Safety Science 106 (2018) 154–161

Contents lists available at ScienceDirect

Safety Science

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

Generating synthetic aviation safety data to resample or establish new datasets

Andrej Lališ[a,*], Vladimír Socha[a], Petr Křemen[b], Peter Vittek[a], Luboš Socha[c], Jakub Kraus[a]

[a] Faculty of Transportation Sciences, Czech Technical University, Prague, Czech Republic
[b] Faculty of Electrical Engineering, Czech Technical University, Prague, Czech Republic
[c] Faculty of Aeronautics, Technical University of Košice, Košice, Slovakia

ARTICLE INFO

Keywords:
Aviation safety
Data resampling
Data simulation
Safety management system
Safety performance
Aerospace Performance Factor

ABSTRACT

Aviation safety data are limited in availability due to their confidential nature. Some aggregated overviews already exist but in order to effectively use the data, it is important to fill the gaps of their existing limitations. For some data, there are not enough data points in order to process them through advanced analysis. For other, only expert assumptions can be obtained. In both cases, these shortcomings can be addressed via proper data resampling or simulation where little effort can make the data suitable for various research and development initiatives. Examples of real aviation safety data made public are demonstrated together with key principles of how to perform their resampling. Then, for cases where only expert assumptions are available, general solution to the transformation of the assumptions into simulated data is introduced. The goal is to demonstrate how to transform accessible data or knowledge about aviation safety into data samples with sufficient granularity. The results provide general solution suitable not only for aviation safety data and knowledge, but also for similar transportation or high-risk industries related data issues, indicating that both the data resampling and simulation provide an option for generating datasets, which can be used for statistical inferential methods, linear regression modelling, recurrent analysis etc. Example of data resampling application is included in Aerospace Performance Factor calculation for years 2008 up to 2015.

1. Introduction

To date, aviation safety is subject of intensive research in terms of new information technology deployment. It is recognised, that further progress in this domain can be achieved by implementing technology, which collects, processes and analyses safety data in order to produce system-wide information of how the system performs on safety (ICAO, 2013). This information is to be used for safety-critical decision-making within safety management system as far as the aviation is concerned, but this principle is generally true for other high-risk industries as well (Niu and Song, 2013; Klein and Viard, 2013). One of the features of the system-wide information is that it cannot be reliably derived by individuals from the data available because aviation became very complex, i.e. hardly manageable for humans. The industry is distributed system of many types of stakeholders (airspace users, organisations, regulators, manufacturers, policy makers etc.) which use different technologies, different procedures and which overlap with each other to various extent. As a result, safety performance of one stakeholder may be severely affected by how safety is managed by other stakeholder and it can be difficult to identify this from either side.

Today's accidents only support this claim. They consist of long chain of events and contributing factors, which typically exceed responsibilities of one stakeholder and its safety management (Socha et al., 2014). From the perspective of managing safety, it is important to have some sort of full picture to be able to apply effective measures to prevent modern accidents. The distributed character of aviation, however, sets constraints for achieving such a full picture. Not only are the data and the full picture to be established distributed in parts among the stakeholders, but also the nature of both is often confidential and may have potential to damage someone's position on the market, if misused. Aviation authorities encourage organisations and other stakeholders to share safety data, experience and safety knowledge (ICAO, 2013; European Commission, 2010) but the degree of these activities is still not perceived to be satisfactory. This paper does not aim to resolve this issue but rather address its consequences, namely limited safety data availability.

Research and development initiatives in the domain of aviation safety are restricted by the safety data not being available. Whilst it may be possible to sign some bilateral confidential agreement between two parties, this is still rather difficult to achieve for multiple

* Corresponding author at: Czech Technical University in Prague, Faculty of Transportation Sciences, Department of Air Transport, Horská 3, 128 03 Prague, Czech Republic.
E-mail address: lalisand@fd.cvut.cz (A. Lališ).

https://doi.org/10.1016/j.ssci.2018.03.013
Received 22 September 2016; Received in revised form 20 February 2018; Accepted 8 March 2018
Available online 21 March 2018
0925-7535/

stakeholders at the time. Fortunately, some of the data are regularly (annually) published by authorities in form of aggregated overview of key safety issues (such as Safety Regulation Commission, 2016; EASA, 2016 or Federal Aviation Administration, 2015), but this is true only for some segments of the industry, e.g. for air navigation services providers (ANSPs). These providers have a lot of advanced technology and data at their disposal and they are typically state-owned monopolies, which are not subject of market competition. The latter was likely the key factor for making some of their data publicly available.

To better understand the issue, it is important to note basic facts of data evolution in this domain. Safety was always measured indirectly, i.e. through its absence (Reason, 2000). It is quite hard to find any effective way to measure it directly as it is the case for conventional measurements related to more tangible issues (Hanakova et al., 2017). Overall safety is intangible system property and even where it is possible to measure it directly, it is often impractical because measuring the things which go right simply means a lot of effort to be spent in order to have meaningful records. Unlike safe state, unsafe outcomes are not only less frequent but they are much more tangible thus considerably easier to track (Hollnagel, 2014). Aviation accidents and incidents attract society from early days of its existence and for decades they were the best driver for safety improvements. As soon as they became rare, the focus just shifted to incidents and safety occurrences with their contributing factors, which, according to investigations, lead to the accidents.

Recently, a new type of data emerged in this domain. Tracking back the root causes of accidents led to the discovery of the so-called organisational factors denoting those contributing factors, which stem from how safety management and safety oversight work (ICAO, 2013). Until the discovery of the importance of how aviation organisations and regulatory bodies are set up as entities, no safety management system nor any sophisticated safety oversight were needed. Progressive requirements for gathering how organisations and regulatory bodies approach safety from management perspective appeared first around the year 2010 (European Commission, 2010; EASA). These requirements established datasets different in their very fundamentals; they assess activities which can hardly be associated with specific unsafe behaviour but which are capable of generating background on which unsafe behaviour emerges. Starting to collect this type of safety data was significant milestone for aviation safety as it brought the industry closer to generate the full picture.

Nowadays, we are closer to the full picture as the content of collected data evolved, but due to the insufficient data sharing and confidentiality restrictions, they are typically not available for research and development initiatives. This inhibits the progress of introducing new technology which could integrate and process the data so that all parties would benefit from industry-wide, open data based knowledge. So has the progress to be achieved the other way. Current research initiatives have to make the best use of public but restricted data samples to come with solutions that aviation organisations may trial and which would expedite establishing the full picture.

