Reliability Engineering and System Safety 209 (2021) 107371

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Bayesian network modeling of accident investigation reports for aviation safety assessment☆

Xiaoge Zhang, Sankaran Mahadevan *

Department of Civil and Environmental Engineering, School of Engineering, Vanderbilt University, Nashville, TN, 37235, USA

ARTICLE INFO

Keywords:
Bayesian network
Safety assessment
Air transportation system
System safety
National Transportation Safety Board

ABSTRACT

Safety assurance is of paramount importance in the air transportation system. In this paper, we analyze the historical passenger airline accidents that happened from 1982 to 2006 as reported in the National Transportation Safety Board (NTSB) aviation accident database. A four-step procedure is formulated to construct a Bayesian network to capture the causal relationships embedded in the sequences of these accidents. First of all, with respect to each accident, a graphical representation is developed to facilitate the visualization of the escalation of initiating events into aviation accidents in the system. Next, we develop a Bayesian network representation of all the accidents by aggregating the accident-wise graphical representations together, where the causal and dependent relationships among a wide variety of contributory factors and outcomes in terms of aircraft damage and personnel injury are captured. In the Bayesian network, the prior probabilities are estimated based on the accident occurrence times and the aircraft departure data from the Bureau of Transportation Statistics (BTS). To estimate the conditional probabilities in the Bayesian network, we develop a monotonically increasing function, whose parameters are calibrated using the probability information on single events in the available data. Finally, we develop a computer program to automate the generation of the Bayesian network in compliance with the XML format used in the commercial GeNIe modeler. The constructed Bayesian network is then fed into GeNIe modeler for accident analysis. The mapping of the NTSB data to a Bayesian network facilitates both forward propagation and backward inference in probabilistic analysis, thereby supporting accident investigations and risk analysis. Several accident cases are used to demonstrate the developed approach.

1. Introduction

The air transportation system has witnessed a steady and fast growth worldwide over the past few decades. According to the International Air Transport Association (IATA), the commercial airline carriers currently transport 1 billion passengers within the U.S. a year [1,2]. In consideration of the catastrophic consequences of aviation accidents, it is highly important to investigate how small incidents escalate as aviation accidents by identifying the propagation pathways of abnormal events in the system; historical aviation accidents can be analyzed to achieve this goal, so that we can better prevent them in the future.

In the air transportation system, the principal mission of safety programs and safety analysts is to ensure that a flight departs from the origin airport and lands safely at the destination airport. A wide range of activities and events affect the safety of this journey (such as aircraft manufacturing, aircraft maintenance, cargo and passengers loading, air traffic control, convective weather, and pilot performance). Complex and intricate interdependencies emerge among the constituent activities related to the safety of the air transportation system [3]. Incidents can occur related to any of these activities anytime, and might escalate into severe accidents alone or in combination with other causal factors (e.g., human fatigue/errors, inadequate training, equipment failure), thereby resulting in catastrophic consequences (e.g., loss of control, fire, collision, etc.) depending on the phase of flight in which they occur (e.g., climb, cruise, taxi, take-off etc.).

To reduce the occurrence rate of the low-probability high-consequence (lp/hc) aviation accidents, there is an imperative need for the development of a systematic approach to examine the interrelationships of causal factors, to guide accident investigations, to identify intervention strategies, and to prioritize potential safety measures (e.g., resource allocation). Towards this end, the knowledge

☆ The source codes of this paper is publicly accessible at GitHub: https://github.com/zxgcqupt/NTSB\protect\T1\textunderscoreBayesian\protect\T1\textunderscoreNetwork.

