. The datasets exhibit clear downward trends (cruise ships) and sharp jumps upwards leading to increased value by factor of 5 (RoPax ships), indicating they could not be reasonably assumed to be sampled from stationary probability distributions. Hence ultimately, both numerical and visual analyses are necessary to conclude which data is suitable for further statistical analysis towards MSS and MOE.

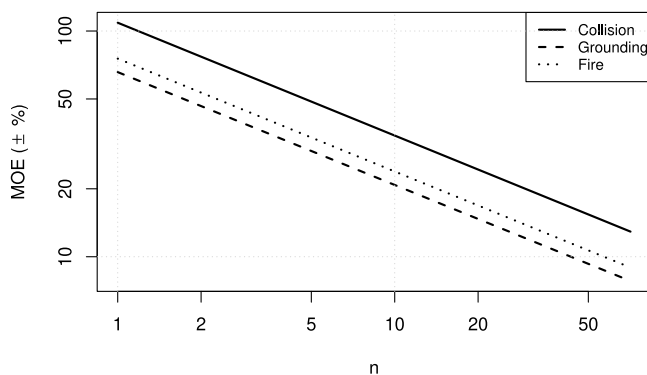
The presented run charts for accident time series provide useful insight into the presence, or absence, of systematic trends in annual



(a) Cruise ships (> 10 GT)



(b) RoPax ships (1k-4k GT)



(c) RoPax ships (>4k GT)

Fig. 8. Log-log plots of margin of error (MOE) at 95% confidence level as a function of sample size.

accident rates (Section 3.2). Thus, it shows that 9 out of 12 datasets have no non-random variations and have been stationary over the course of 40 years (1980–2020), meaning that neither statistically significant surge nor decline of safety records can be observed. Thus arguably, most of the large passenger ships have been running into accidents at the same annual rate over the last 40 years. This is contrary to the conclusions often found in the literature that the corresponding

frequencies have been either increasing or decreasing, e.g. [3]. In other words, the run charts along with statistical tests introduced in Sections 2.2.1 and 2.2.2 could potentially serve as a good alternative for trend analysis—something to be confirmed in future research.

The observation of the stable average accident frequencies is useful, although in different ways, to both analysts and policy makers. The former are interested in stationary statistical properties of data, for it allows for statistical inference with, as suggested Section 2.2.3, a negative binomial distribution.⁴ In turn, the policy makers are interested in the impact of implemented safety policies. Arguably, the impact on accident frequencies has been negligible most of the time, although future research needs to confirm this.

5. Caveats and limitations

As already indicated in Section 3.1, the considered time exposure of 40 years was somewhat arbitrary, guided by the perceived availability of accident records within the casualty database. The general rule we followed was to have the longest time exposure possible, being cautious that the further back in time we go (say beyond 30 years from now), the lower accuracy of accident data could be expected. We did not analyse or factor in the accuracy of annual accident rates (e.g., due to under-reporting [14,20]), because pertinent information was simply unavailable. We also had no access to information on how the redefinition of a collision and contact accidents in 2012 by the UK [53], influenced the numbers of cases in each category.

We are aware that accident underlying factors, such as operational conditions and used technology, are subject to significant change. This poses a challenge to the predictive power of the presented statistical analysis and hence the results should be used with care.

The run charts in Figs. 5 to 7 show that shorter time exposures could be considered for the statistical analysis conducted. Thus for instance, one can notice that the upward trend for RoPax contact accidents may have become stabilised over the last 15 years and the statistical analysis could be valid for this shorter period. However, we argue that the shortened time exposure (i.e., the sample size of some 15 points) would fly in the face of the MSS calculation results (recall the lowest MSS is 57, Table 4), and it hence was dismissed.

6. Conclusions

The paper provides a framework for systematically identifying MSS and MOE, suitable for QRA analysts and policy makers. The framework answers the practical questions associated with MSS for a given dataset (maritime accident and ship types) and the associated MOE.

The paper has specifically synthesised the knowledge on statistical errors into a generic framework for the MSS and MOE determination and applied the framework to specific accident datasets.

The resulting MSS values have been summarised in a tabular form, whereas the MOE is determined from the MOE plots as functions of the available sample size (time exposure in years). The lowest value for the MSS corresponds to fire accidents with RoPax ships above 4k GT, the value is between 57 and 79 sample years. The highest MSS, between 193 and 277 sample years, is associated with fire accidents on cruise ships. Thus, given that no accident database exists to contain at least 57 years of records, the \bar{x} will always contains the MOE above 10%. To our best knowledge such study for the maritime domain has not been performed earlier, making the presented work novel and relevant.

Future studies should look at other critical ship types, such as oil tankers, to raise awareness among QRA analysts and their end users.

⁴ Note, the distribution parameters, the success probability and the number of failures before a prespecified number of successes is reached, are directly calculable from the sample mean and variance.