Data scarcity, however, is not a new issue. There are several studies available to date, which propose methodologies to overcome this issue in different applications. In fact, very few deal with this problem in scope of safety (such as Yu et al., 2017; El-Gheriani et al., 2017, which are only oriented to major accidents); much more frequent are studies oriented to system reliability, failure and risk assessment in terms of data uncertainty and its reduction. Both safety and reliability oriented studies are typically using Bayesian approach in some variations to produce a posterior distribution by combining data, expert knowledge or various simulation results. Among other methods, first order reliability method and Monte Carlo simulation (Awadallah et al., 2016), or grey system theory (Wen et al., 2011) are used in respective applications. Special attention in the literature is paid to expert elicitation, which was already formalised in several publications (such as Meyer, 2001; Keeney and von Winterfeldt, 1991 or Aven and Guikema, 2011). All the methodologies are, however, difficult to apply directly on the problem in this work as they require various inputs which are out of the scope of this paper. The problem here is of more generic nature, even though it can be complemented with the methods from other studies.

With respect to the afore-mentioned, this article describes the public aviation safety data in detail and provides solutions for how to overcome their limitations. It suggests generating either synthetic aviation safety data or resampling the data already available. The motivation to use data resampling is based on the need to decompose existing signals to increase their granularity for the purpose of further processing and analysis. Data simulation complements this approach by extending the possibility to generate entirely synthetic signals.[1] Synthetic data have their apparent limitations but the important aspect is that they can enable application of advanced analyses, even for experimental or learning purposes only, where real data do not allow it. Direct application of mathematical tools and methods, such as statistical inferential procedures, autoregression or recurrent analysis, to make inferences about safety performance (the full picture) would be otherwise impossible. To enable the tools and methods, it is important to resample the data, i.e. to transform annual figures into month, week or day distribution. For cases where no data are available, simulation based on expert assumptions can provide the solution.

Taking into account the goal, this paper deals with methodology of both data resampling and simulation. It describes data and identifies the gap for improvement. The methods are applied on selected figures from real datasets in the domain of aviation safety. At the end, aviation safety performance is computed using the resampled data to exemplify the contribution of the proposed solution.

2. Methods

This section details the proposed methodology to achieve the goal of this paper. At first, aviation safety data are specified, including their sources, relevant issues and examples. At second, data resampling follows with description of key principles of how to combine expert knowledge and real datasets to increase data granularity by the means of mathematical functions. Lastly, after the outline of data resampling principles, the methodology further specifies data simulation in order to extend the principles of generating synthetic data to situations where no real data are available.

2.1. Data characteristics

Aviation safety data comprise accidents, incidents and safety occurrences. The data are available in form of aggregated figures denoting number of observations of respective accident, incident or occurrence during given time interval. Additionally, new data types were recently introduced to aviation through the European Union-wide (EU-wide) safety key performance indicators (SKPIs) (European Commission, 2013), which are based on the so-called organisational factors. However, these are using artificial scores and due to their novelty, inherent bias and lack of relevant expert assumptions, they are not considered in the methods of this study.

Aviation accident records were gathered reliably till now and they are publicly available together with investigation reports, including conclusions and corrective measures. These data can be found on website of responsible body for respective investigation.[2] But because aviation accidents became rare, they solely cannot be used for safety management today. In terms of any research and development initiatives, much more valuable are data concerning incidents and safety

[1] For further reading on data resampling and simulation methods refer to Lahiri (2003) and Carsey (2014)).

[2] Such as Air Accidents Investigation Institute (2017) in the Czech Republic or Bundesstelle für Flugunfalluntersuchung (2017) in Germany.

occurrences. These data are published on websites of some aviation authorities, but there are not many yet.

One of the most useful data repositories is provided by European Organisation for the Safety of Air Navigation (EUROCONTROL) on its dedicated performance monitoring websites [3,4,5] and in annual safety- and operations-related reports (Safety Regulation Commission, 2016; EUROCONTROL, 2016). EUROCONTROL provides EU-wide aggregated overview of the most common safety issues in the domain of Air Traffic Management (ATM) and ANSPs in form of interactive dashboards (see Fig. 1) together with many other overall performance-related indicators, such as complexity scores, flight delays, traffic distribution etc. Because the most detailed aviation safety data are provided from this domain, they are used to exemplify generating synthetic data.

At the highest level of detail, ATM related safety data are published on (a) Separation Minima Infringements; (b) Unauthorised Penetrations of Airspace; (c) Runway Incursions and (d) ATM Specific Occurrences. The data include severity distribution for the most severe events (severity A - serious incident and severity B - major incident, as defined in (EUROCONTROL, 1999)) and are available back to the year 2004. Federal Administration Authority (FAA) publishes regularly reports of similar quality in the U.S., but unlike in Europe, no data on organisational factors (structure) are provided. In Europe, the data can also be obtained directly from providers' annual reports but these are not all consistent in their content. Some providers are more advanced in safety and others are less, which results in each ANSP publishing different data.

Table 1 demonstrates the Separation Minima Infringements (SMI) in total numbers from year 2008 to 2015. It shows EU-wide figures of these occurrences, where severity A and B Infringements are extracted and stated separately because they represent the most severe outcomes of this type of occurrence.

For aviation organisations other than ANSPs there are almost no data accessible. Owing to the recent initiatives to establish common reporting scheme in the EU (European Commission, 2014), some data from other organisations are already available on the EU level and basic statistics and knowledge were extracted into newest annual safety review by European Aviation Safety Agency (EASA) (EASA, 2016). Compared to the EU-wide data published by EUROCONTROL, however, it does not provide much level of detail, such as distribution per year and month, or per country and airport.

With respect to the mentioned facts, this study demonstrates the basic principles of resampling using data from EUROCONTROL's repositories, which relate to the listed occurrences measured at the highest level of detail. In fact, its data exclusively can be resampled with no need for complementing them with data from other stakeholders to be able to test, for instance, statistical and stochastic tools to analyse the data.

2.2. Data processing

Data processing can be performed using two methods: data resampling and data simulation. The selection of appropriate method depends on following conditions. The first is real data accessibility and the second is expert assumptions availability. In this study, data resampling is used only if real data are accessible and at least some expert assumptions are provided. Data simulation is used to synthesise data vectors where no real data are accessible but expert assumptions exist. It is possible to use the simulation also in case where no data nor any expert assumptions are available but then the output may be highly questionable.