* Corresponding author.
E-mail addresses: zxgcqupt@gmail.com (X. Zhang), sankaran.mahadevan@vanderbilt.edu (S. Mahadevan).

https://doi.org/10.1016/j.ress.2020.107371
Received 21 February 2020; Received in revised form 15 October 2020; Accepted 26 November 2020
Available online 17 December 2020

acquired from the past accidents might be valuable in managing the risk and preventing hazardous event escalation for future flights if the relationships among the causal factors are represented systematically and the uncertainty arising from multiple sources (e.g., data scarcity, or lack of knowledge) is characterized and propagated in the inference quantitatively. The Bayesian network has emerged as a popular probabilistic approach of capturing causal relationships among random variables [4–8], in which random variables and their dependencies are represented as nodes and links in the network, and it offers a rigorous mechanism for probabilistic inference and a systematic framework for uncertainty representation and propagation in complex systems [9–13].

Compared with traditional risk analysis approaches, such as Fault Trees and Event Trees, the Bayesian network has an important advantage: when additional information (e.g., data, evidence) on some random variables is available, the Bayesian network allows us to propagate such information in both forward propagation and backward inference, thereby supporting dynamic update of our belief on the system state. Due to its promising features, Bayesian networks have been widely studied in the literature to analyze the safety of a wide range of engineering systems [14–17]. For example, Zhang et al. [18] constructed a Bayesian network from historical data to analyze the characteristics of navigational risk in the Yangtze River, and identified the parameters and conditions that had the greatest impact on the accident outcomes. Wang and Yang [19] developed a Bayesian networks-based approach to assess the accident severity in waterborne transportation. Chen et al. [20] proposed a decision-making framework based on Bayesian networks for rear-end collision risk prediction in automobiles considering influential factors, such as weather, road, vehicle speed, and driver reaction capacity.

Likewise, in the domain of aviation, Bayesian network-based causal models have been extensively studied for system risk management and enhancing aviation safety [21–23]. For example, Luxhøj et al. [24] developed an Aviation System Risk Model (ASRM) for the risk assessment of organizational factors in the aviation system. Ale et al. [25,26] developed a causal model for air transport safety (CATS) to provide insights into cause–effect relationships in the event sequences leading up to potential incidents and accidents. Greenberg et al. [27] demonstrated the use of the Bayesian network in evaluating the accident probability of large passenger aircraft, where environmental factors, temporary and permanent characteristics of each pilot were considered in a structured way. Stamatelatos et al. [28] demonstrated the quantification of various uncertainties in a probabilistic risk assessment (PRA) analysis through the Bayesian network. Ancel et al. [29] developed an object-oriented Bayesian network to integrate safety risks contributing to in-flight loss-of-control aviation accidents. In spite of the tremendous progress that has been made so far, several major issues remain unresolved when analyzing the safety of air transportation system.

1. An end-to-end approach that is publicly accessible is missing in the literature. Such an approach will help to identify the dominant contributory factors (or common causes) leading to aviation incidents/accidents as well as the modeling of the relationships among the contributory factors. It is essential for researchers to compare their analysis against publicly accessible models and understand the occurrence mechanisms of aviation accidents and their escalation pathways through the system.
2. Most of the existing studies focus on a limited number of factors (e.g., human factors, structural failure, organization factors) or accident types (e.g., loss of control, see Ref. [29]). In fact, aviation is a system of systems, but the relationships across different causal factors or different accidents have been rarely explored systematically.

Under this circumstance, an end-to-end, open-source structured, systematic approach is needed to understand the role of each contributing factor in incident and accident occurrence and their escalation pathways through the system. In this paper, we are motivated to address this need by developing a Bayesian network-based approach to analyze the past aviation accidents, thereby offering insights on the mitigation of accident risk in the air transportation system. To achieve this goal, we analyzed the commercial airliner accidents (FAR 121) that happened from 1982 to 2016 as reported in the National Transportation Safety Board (NTSB) database [30]. Air carriers authorized to operate under a Federal Aviation Regulations (FAR) 121 certificate are generally large, U.S.-based airlines, regional air carriers, and all cargo operators [31].