CRedit authorship contribution statement

Romanas Puisa: Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jakub Montewka:** Writing – original draft, Methodology, Conceptualization. **Przemyslaw Krata:** Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgment

The access to the Sea-web causality database was provided by the Maritime Safety Research Centre of University of Strathclyde during the first author's employment there. This is kindly acknowledged.

References

- [1] Vanem Erik, Antao Pedro, Østvik Ivan, de Comas Francisco Del Castillo. Analysing the risk of LNG carrier operations. *Reliab Eng Syst Saf* 2008;93(9):1328–44.
- [2] EMSA. Study investigating cost effective measures for reducing the risk from fires on ro-ro passenger ships (FIRESAFE). Appendix: Sensitivity and Uncertainty Analyses. 2016.
- [3] Eliopoulou Eleftheria, Papanikolaou Apostolos, Voulgarellis Markos. Statistical analysis of ship accidents and review of safety level. *Saf Sci* 2016;85:282–92.
- [4] Merrick Jason RW, Van Dorp Rene. Speaking the truth in maritime risk assessment. *Risk Anal Int J* 2006;26(1):223–37.
- [5] EMSA. FIRESAFE II. Combined assessment. Final report, Version 1.1. 2018.
- [6] Montewka Jakub, Goerlandt Floris, Kujala Pentti. On a systematic perspective on risk for formal safety assessment (FSA). *Reliab Eng Syst Saf* 2014;127:77–85.
- [7] Du Lei, Goerlandt Floris, Kujala Pentti. Review and analysis of methods for assessing maritime waterway risk based on non-accident critical events detected from AIS data. *Reliab Eng Syst Saf* 2020;200:106933.
- [8] Mauro Francesco, Vassalos Dracos, Paterson Donald. Critical damages identification in a multi-level damage stability assessment framework for passenger ships. *Reliab Eng Syst Saf* 2022;228:108802.
- [9] Ung ST. Navigation Risk estimation using a modified Bayesian Network modeling-a case study in Taiwan. *Reliab Eng Syst Saf* 2021;213:107777.
- [10] Mazurek Jolanta, Lu Liangliang, Krata Przemyslaw, Montewka Jakub, Krata Hubert, Kujala Pentti. An updated method identifying collision-prone locations for ships. A case study for oil tankers navigating in the Gulf of Finland. *Reliab Eng Syst Saf* 2022;217:108024.
- [11] Zhang Mingyang, Kujala Pentti, Hirdaris Spyros. A machine learning method for the evaluation of ship grounding risk in real operational conditions. *Reliab Eng Syst Saf* 2022;226:108697.
- [12] Weng Jinxian, Yang Dong. Investigation of shipping accident injury severity and mortality. *Accid Anal Prev* 2015;76:92–101.
- [13] Siqueira Paulo Gabriel, das Chagas Moura Márcio, Duarte Heitor Oliveira. A Bayesian population variability based method for estimating frequency of maritime accidents. *Process Saf Environ Prot* 2022;163:308–20.
- [14] Hassel Martin, Asbjørnslett Bjørn Egil, Hole Lars Petter. Underreporting of maritime accidents to vessel accident databases. *Accid Anal Prev* 2011;43(6):2053–63.
- [15] Li Guorong, Weng Jinxian, Wu Bing, Hou Zhiqiang. Incorporating multi-scenario underreporting rates into MICE for underreported maritime accident record analysis. *Ocean Eng* 2022;246:110620.
- [16] Indira V, Vasanthakumari R, Jegadeeshwaran R, Sugumaran V. Determination of minimum sample size for fault diagnosis of automobile hydraulic brake system using power analysis. *Eng Sci Technol Int J* 2015;18(1):59–69.
- [17] Izdebski Mariusz, Jacyna-Golda Ilona, Golda Paweł. Minimisation of the probability of serious road accidents in the transport of dangerous goods. *Reliab Eng Syst Saf* 2022;217:108093.
- [18] Kvaløy Jan Terje, Aven Terje. An alternative approach to trend analysis in accident data. *Reliab Eng Syst Saf* 2005;90(1):75–82.
- [19] Zhang Jinfen, Wan Chengpeng, He Anxin, Zhang Di, Soares C Guedes. A two-stage black-spot identification model for inland waterway transportation. *Reliab Eng Syst Saf* 2021;213:107677.
- [20] Psarros George, Skjong Rolf, Eide Magnus Strandmyr. Under-reporting of maritime accidents. *Accid Anal Prev* 2010;42(2):619–25.
- [21] Ibrion Michaela, Paltrinieri Nicola, Nejad Amir R. Learning from failures in cruise ship industry: The blackout of Viking Sky in Hustadvika, Norway. *Eng Fail Anal* 2021;125:105355.
- [22] Montgomery Douglas C, Runger George C. Applied statistics and probability for engineers. John Wiley & Sons; 2010.