[3] http://ansperformance.eu/.
[4] http://www.eurocontrol.int/prudata/dashboard/eur_view_2014.html.
[5] http://www.eurocontrol.int/prudata/dashboard/rp2_2015.html.

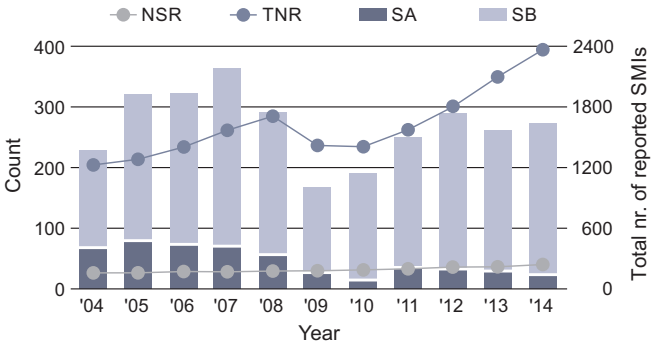

Fig. 1. Separation Minima Infringements (SMI) dataset with number of states reporting (NSR), total number of records (TNR), number of reports with severity "A" type (SA) and severity "B" type (SB).[1] Note that only TNR refers to the right-most y-axis.

Table 1
Separation Minima Infringements (SMI) distribution per year and highest severity.

	2008	2009	2010	2011	2012	2013	2014	2015
SMI severity A	56	24	16	35	29	30	23	20
SMI severity B	236	141	178	217	258	232	250	228
SMI total rep.	1711	1418	1402	1571	1796	2161	2359	2316

2.2.1. Data resampling

First method transforms real data into desired distribution with the help of expert assumptions. Typically, resampling is needed when data granularity is to be increased; even the most detailed data on safety occurrences in aviation are public only as annual figures of occurrence observations but at least distribution by month or week is needed for the deployment of advanced mathematical methods. To resample the data, expert assumptions are to be made before the resampling process starts. In general, regarding safety occurrences similar to SMI, it is true that (a) occurrence rate is higher in summer than in winter; (b) the higher the total amount of reports per year the bigger the difference between peak and trough values; (c) occurrence observations should correspond to the traffic distribution, i.e. maximum number of observations is most likely in July and minimum in January and (d) values are to be natural numbers or zero.

The assumptions are based on following facts. Occurrence rate assumption stems from the fact, that the higher the traffic saturation, the higher the probability of a conflicting situation. This is especially true for today's volumes of traffic reaching maximum capacity of existing airspaces in Europe (Lehouillier et al., 2016) and it is indicated by increasing complexity scores of Europe's ANSPs (EUROCONTROL, 2017). Traffic saturation is known to be seasonal, what can be inferred from traffic figures clearly indicating regular peak values around July and troughs around January, justifying the second assumption. Absolute difference between peak and trough values of occurrence observations during a year can hardly be constant for all safety occurrences; occurrences with hundreds of observations per year should have the difference amplified by their magnitude, causing it to increase. Occurrences with no more than 10 observations per year must remain within their limits. Last assumption relates to the format of occurrences. Any occurrence is a binary variable; either there is an occurrence or there is no occurrence. It is clear that there cannot be negative number of observations and any other than natural number or zero is not conceivable in real world. All these assumptions are general enough to be universal and valid for all safety occurrences similar to the SMI, i.e. for all occurrences on which the data are currently accessible. This is mainly due to the binary property of monitored safety occurrences and their close relation with traffic saturation, especially when reaching limits of given airspace.

Taking into account all these assumptions, following equation

provides basic solution to the problem, where the goal is to resample data into any other distribution with higher granularity:

$$N = \int_0^M k \cdot f(x)dx, \tag{1}$$

where N is annual total number of selected occurrence observations, M is scale determined by the new distribution to be produced, k is coefficient of seasonal variance, x represents time and $f(x)$ is time-dependent mathematical function capturing expert assumptions. Scale M is determined by total number of data points to be produced (e.g. $M = 12$ for data distributed by month whilst N is the total figure per year). Coefficient k may be calculated in many ways but it should be in line with provided expert assumptions. If there are no expert assumptions on the coefficient, one of the possible ways to calculate it is to use average occurrence rate N/M as a starting point because the variance typically depends on this rate: the higher the number of observations per event type, the higher the variance. For datasets by EUROCONTROL, the variance is unknown and no expert assumptions can be considered, so the problem is then shifted to the coefficient k. Based on empirical testing in MATLAB environment (MATLAB R2015b, MathWorks, Inc., Natick, MA,USA) for the purpose of this study, reasonable results with the data samples from Table 1 were achieved with $k = 0.25 \cdot N/M$ (k amplifies $f(x)$ by 25% of the average occurrence rate N/M). The coefficient may be set differently at ones discretion so that the results copy as much as possible what is supposed to be real.

With regard to the Eq. (1), new data distribution can be calculated as follows:

$$N_i = [k \cdot F(x)]_{i-1}^{i}, \quad i = 1,2,...,M, \tag{2}$$

where N_i is number of occurrences during selected time interval i and $F(x)$ is anti-derivative of the integrand (function $f(x)$). Obviously, N_i needs to be rounded in order to fulfil the last assumption about natural numbers or zero. If deemed appropriate, Eq. (2) may be complemented with white noise, which makes the resampling more realistic. The noise can be of any distribution but because Gaussian white noise is good approximation of many real-world situations (Yanushevsky, 2007), it is preferred in this study. Gaussian white noise can be generated using pseudorandom component of Gaussian distribution with mean 0 and variation equal to 1 (such pseudorandom numbers can be produced by MATLAB or similar software). The component is based on the following equation:

$$\vec{\epsilon} = p \cdot \vec{u_i}, \quad \vec{u_i} \sim N(\mu,\sigma^2), \tag{3}$$

where $\vec{\epsilon}$ is vector of final white noise components, $\vec{u_i}$ is vector of pseudorandom Gaussian distributed numbers with mean $\mu = 0$ and variance $\sigma^2 = 1$ and p is noise effect coefficient. The coefficient p amplifies the noise as needed. If the expert assumptions do not include any information about the noise, the variable p should be so that the output will be reasonable, i.e. no extreme differences between each two consecutive resampled points are achieved but on the other hand, the function $f(x)$ should not be clearly visible. In addition, the coefficient needs to be variable with the magnitude of occurrence observations, because the same noise cannot influence data with hundreds of occurrences per given time period in the same way as those with no more than ten. Therefore, p needs to be expressed rather as ratio, dependent on the average number of occurrences of given event type, multiplied by constant as follows:

$$p = r \cdot \frac{N}{M}, \tag{4}$$

where N is number of selected occurrence observations of original distribution, M is scale determined by the new distribution to be produced and r is constant to be set. Experiments performed in this study estimated the value for $r = 0.125$ to fit well the EUROCONTROL data repositories but it may be set different for other cases. The sum of all N_i may not precisely be equal to the real values of N due to rounding the results and adding the noise, but it should remain acceptably close for all cases. This also means that there should not be too much noise added, otherwise the resampling output may exceed reasonable limits.