The sequences leading to each accident event is represented as a single graph, where the causal-effect relationships are represented as edges in the graph. Next, all the single graphs representing the causal-effect relationships of each accident event are aggregated into one large Bayesian network to capture the interrelationships among different events across all the accidents. To construct the Bayesian network, we estimate the prior probabilities as the ratio of occurrences of each event in the database to the number of aircraft departures over the same period. Due to data scarcity, it is impossible to estimate all the conditional probabilities from the data. To address this issue, a nonlinear monotonically increasing function is developed to approximate the conditional probabilities, where the contribution of each individual factor specific to event outcome is considered in this function. The parameters of this function are calibrated with the available conditional probabilities when the outcome is conditional on a single event. At last, we develop a program to automate the generation of such a large scale Bayesian network in XML format. The generated Bayesian network can be fed into the GeNIe modeler from BayesFusion, LLC to support accident analysis and investigation directly [32]. The contributions made in this paper are briefly summarized as below:

1. To the best of our knowledge, this is the first time that an end-to-end approach is developed to **automatically** construct a Bayesian network from the event sequences in the NTSB aviation accident database. The Bayesian network captures the causal relationships among different events through conditional probabilities in a systematic way, thereby providing decision support for aviation accident analysis. Considering that the Bayesian network consists of more than 500 nodes and 1000 edges, it is time-consuming to manually create such a high-dimensional Bayesian network. As can be seen in CATS [3,33,34], even larger BBNs can be built by hand and adapted to new information, but doing this automatically is less time-consuming. To overcome this problem, we develop a novel procedure to automate the generation of the Bayesian network in XML format, by addressing several technical challenges as explained in Section 4.4 below.
2. Along with the collected and processed data, innovative methodologies are developed to estimate the prior probabilities and approximate conditional probabilities. The prior probabilities are estimated from the data as the ratio of event occurrences to the number of aircraft departures. A nonlinear monotonically increasing function is developed to assist the estimation of conditional probabilities, where the parameters of the function are calibrated with the available conditional probabilities derived from NTSB accident database.

The rest of the paper is structured as follows. Section 2 provides a brief introduction to the NTSB commercial airliner accident investigation data (FAR 121), and describes the major components in each accident record. Section 3 introduces the basic concepts of Bayesian networks. Section 4 develops the proposed methodology to construct the Bayesian network for the NTSB accident data. Section 5 explains the parameter calibration for the function used to approximate conditional probabilities and verifies the generated Bayesian network. Section 6 provides concluding remarks and discusses future research directions.

2. NTSB accident investigation data

The National Transportation Safety Board (NTSB) is an independent agency tasked with the mission of increasing transportation system safety by investigating every accident in civil aviation as well as in other modes of transportation (e.g., highway, railroad, marine) in the United States. In investigating each accident, NTSB determines the probable cause of the accident and issues safety recommendations with the aim of preventing future accidents. Typically, when an investigation is completed, a final description of the accident and its probable cause are made available to the public in the NTSB website. Over the past few decades, reports on civil aviation accidents and selected incidents within the United States have been stored in Microsoft Access (MDB) format to cover accident information from 1982 to present. The complete dataset for each year beginning from 1982 is available at the website https://app.ntsb.gov/avdata/.

Since we are interested in the safety of commercial flights in this paper, we select all the accidents with aircraft falling under the category of FAR Part 121. Air carriers authorized to operate under a Part 121 certificate are generally large, U.S.-based airlines, regional air carriers, and all cargo operators [31]. Under Part 121, a total number of 2243 accidents are included in the NTSB database from 1982 to 2019, among which 102 accidents resulted in fatality, 534 accidents led to serious injury, and 88 had severe damage or destruction of the aircraft.