- [23] Taleb Nassim Nicholas. How much data do you need? An operational, pre-asymptotic metric for fat-tailedness. *Int J Forecast* 2019;35(2):677–86.
- [24] Perrow Charles. Normal accidents. Princeton University Press; 2011.
- [25] Kaptan Mehmet, Uğurlu Özkan, Wang Jin. The effect of nonconformities encountered in the use of technology on the occurrence of collision, contact and grounding accidents. *Reliab Eng Syst Saf* 2021;215:107886.
- [26] Puisa Romanas, Williams Stuart, Vassalos Dracos. Towards an explanation of why onboard fires happen: The case of an engine room fire on the cruise ship “Le Boreal”. *Appl Ocean Res* 2019;88:223–32.
- [27] Puisa Romanas, Lin Lin, Bolbot Victor, Vassalos Dracos. Unravelling causal factors of maritime incidents and accidents. *Saf Sci* 2018;110:124–41.
- [28] Dekker Sidney, Pruchnicki Shawn. Drifting into failure: theorising the dynamics of disaster incubation. *Theor Issues Ergon Sci* 2014;15(6):534–44.
- [29] Rasmussen Jens. Risk management in a dynamic society: a modelling problem. *Saf Sci* 1997;27(2–3):183–213.
- [30] Rae Andrew, McDermaid John, Alexander Rob. The science and superstition of quantitative risk assessment. *J Syst Saf* 2012;48(4):28.
- [31] Chambers John M. Graphical methods for data analysis. CRC Press; 2018.
- [32] Anhoj Jacob, Olesen Anne Vingaard. Run charts revisited: a simulation study of run chart rules for detection of non-random variation in health care processes. *PLoS One* 2014;9(11):e113825.
- [33] Schilling Mark F. The surprising predictability of long runs. *Math Mag* 2012;85(2):141–9.
- [34] Chen Zhenmin. A note on the runs test. *Model Assist Stat Appl* 2010;5(2):73–7.
- [35] Anhoj Jacob. Run charts with R. 2021, [Online] <https://cran.r-project.org/web/packages/qicharts/vignettes/runcharts.html>. (Accessed 1 July 2021).
- [36] Hampel Frank, Zurich Eth. Is statistics too difficult? *Canad J Statist* 1998;26(3):497–513.
- [37] Gagniuc Paul A. Markov chains: from theory to implementation and experimentation. John Wiley & Sons; 2017.
- [38] Harvey Andrew C. Time series models. MIT Press; 1993.
- [39] Nuzzo Regina. Statistical errors. *Nature* 2014;506(7487):150.
- [40] Dare Jayeola, Patrick Aye O, Oyewola David O. Comparison of stationarity on Ljung box test statistics for forecasting. *Earthline J Math Sci* 2022;8(2):325–36.
- [41] Serinaldi Francesco, Kilsby Chris G. Stationarity is undead: Uncertainty dominates the distribution of extremes. *Adv Water Resour* 2015;77:17–36.
- [42] Kristiansen Svein. Maritime transportation: safety management and risk analysis. Routledge; 2013.
- [43] Arbous Adrian Garth, Kerrich JE. Accident statistics and the concept of accident-proneness. *Biometrics* 1951;7(4):340–432.
- [44] Edwards Carol B, Gurland John. A class of distributions applicable to accidents. *J Amer Statist Assoc* 1961;56(295):503–17.
- [45] Dekking Frederik Michel, Kraaikamp Cornelis, Lopuhaä Hendrik Paul, Meester Ludolf Erwin. A modern introduction to probability and statistics: understanding why and how. Springer Science & Business Media; 2005.
- [46] Kosar Tomaž, Bohra Sudev, Mernik Marjan. A systematic mapping study driven by the margin of error. *J Syst Softw* 2018;144:439–49.
- [47] Cauty Angelo, Ripley BD. boot: Bootstrap R (S-Plus) functions. 2021, R package version 1.3-28.
- [48] Athreya KB. Bootstrap of the mean in the infinite variance case. *Ann Statist* 1987;724–31.
- [49] MSC. Formal Safety Assessment - Cruise ships. Submitted by Denmark. MSC 85/INF.2. 2008.
- [50] MSC. Formal Safety Assessment - RoPax ships. Submitted by Denmark. MSC 85/INF.3. 2008.
- [51] Formal safety assessment - cruise ships. MSC 85/INF.2. Technical report.
- [52] Formal safety assessment - RoPax ships. MSC 85/INF.3. Technical report.
- [53] The European Parliament and the Council. Directive 2009/18/EC of the European Parliament and of the Council of 23 April 2009 establishing the fundamental principles governing the investigation of accidents in the maritime transport sector and amending council directive 1999/35/EC and Directive 2002/59/EC of the European Parliament and of the Council (Text with EEA Relevance). Brussels: The European Parliament and the Council; 2009.