For the particular expert assumptions introduced in this chapter, data seasonality may be modeled by sinus function:

$$N = \int_0^M k \cdot \sin\left(x \cdot \frac{2\pi}{M} - \frac{\pi}{2} - \frac{2\pi}{M}\right)dx. \tag{5}$$

The sinus uses the expression

$$x \cdot \frac{2\pi}{M} - \frac{\pi}{2} - \frac{2\pi}{M}, \tag{6}$$

to move the extreme values on the interval $\langle 0,M \rangle$ so that its maximum is achieved at the point of $7 \cdot M/12$ (July data) and the minimum at $M/12$ (January data). The sinus function is shifted upwards by constant of integration so that no values are negative. Recalculating new occurrence distribution during given year will, therefore, follow the equation:

$$N_i = \left[-k \cdot \cos\left(x \cdot \frac{2\pi}{M} - \frac{\pi}{2} - \frac{2\pi}{M}\right)\right]_{i-1}^{i}, \quad i = 1,2,...,M, \tag{7}$$

where N_i is number of selected occurrence observations during month i in selected year, M is scale determined by the new distribution to be produced, k is coefficient of seasonal variation and i is successive time step of the series from new distribution.

However, problem may arise as soon as specific requirement exists for resampled data distribution. No data distribution is assured by Eq. (7) but empirical testing showed that Gaussian and various skewed distributions are randomly obtained with the sinus function and k. Safety occurrences in aviation are assumed to follow non-Gaussian distribution (Seshadri, 1998; Wang et al., 2014) which also seems to be the case in other industries, where inverse Gaussian distribution fits incidents and lognormal distribution fits less severe but more frequent non-conformances (Love et al., 2015). Unfortunately, inverse Gaussian distribution could not be obtained from Eq. (7) and there is no general transformation function by which such distribution could be obtained e.g. from Gaussian distributed random variable (Chhikara, 1988), which is frequent product of the equation. The basic solution in Eq. (2) may produce different data distribution with different $f(x)$ and k and so has the investigator first check the distribution of the output from Eq. (2) and then add white noise with appropriate distribution in order to obtain desired distribution of the resampled data. Due to the complexity, however, this may not always be possible.

To demonstrate the resampling method as applied on aviation safety occurrences (Eq. (7)), at this point there is missing only a real figure of annual occurrences of selected event type (variable N_i) and the final decision about how many points are to be obtained from the figure (variable M). In this paper, SMI severity B recorded number of occurrences for year 2011 within the EUROCONTROL region was randomly selected ($N_i = 217$ occurrences) and this figure was resampled into monthly-distributed dataset of 12 figures ($M = 12$) for each month during 2011. The results are in shown in Section 3.

2.2.2. Data simulation

As long as there are no data available and it is important to generate some, assumptions have to additionally include what is available for resampling, i.e. occurrence observation figures. Experts to provide such assumptions are preferred to be front line personnel as they can usually estimate how frequent some occurrences are. For example, an Air Traffic Control Officer (ATCO) can estimate how many times a day or a week does he or she experience Short Term Conflict Alerts (STCA), alerting him or her to some aircraft being on collision course, whether horizontally, vertically or both. Usually, ATCO can also estimate how much does this value vary during a year, providing an estimation for variability as well.

Key principles of the simulation remain the same as for data

resampling, but this time the core lies with pseudorandom number generation. Concerning data simulation, the pseudorandom component will not simulate noise only, but the entire dataset. The distribution parameters are to be fitted to the expert or front line personnel assumptions on the occurrence. The basic solution for data simulation is then as follows:

$$E_i = ||\vec{e_i}|| = \sum_{j=1}^{D_i} e_{ij}, \quad \vec{e_i} \sim IG(\mu_i,\lambda_i) \wedge e_{ij} \in \mathbb{N}^0, \tag{8}$$

where E_i is sum of occurrence observations of event type E during time period i,$\vec{e_i}$ is vector of observations during time period i,D_i is number of data points during time period i and e_{ij} is j^{th} element from the vector $\vec{e_i}$. Vector elements are assumed to be natural numbers or zero and obeying inverse Gaussian distribution with mean μ_i and shape parameter λ_i (both variable with i). In this case, no real data exist which would restrict the simulation and so it can be based on truly inverse Gaussian distribution.

The vector $\vec{e_i}$ is to be generated using pseudorandom numbers as MATLAB or similar software can produce. Average value μ_i and its estimated variance can be provided by an expert, but shape parameter λ_i is difficult to obtain. It can only be reliably inferred from real data samples of similar occurrences. Because data in EUROCONTROL's repositories are not sufficient for such analysis, parameter λ_i will be replaced for the purposes of this study to produce single parameter inverse Gaussian distribution as follows:

$$\lambda_i = \mu_i^2. \tag{9}$$

This distribution allows overcoming the issue with unknown λ_i, but eventually it may not be so different from the distribution based on real data. For lower numbers of occurrence observations (μ_i less than approximately 25), the probability density function is similar in shape to how the distribution of aviation safety occurrences is described by Wang et al. (2014), whilst for larger numbers (μ_i more than 25) it is approaching normal (Gaussian) distribution. Mean μ_i greater than 25 can be prevented simply by utilising the pseudorandom element to simulate data "on daily basis" as it is the case in Eq. (8), where weekly or monthly data can be produced as a sum of daily simulation. This is possible due to additive property of the distribution according to which the sum of inverse Gaussian distributed random variables produces another inverse Gaussian distributed variable under given conditions (Chhikara, 1988). According to all EUROCONTROL's datasets, it is very unlikely, that there would be on average 25 or more observations per an occurrence a day. This way, the desired properties of inverse Gaussian distribution are preserved and can be used to simulate synthetic data. However, the noise induced by the omission of actual λ_i may be significant in some cases, and so should such a simulation be used only when necessary and only for testing of mathematical models, analytical tools etc. The desired noise is added by rounding the values to achieve natural numbers or zero but this may change the distribution. It is therefore highly advisable to perform tests of the produced statistics before the data are used.