2.1. Statistical indicators

In the NTSB aviation database, the severity of an aviation accident or incident is classified as the combination of the highest level of injury sustained by the personnel involved (that is, fatal, serious, minor, or none), which is used to indicate the highest level of injury among all injuries sustained as a result of the event, and the level of damage to the aircraft involved (that is, destroyed, substantial, minor, or none). Fig. 1 illustrates the highest level of injury and the level of damage to the aircraft for all the accidents with aircraft belonging to the FAR Part 121 as reported in the NTSB aviation database. As can be observed from Fig. 1(a), over the past two decades, the total number of accidents has dropped in a steady manner while the number of accidents with serious injuries fluctuates dramatically, possibly due to the introduction of new air traffic management policies or regulation rules as well as the advance of technologies used by aircraft and airport. Yet, the number of fatal injury is maintained at an extremely low rate since 1982. Fig. 1(b) illustrates a similar trend on the level of damage to the aircraft since 2001. The number of accidents with substantial damage to the aircraft has been reduced significantly in 2019 when compared to 2001. In the subsequent Bayesian network analysis, the outcomes of aviation accidents will be analyzed with respect to the highest level of personnel injury and the level of damage to the aircraft.

2.2. Event sequences and occurrences

The NTSB aviation database shows important information regarding the propagation of the effects of initiating events in two tables: seq_of_events and occurrences. Tables 1 and 2 illustrate the event sequences and occurrences for the accident with event ID 20001213X29335. In both tables, a unique 10- or 11-digit alphanumeric code ev_id is assigned to each accident event as ID. Another attribute in both tables is the aircraft key, and it is used in conjunction with the event ID (ev_id) to distinguish between aircraft and accident in the event of a collision between two or more aircraft. For example, a midair collision between two aircraft is recorded as a single accident with a single event ID, but with one aircraft given the aircraft key of 1 and the other given the aircraft key of 2. In NTSB accident database, each event record typically includes multiple occurrences. Each occurrence is coded sequentially in order and assigned an occurrence number signifying that order. For example, the first occurrence is coded as occurrence_no = 1. Each occurrence includes a phase_of_flight code. The meanings of the codes can be obtained from the NTSB coding manual (e.g., occurrence code, subject code, modifier code, phase of flight) [35].

In the seq_of_events Table 2, a unique numerical code is assigned to each seq_events subject code (subj_code) when multiple findings are cited for a single occurrence. For example, an event record may include occurrence_code = 1 for the first occurrence, and associated with that occurrence may be seq_event_no = 1, seq_event_no = 2, and seq_event_no = 3. If that accident included more than 1 occurrence, then occurrence_no = 2 may also have data associated with a seq_event_no = 1, seq_event_no = 2, etc. Within the list of accident findings, the column cause_factor designations are used to provide greater detail about the level of contribution each finding had to the outcome. Cause/factor designations are used in conjunction with the subject/modifier/person codes. Each mishap may have multiple causes and/or factors. For example, for the accident shown in Table 2, the improper trim setting is coded as a combination of subj_code 22120 (trim setting) and modifier_code 3109 (improper). The subject codes (Subj_Code) are used to identify the individuals, equipment, processes, or phenomena that contributed to the mishap event. Subject codes are used in conjunction with the modifier codes, person codes, and the cause/factor designator to provide the details of the mishap event and the probable cause. Person codes are used in conjunction with subject and modifier codes to indicate the individual or group associated with a finding. For example, if a pilot was delayed in aborting takeoff, it could be coded as the combination of subj_code 24505 (aborted takeoff), modifier_code 3104 (delayed), and person_code 4000 (pilot in command).

The two tables seq_of_events and occurrences keep track of the escalation of initiating events into aviation accidents in detail, where the significant contributory factors leading to the occurrence have been appropriately recorded in the two tables. In Section 4, we will demonstrate how the data present in the two tables can be leveraged to develop a Bayesian network in modeling the causal and dependent relationships between contributory factors and event outcomes.