To demonstrate the simulator, fictional assumptions (a) STCA is experienced on average 2 to 3 times a day; (b) the average occurrence rate during peak days is by 1 occurrence more a day, and vice versa, the average during trough day is by 1 occurrence less a day; will serve as the basis to synthesise new data.

STCA is a safety occurrence similar to SMI. In fact, it relates to SMI because it is supposed to alert ATCO to prevent SMI or similar situations in advance, but obviously there must be more STCA warnings than SMIs, because STCA under normal operational conditions precedes SMI and only after the conflict is unresolved by ATCO, SMI can emerge. Other assumptions are therefore the same as in the example with data resampling.

The assumptions are to be taken into account in similar way as for data resampling, i.e. by the means of mathematical functions, which quantify the assumptions. For the STCA assumptions, Eq. (8) was complemented with the following equations, using the same sinus function to model data seasonality:

$$\mu_i = k\cdot\sin\left(i\cdot\frac{2\pi}{M}-\frac{\pi}{2}-\frac{2\pi}{M}\right) + x_E, \quad i = 1,2,...,M, \tag{10}$$

$$k = \frac{\max(x_e)-\min(x_e)}{2}, \tag{11}$$

where μ_i is the distribution mean during time step i,k is distribution mean variation, M is scale determined by the new distribution to be produced and x_E is expert assumption on process intercept. Given the assumption on occurrence rate for STCA, the process intercept value (x_E) is 2.5 per day. Coefficient of seasonal variation k can be calculated as the difference between its maximum and minimum estimated value, as follows (by the means of Eq. (11)).

$$k = \frac{3.5-1.5}{2} = 1 \tag{12}$$

If the goal is to generate data distributed by month, D_i corresponds to number of days per each month from January to December and M is equal to 12. Likewise, many other assumptions may be included, which can set different λ_i,x_E and μ_i or even set requirement for different probability distribution of the simulated data. The simulator does not aim to provide solution for every possible scenario but rather provide key principles on how to simulate safety data by reusing the principle of data resampling from previous section.

At this point, the simulator Eqs. (8)–(11) can be used to generate synthetic data for STCA event type according to the afore-mentioned expert assumptions (a) and (b). Variable M was set to 12 as in case of data resampling, D_i included number of days of each month during average year (28 for February) and variable k was set to 1 in line with Eq. (12). The results are in shown in Section 3.

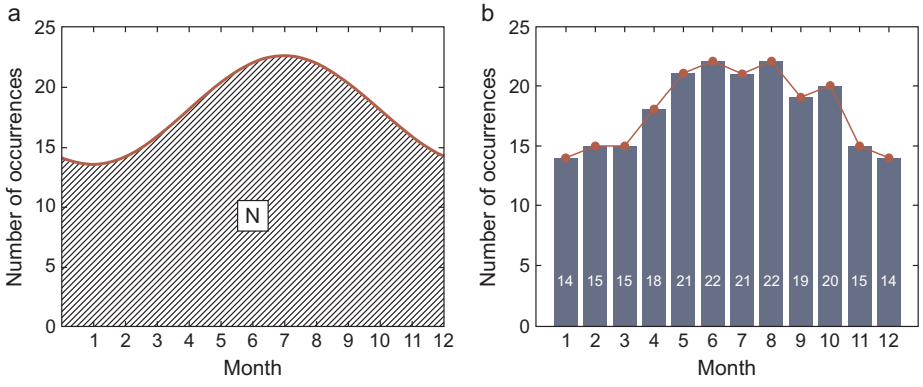

Fig. 2. Basic solution for given assumptions and SMI severity B in 2011 (a) and generated monthly-distributed data according to the basic solution for SMI severity B in 2011 (b).

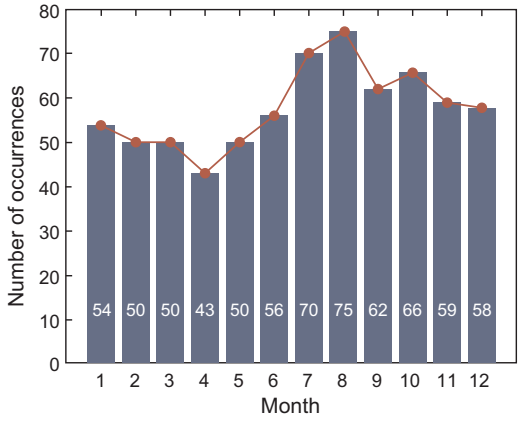

Fig. 3. Example of simulated data for STCA event type.

3. Results

The results of data resampling are depicted in Fig. 2. They are based on SMI severity B from year 2011. Fig. 2a demonstrates the sinus function behind the data resampling equations, where the area below the sinus curve and x-axis equals total number of SMI severity B observations in the year 2011. Fig. 2b depicts resampled data. Pseudorandom component (as per Eqs. (3) and (4) with $r = 0.125$) was added and so the distribution does not follow the sinus function too precisely as it is assumed to be the case during real operations. The data represent occurrence observations distributed by month of the year 2011. Data distribution remains random in this case.

The results of data simulation are on Fig. 3. The simulation is based on the fictional assumptions about STCA event type and the results are distributed by month of a fictional year. For February, 28 days are assumed in this example and the data obey inverse Gaussian distribution.

4. Discussion

The sum of all resampled occurrences on Fig. 2b is 231 which, compared to the real data of 217 occurrences in 2011, shows that the sum of the error induced by rounding and adding the noise was 6.45% thus not so significant.

Concerning data simulation, due to quite a lot of uncertainty put in the simulator (all the assumptions together), each time it runs it usually produces a notably different curve or histogram. This may not always correspond to the reality and thus shall not be preferred over data resampling, but the results make it possible to learn how to build or to trial different methodologies or advanced models where no other options exist. In any case, it is advisable to check on regular basis with experts or front line personnel how the trends evolve in order not to have the simulated data based on obsolete assumptions.

It is possible to use different or even multiple functions $f(x)$ instead of the sinus used in this study for data resampling or simulation equations. However, in such case, it is important to carefully quantify qualitative statements, which produce such need and insert them into the equations as either coefficients, constants or mathematical functions. This study does not aim to provide solutions for any possible case that may exist, but it rather outlines and exemplifies how such simulator and data resampling works, providing general solution for most common issues. On the other hand, the general nature of the proposed methods provides an option for their implementation in other transportation domains or high-risk industries.

When considering the results in terms of other research performed especially in reliability engineering, where similar problems with data unavailability appeared, the overlap with this study regards using more robust functions $f(x)$ or parameter estimation of more optimal function than sinus used in this work. Bayesian approach of integrating different data sources and expert knowledge works with probability density functions of parameters typically pertaining different variables (inputs) composing a regression model to predict future output. This study is, however, focused only on the mathematical models and their application on increasing data granularity. The core principles are also demonstrated on data simulation, but the input available in this study is very limited to allow for robust approach in producing mathematical models. If the inputs necessary to use such modelling are available, Bayesian approach and other methodologies applied to data scarcity can be used to produce more complex and precise model for generating or resampling data. Likewise, additional improvement can be achieved by robust expert elicitation, following the published frameworks suitable for particular application.

Both resampled and simulated data are suitable for applications only as entire datasets. This is because local differences between two consecutive data points may not correspond to the reality at all either due to inaccuracies in expert assumptions or due to added noise, and if only a selection of such data is used for building mathematical models, this may be completely misleading. Therefore, it is highly recommended to use entire datasets and not only their subsets.

As an example of application of the methods in this study, reconstruction of Aerospace Performance Factor (APF) according to the methodology developed by FAA (Lintner et al., 2009) will be demonstrated. Generally, the APF is one of the system-wide information, which can be produced by composing safety data into a single data point, which is intended to quantify level of system's safety performance. Concerning the data published by EUROCONTROL, the APF signal can be reconstructed for several years and in this example, it is calculated from 2008 up to 2015. According to the APF methodology, required are (a) resampled data on selected EU-wide safety occurrences into distribution by month and (b) EU-wide traffic distribution data by month in total hours flown format.

For the selected time period, data on safety occurrences from all EUROCONTROL data repositories were subjected to resampling. The real data sample comprises only 8 data points per each occurrence (distributed by year), i.e. 96 data points per each event type were achieved by the resampling process. It is to be noted here, that EUROCONTROL used for their APF calculation larger datasets concerning the safety occurrences included in the calculation (Neubauer and Lintner, 2010) whilst in this study only the occurrences provided in the public data repositories were used. The reason for omitting majority of safety occurrences used by EUROCONTROL is the complete unavailability of respective safety data. The missing data were not simulated due to that only rough assumptions could be provided. On the other hand, the most critical safety occurrences are included in the public repositories and so simulating the rest of the data could actually introduce more noise to the reconstructed signal than just omitting the data. Therefore, in this case, data resampling was preferred over data simulation. The impact of this decision can be verified through the comparison of the reconstructed signal with EUROCONTROL's output (Fig. 4), because both include year 2008.

With regard to the traffic distribution, the public repositories include annual total figures for flight hours in the EU region, but only for year 2015 the distribution by month is available. On the other hand, the same data for traffic distribution in the EU region are available using different unit, namely average daily movements, for the entire time period both as annual figures as well as figures distributed by month. The procedure to obtain distribution by month for years 2008 up to 2014 needs no resampling because the real data are there and just need to be converted into different units. Most important is to obtain the ratio between the figures as follows:

$$R_m = \frac{ADM_m}{THF_m} \tag{13}$$

where R_m is the calculated ratio, *ADM* is traffic in average IFR daily movement format, *THF* is the same figure in total flight hours format

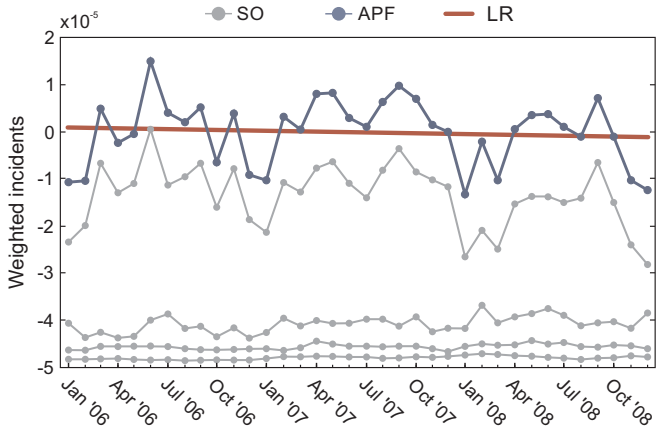

Fig. 4. APF signal based on real data from 2006 to 2008 (Lintner et al., 2009).

and m stands for respective month. Because the traffic figures in total flight hours format distributed by month are accessible for year 2015 only, the ratios R_m can be calculated using data from year 2015 only. Obtained ratios serve then as coefficients to recalculate all the years backwards using Eq. (13) to obtain monthly-distributed traffic data in total hours flown format.

At this stage, all variables are known and the APF signal can be reconstructed. Fig. 5 depicts the results. For the year 2008, reconstructed APF signal is similar to the one on Fig. 4. EUROCONTROL used relative APF figures, which are adjusted to the process mean whilst Fig. 5 demonstrates absolute APF figures, which cause shifting the scale of y-axis. Some difference can be observed, which is certainly attributable to the difference between the data behind each calculation, but comparing the outputs for the year 2008, the two signals are convincingly similar in shape and magnitude.

Last point to discuss is the new type of data on organisational factors. They are publicly available in Europe only, measured from 2012 and referring to the three EU-wide SKPIs, measured at both national and ANSP level. The data contain information on (a) Effectiveness of Safety Management; (b) Application of Just Culture and (c) Risk Analysis Tool (RAT) methodology usage.

This dataset is limited compared to the accidents and occurrences due to its novelty. It is available on the same EUROCONTROL websites together with accidents, incidents and occurrences but methodology and format of these data is obviously different from safety occurrences. To evaluate these SKPIs, artificial scores are used, represented by percentage derived from self-assessment questionnaires (see EASA). These questionnaires, however, provide certain room for bias, and so the data are not comparably accurate to the safety occurrence records. Considering this new type of safety data, no such data resampling or simulation can be used as for safety occurrences. These data are not seasonal nor do they depend on the volume of traffic etc. Their dynamicity is very low; according to the dashboards at EUROCONTROLs websites they tend to change a bit year after a year, but it is quite normal as they refer to things, which are hard to change (such as fundamentals of safety management system), and which are seasonally independent. At this point, resampling the data would more or less just follow even distribution with some linear trend during all the year thus it makes no sense to pay special attention to them. Should this change in the future and new assumptions could be drawn, then similar principles as used in the examples in this paper can be reused to build dedicated simulator for these datasets.