3. Bayesian networks

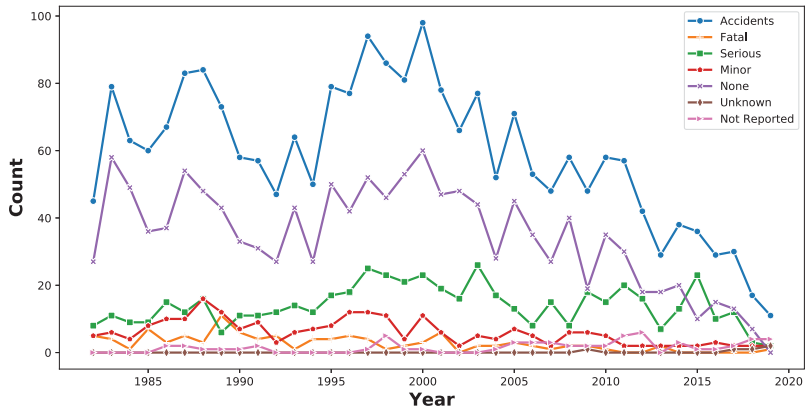
A Bayesian network is a directed acyclic graph (DAG) comprised of a set of links and probability distributions on the nodes to represent the probabilistic dependence among random variables [36–38]. In a Bayesian network, the nodes with edges directed into them are called "child" nodes, the nodes with outgoing links are called "parent" nodes, and the nodes without any parent nodes are called "root" nodes. Fig. 2 shows a four-node Bayesian network, where the random variables x_1 and x_2 are the parent nodes of x_3, i.e., node x_3 is the child node of x_1 and x_2. Further, x_3 is the parent node of x_4, or x_4 is the child node of x_3. Since x_1 and x_2 have no parent node, they are the root nodes in this Bayesian network.

The Bayesian network captures the dependence among a set of random variables through the directed edges in the network. An arc from A to B indicates that there is dependence of B on A, which is represented through the conditional probability $P(B|A)$ in probability theory. Suppose a Bayesian network consists of n random variables $x_1, x_2, \ldots, x_n$, we can write the full joint probability distribution as:

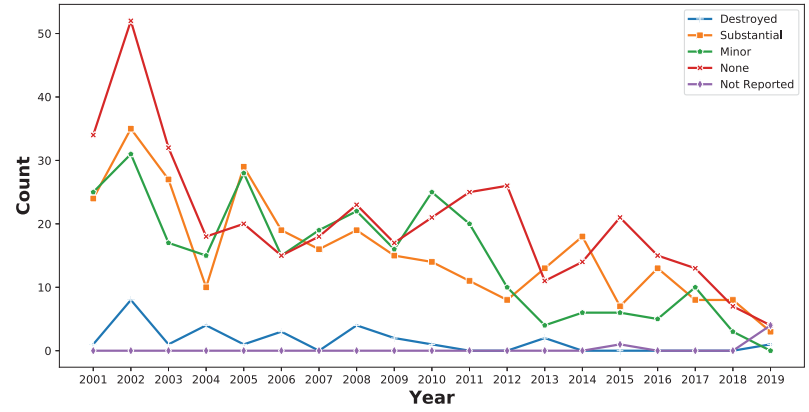
$$P(x_1, x_2, x_3, \ldots, x_n) = P(x_1|x_2, x_3, \ldots, x_n)\, P(x_2|x_3, x_4, \ldots, x_n) \cdots P(x_{n-1}|x_n)\, P(x_n) \quad (1)$$

and it can be reformulated as

$$P(x_1, x_2, x_3, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|x_{i+1}, \ldots, x_n) \quad (2)$$



(a) Variation of the highest level of injury 1982-2019



(b) Level of aircraft damage since 2001 (the data on aircraft damage before 2000 is not available in the NTSB aviation database)

Fig. 1. Damage severity as a result of aviation accident from the NTSB database.

Table 1
Preliminary data on accident event 20001213X29335.

ev_id	Aircraft_Key	Occurrence_No	Occurrence_Code	Phase_of_Flight	Altitude
20001213X29335	1	1	260	521	0
20001213X29335	1	2	340	523	0
20001213X29335	1	3	310	523	0

Table 2
Event sequences in accident record 20001213X29335.

ev_id	Aircraft_Key	Occurrence_No	seq_event_no	Subj_Code	Cause_Factor	Modifier_Code	Person_Code
20001213X29335	1	1	1	22120	C	3109	4000.0
20001213X29335	1	2	1	24505	C	3104	4000.0
20001213X29335	1	3	1	20200		2506	0