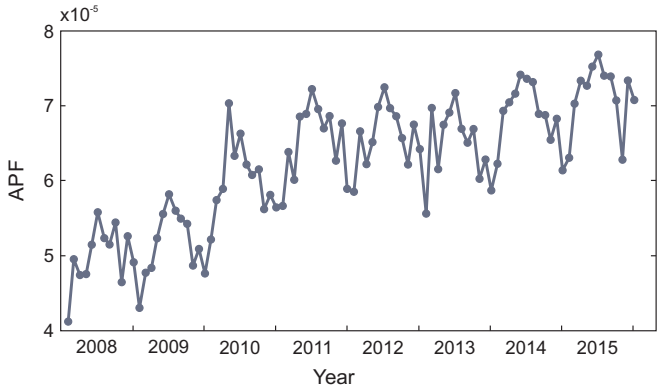

Fig. 5. APF signal based on public data repositories with applied data resampling from 2008 to 2015.

5. Conclusions

Restrictions concerning aviation safety data and their availability lead to the search for solutions, which are capable to overcome them. There are cases in which almost all safety data are accessible and just few data points need to be acquired via data resampling; in other cases there are very little or no safety data available and so they need to be simulated using expert assumptions only. The former can help to verify new methodologies or advanced modelling as they are likely to achieve comparable results with real data; the latter makes it possible to learn how to build or to trial the same methodologies or advanced models as in the former case. Both cases are usable for modern research and development activities in the domain of aviation safety but due to their general nature, they can find application in other transportation domains or high-risk industries. Because the data to be simulated or resampled in aviation are related to socio-technical system, expert assumptions are often of critical importance and are to be considered adequately.

This paper drew basic principles and solutions to the above-mentioned problems. Using basic mathematical functions, expert assumptions were transformed into sets of equations. Were real data were accessible, the equations considered them. Typical problem with such data in aviation is that it is available only annually as total figures whilst month or day distribution is desired. Introduced sets of equations were used for data resampling whilst annual total figures were obeyed. Where no data are available, the solution is based on pseudorandom number generation, such as modern computational software can generate. Mathematical functions then complement the pseudorandom number so that it produces conceivable outcome in accordance with expert assumptions.

It is clear that the synthetic data used to fill the gaps of existing limitations will never contain anything outside of what is inserted in the very equations behind the simulation. Even though they are based on expert assumptions and account for randomness, it is not possible to include all the variables, which affect the values of measured aviation safety data. The less real data and expert assumptions there are, the more inaccurate the resampled or simulated data and vice versa. It is important to note that because the aviation is a socio-technical system, it is unlikely that the system is deterministic. Therefore, there is no ultimate set of assumptions and equations, which describe the system completely and so real data should always be preferred. On the other hand, some of these limitations may be reduced by further research, applying methods from different studies dealing with data scarcity, such as Bayesian approach or Monte Carlo simulation, to refine and perfect the mathematical functions used to generate synthetic data in specific applications.

Despite the limitations, the synthesised data make it possible to implement, verify and validate advanced methodologies or analytical tools, which are highly dependent on data sample size. There are constraints stemming from the confidential nature of aviation safety data but because no aviation stakeholder is willing to share them extensively, under the risk of their misuse and with no assurance what will the benefits be, it seems unlikely that this will improve soon. On

the other hand, a chance exists to improve the situation with new technologies and inventions. At this stage, these can be pre-set up and checked using simulated data and then, if proven, used to demonstrate their capabilities to aviation stakeholders, including regulatory bodies. This may eventually resolve the general unwillingness to share and work with safety data jointly and to establish the full picture of aviation safety. As soon as some technology is proven at least on partially real data, it may eventually convince aviation stakeholders to trial it.

Acknowledgement

This work was supported by the Czech Technical University in Prague [junior research Grant No. SGS16/188/OHK2/2T/16]

References

Air Accidents Investigation Institute, 2017. Reports of accidents and incidents (September 2017). <http://www.uzpln.cz/en/reports>.

Aven, T., Guikema, S., 2011. Whose uncertainty assessments (probability distributions) does a risk assessment report: the analysts or the experts? Reliab. Eng. Syst. Saf. 96 (10), 1257–1262. http://dx.doi.org/10.1016/j.ress.2011.05.001.

Awadallah, A.G., Saad, H., Elmoustafa, A., Hassan, A., 2016. Reliability assessment of water structures subject to data scarcity using the SCS-CN model. Hydrol. Sci. J. 61 (4), 696–710. http://dx.doi.org/10.1080/02626667.2015.1027709.

Bundesstelle für Flugunfalluntersuchung, 2017. Investigation reports (September 2017). <https://www.bfu-web.de/EN/Publications/Investigation%20Report/reports_node.html>.

Carsey, T., 2014. Monte Carlo Simulation and Resampling Methods for Social Science. Sage, Los Angeles.

Chhikara, R., 1988. The Inverse Gaussian Distribution: Theory: Methodology, and Applications (Statistics: A Series of Textbooks and Monographs). CRC Press ISBN: 9780824779979.

EASA, 2016. Annual Safety Review 2016, European Aviation Safety Agency, Cologne, Germany, ISBN: 978-92-9210-202-9. doi:http://dx.doi.org/10.2822/541561. <https://www.easa.europa.eu/document-library/general-publications/annual-safety-review-2016#group-easa-downloads>.

EASA. Safety Key Performance Indicators (SKPI)/Acceptable Means of Compliance (AMC) Amendment 1/Guidance Material (GM) Amendment 1.

El-Gheriani, M., Khan, F., Chen, D., Abbassi, R., 2017. Major accident modelling using spare data. Process Saf. Environ. Prot. 106, 52–59. http://dx.doi.org/10.1016/j.psep.2016.12.004.

EUROCONTROL, 1999. ESARR 2 Guidance to ATM Safety Regulators: Severity Classification Scheme for Safety Occurrences in ATM. <https://www.eurocontrol.int/sites/default/files/article/content/documents/single-sky/src/esarr2/eam2-gui1-e1.0.pdf>.

EUROCONTROL, 2016. Annual network operations report 2015 (June 2016). <http://www.eurocontrol.int/sites/default/files/publication/performance/2015_annual/final-edition/annual_network_operations_report_2015_main_report_final.pdf>.

EUROCONTROL, 2017. Pan-European ANS Performance data repository (September 2017). <http://ansperformance.eu/data/performancearea/>.

European Commission, 2010. Regulation (EU) No 996/2010 of the European Parliament and of the Council of 20 October 2010 on the investigation and prevention of accidents and incidents in civil aviation and repealing Directive 94/56/EC. Text with EEA relevance. Official J. Eur. Union OJ L 264, 25–27. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:295:0035:0050:EN:PDF>.

European Commission, 2010. Regulation (EU) No 996/2010 of the European Parliament and of the Council of 20 October 2010 on the investigation and prevention of accidents and incidents in civil aviation and repealing Directive 94/56/EC. Text with EEA relevance. Official J. Eur. Union OJ L 264, 25–27. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:295:0035:0050:EN:PDF>.

European Commission, 2013. Commission Implementing Regulation (EU) No 390/2013 laying down a performance scheme for air navigation services and network functions. Official J. Eur. Union OJ L 128, 1–30. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:128:0001:0030:EN:PDF>.

European Commission, 2014. Regulation (EU) No 376/2014 of the European Parliament and of the Council on the reporting, analysis and follow-up of occurrences in civil aviation, amending Regulation (EU) No 996/2010 of the European Parliament and of the Council and repealing Directive 2003/42/EC of the European Parliament and of the Council and Commission Regulations (EC) No 1321/2007 and (EC) No 1330/2007. Official J. Eur. Union OJ L 122, 18–43. <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014R0376&from=EN>.

Federal Aviation Administration, 2015. Air Traffic Organization: 2015 Safety Report. <https://www.faa.gov/about/office_org/headquarters_offices/ato/service_units/safety/media/2015_safety_report.pdf>.

Hanakova, L., Socha, V., Socha, L., Szabo, S., Kozuba, J., Lalis, A., Vittek, P., Kraus, J., Rozenberg, R., Kalavsky, P., Novak, M., Schlenker, J., Kusmirek, S., 2017. Determining importance of physiological parameters and methods of their evaluation for classification of pilots psychophysiological condition. In: 2017 International Conference on Military Technologies (ICMT). IEEE. http://dx.doi.org/10.1109/miltechs.2017.7988810.

Hollnagel, E., 2014. Safety-I and Safety-II: The Past and Future of Safety Management. CRC Press ISBN: 978-1-4724-2305-4.

ICAO, 2013. Safety management manual (SMM), third ed., International Civil Aviation Organization, Montreal, Quebec, ISBN: 978-92-9249-214-4.

ICAO, 2013. Global Aviation Safety Plan 2014-2016, first ed., International Civil Aviation Organization, Montreal, Quebec, ISBN: 978-92-9249-355-4.

Keeney, R., von Winterfeldt, D., 1991. Eliciting probabilities from experts in complex technical problems. IEEE Trans. Eng. Manage. 38 (3), 191–201. http://dx.doi.org/10.1109/17.83752.

Klein, T., Viard, R., 2013. Process safety indicators in chemical industry – what makes it a success story and what did we learn so far? Chem. Eng. Trans. 31, 391–396. http://dx.doi.org/10.3303/CET1331066.

Lahiri, S.N., 2003. Resampling Methods for Dependent Data. Springer, New York, New York, NY.

Lehouillier, T., Soumis, F., Omer, J., Allignol, C., 2016. Measuring the interactions between air traffic control and flow management using a simulation-based framework. Comput. Indust. Eng. 99, 269–279. http://dx.doi.org/10.1016/j.cie.2016.07.025.

Lintner, T., Smith, S., Lieu, A., Cioponea, R., Stewart, S., Majumdar, A., Dupuy, M.-D., 2009. The measurement of systemwide safety performance in aviation: three case studies in the development of the aerospace performance factor (apf), vol. 2, pp. 1060–1104. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-77952276172&partnerID=40&md5=09f92af816dc732b263f9e5d7c9b1465>.

Love, P.E., Teo, P., Carey, B., Sing, C.-P., Ackermann, F., 2015. The symbiotic nature of safety and quality in construction: incidents and rework non-conformances. Saf. Sci. 79, 55–62. http://dx.doi.org/10.1016/j.ssci.2015.05.009.

Meyer, M., 2001. Eliciting and Analyzing Expert Judgment: A Practical Guide. Society for Industrial and Applied Mathematics and American Statistical Association, Philadelphia, PA.

Neubauer, K., Lintner, T., 2010. The APF: Using the aerospace performance factor to measure safety performance, pp. 319–372. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-77958496643&partnerID=40&md5=5405cab56d6cfd2caf2db2bde762a891>.

Niu, D.X., Song, Z.Y., 2013. Research on nuclear power plant operational safety performance based on confidence level and fussy evaluation model. Appl. Mech. Mater. (AMM) 475–476, 1721–1724. http://dx.doi.org/10.4028/www.scientific.net/amm.475-476.1721.

Reason, J., 2000. Safety paradoxes and safety culture. Inj. Control Saf. Promot. 7 (1), 3–14. http://dx.doi.org/10.1076/1566-0974(200003)7:1;1-v;ft003.

Safety Regulation Commission, 2016. Src document 55: Annual safety report 2015 (January 2016). <https://www.eurocontrol.int/sites/default/files/article/content/documents/single-sky/src/src-docs/src-doc-55-e1.0.pdf>.

Seshadri, V., 1998. The Inverse Gaussian Distribution: Statistical Theory and Applications (Lecture Notes in Statistics). Springer ISBN: 978-0-387-98618-0.

Socha, V., Socha, L., Szabo, S., Nemec, V., accidents, Air, 2014. their investigation and prevention. eXclusive e-J. 1–9. <http://www.exclusiveejournal.sk/files/4-2014/1-socha-socha-szabo-nemec.pdf>.

Wang, C., Drees, L., Holzapfel, F., 2014. Incident prediction using subset simulation. In: Proc. of ICAS 2014 29th Congress of the International Council of the Aeronautical Sciences, International Council of the Aeronautical Sciences, pp. 1–8, ISBN: 3-932182-80-4.

Wen, Z.H., Zhou, J., Jia, M.X., 2011. Study on relation of structural reliability calculation and fuzzy mathematics. Adv. Mater. Res. 243–249, 5739–5744. http://dx.doi.org/10.4028/www.scientific.net/amr.243-249.5739.

Yanushevsky, R., 2007. Modern Missile Guidance. CRC Press ISBN: 9781420062267.

Yu, H., Khan, F., Veitch, B., 2017. A flexible hierarchical bayesian modeling technique for risk analysis of major accidents. Risk Anal. 37 (9), 1668–1682. http://dx.doi.org/10.1111/risa.12736.