Suppose Parents (x_i) denotes the set of parent nodes of node x_i, then we can simplify the joint probability distribution shown in Eq. (2) by using our knowledge of what the parents of each node are:

$$P(x_1, x_2, x_3, \ldots, x_n) = \prod_{i=1}^{n} P\left(x_i \middle| \text{Parents}(x_i)\right) \tag{3}$$

Given Eq. (3), the full joint probability distribution of the Bayesian network shown in Fig. 2 can be rewritten as

$$P(x_1, x_2, x_3, x_4) = P\left(x_4 \middle| x_3\right) P\left(x_3 \middle| x_1, x_2\right) P(x_1) P(x_2) \tag{4}$$

The calculation of the full joint probability $P(x_1, x_2, x_3, \ldots, x_n)$ requires knowledge of the marginal probabilities and conditional probabilities, i.e., $P(x_n)$, $P\left(x_{n-1} \middle| x_n\right)$ etc. For the Bayesian network shown in Fig. 2, the marginal distributions of root nodes can be inferred from historical data. With respect to the event *electrical system wiring overheating*, we can count how many times overheating occurred in the wiring out of the total number of flights. For the sake of demonstration, assume that random variables x_1 and x_2 have the marginal distributions shown in Table 3.

Another important quantity is the conditional probability table (CPT) for the child nodes. The CPT measures the probability of each value of one variable if we know the values taken by the other variables. Table 4 shows the CPT of the random variable x_3, we observe

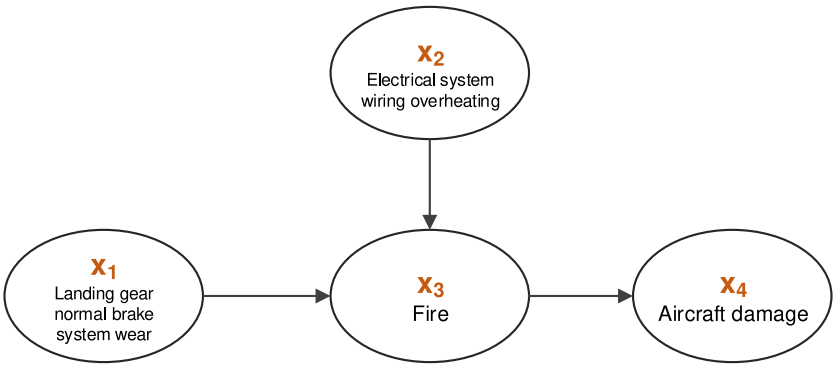


Fig. 2. A simple Bayesian network illustrating the occurrence of an aviation accident.

Table 3
Marginal distributions of random variables x_1 and x_2.

	Landing gear normal brake system wear		Electrical system wiring overheating
$P(x_1 = 1)$	0.0001	$P(x_2 = 1)$	0.0002
$P(x_1 = 0)$	0.9999	$P(x_2 = 0)$	0.9998

1: Yes, 0: No.

Table 4
Conditional probability table (CPT) of random variable x_3 shown in Fig. 2.

Landing gear normal brake system wear	Yes		No	
Electrical system wiring overheating	Yes	No	Yes	No
$P(x_3 = 1)$	0.99	0.93	0.95	2e−9
$P(x_1 = 0)$	0.01	0.07	0.05	1-(2e−9)

Table 5
Conditional probability table (CPT) of random variable x_4.

Fire	Yes	No
$P(x_4 = 1)$	0.92	0
$P(x_1 = 0)$	0.08	1

that if the landing gear normal brake system is worn out and the electric wiring is overheated, then there is a probability of 0.99 for the fire occurrence. Similarly, Table 5 illustrates the CPT for the random variable x_4. When a fire happens, there is a 92% chance that the aircraft will be damaged.

In the Bayesian network, if any additional evidence d is available on some random variable x_i, then the occurrence probability of event x_i can be updated following Bayes' theorem as shown in Eq. (5) [39]. The updated occurrence probability $P(x_i|d)$ can then be propagated through the entire Bayesian network both forward and backward to update our belief about the occurrence probabilities of other events relating to the random variable x_i across the Bayesian network.

$$P(x_i|d) = \frac{P(d|x_i) \times P(x_i)}{P(d)} \tag{5}$$

Take the Bayesian network in Fig. 2 as an example, Fig. 3 demonstrates the updating of our belief in a few cases when evidence is observed. In Fig. 3(b), we observe that if the landing gear normal brake system is worn out, the fire occurrence probability increases to 93%. As a result of the increased probability of fire occurrence, the probability that the aircraft will be damaged increases to 86% accordingly. Fig. 3(c) shows that if we observe that the fire happened, the occurrence probability of root causes are updated accordingly. There is a 67% probability that the wiring is overheated, and 33% probability that the landing gear is worn out. As a result, the probability of aircraft damage increases to 92%.

In a Bayesian network, the uncertainty caused by limited data or lack of knowledge is propagated in a quantitative manner. The solid mathematical foundation for reasoning under uncertainty has made the Bayesian network emerge as a promising information fusion tool in support of effective diagnosis, prediction, risk analysis, and decision making in the face of limited information [40–44].

4. Proposed methodology

In this section, we introduce the proposed methodology for the development of a Bayesian network-based representation to capture the causal relationships embedded in the historical NTSB aviation accidents. The proposed methodology follows a four-step procedure. In the first step, for each accident, we develop a graphical representation of the progression from initiating events to an accident. In the second step, we develop an approach to estimate the prior probabilities of each event from the historical data. In the third step, since the accident data is limited, we develop an approach to approximate the conditional probabilities if the event is conditional on multiple contributory causes. In the fourth step, we generate an aggregated Bayesian network that combines information on all the accidents. The Bayesian network is generated in the XML format compatible with GeNIe modeler in an automated manner. The generated aggregated Bayesian network can then be fed into GeNIe modeler to support accident analysis and investigation directly.

4.1. Graphical representation of event sequence in each accident

Section 2.2 explains the significant contributory factors that occur sequentially the occurrence of aviation accident, including human factors, weather conditions, airport conditions, communications etc. In this section, we combine the event sequences in the NTSB accident database with the coding manual (e.g., occurrence code, subject code, modifier code) to retrieve information concerning the findings of aviation accidents. By doing this, we create an end-to-end graphical representation for each aviation accident. The graphical representation covers how the event happened, and what is the outcome of

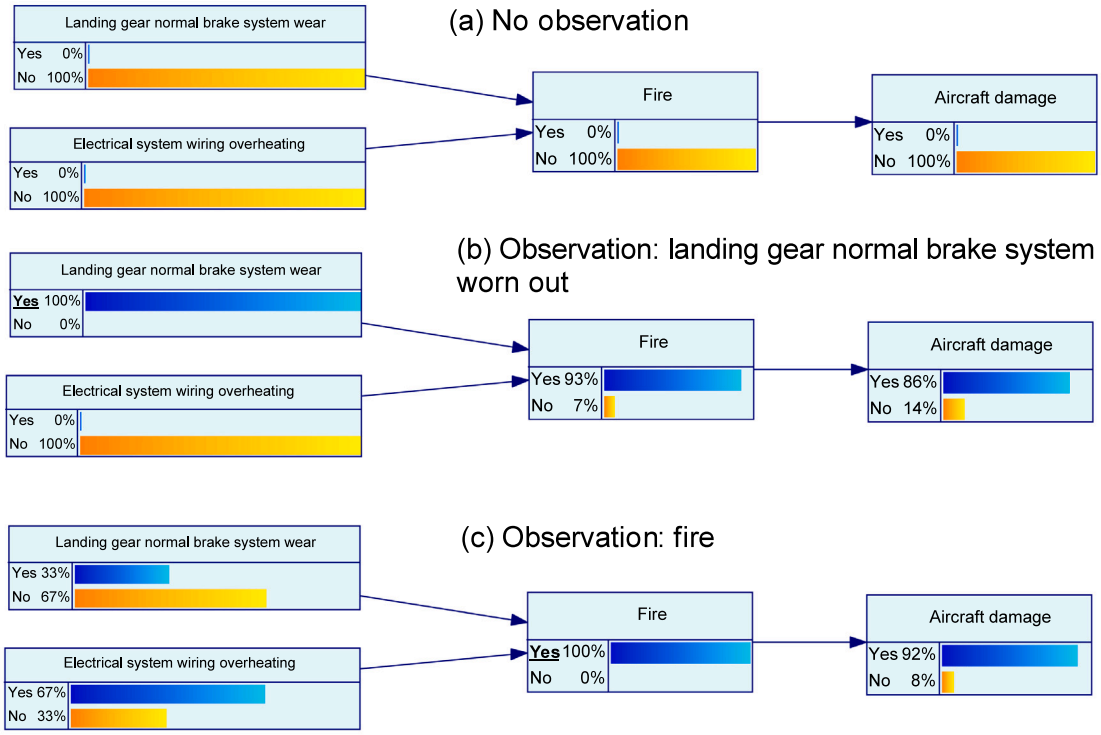


Fig. 3. Bayesian network inference demonstration.

the accident (i.e., aircraft damage level, personnel injury level etc.) following the order of event occurrences, where the causes and factors contributing to each occurrence are appropriately represented in the graph. In this example, the pilot errors are represented as parent nodes for each occurrence. By combining the event information with the aircraft damage and personnel injury table, we derive the outcome of the accident as represented by the degree of damage to aircraft and the level of injury to personnel in the flight.

Fig. 4 illustrates the graphical representation of the sample accident event 20001213X29335, where the three green circles denote the three occurrences as shown in Table 1, and the two gray cells denote the outcome of the accident event in terms of aircraft damage and person injury. From this figure, we can see that the pilot-in-command configured the trim improperly, which caused the loss of control of the aircraft on the ground. Next, the pilot delayed aborting the takeoff, which led to overrun of the aircraft beyond the runway. The overrunning aircraft collided with the approach light and navigational aid (NAVAID) on the ground. As a result, the aircraft was destroyed and several fatal injuries occurred.

The graphical representation of accident event sequence greatly facilitates our understanding on the escalation of initiating events in the aviation system. The visualization also helps with the explanation of the accident occurrence.

4.2. Prior probability estimation

The previous section dealt with the graphical representation of a single accident event. In this section, we focus on the development of an aggregated Bayesian network for all the accidents in the NTSB FAR Part 121 accident investigation database. Each aviation accident in the database corresponds to a unique graphical representation similar to Fig. 4. In the aggregated Bayesian network, the prior probabilities and the conditional probability tables need to be properly estimated before it can be used for inference. Typically, the prior probability can be estimated as the ratio of the occurrence times of a given event to the total number of flights over the same period. The total number of flights within the U.S. can be calculated from the data available in the Bureau of Transportation Statistics website [45]. Table 6 reports the number of aircraft departures for all the scheduled and nonscheduled service by large certified U.S. air carriers at all airports served within the 50 states and the District of Columbia from 1975 to 2018. As shown in Table 6, not all the scheduled services are actually performed. Moreover, for several years, the number of total performed departures exceeds the number of total scheduled departures because nonscheduled departures are included in the totals. Considering that the accidents in the NTSB database we analyzed in this paper occurred between 1982 to 2006, the sum of total performed aircraft departures within the date ranging from 1982 to 2016 gives the total number of flights, and it acts as the denominator when calculating the prior probability for each event.

As illustrated in Table 6, the number of aircraft departures over some years are missing, such as 1981–1984, 1985–1989. To address this issue, we impute the number of aircraft departures over the missing years with linear interpolation. Fig. 5 illustrates the number of total performed flights from 1982 to 2018. With the interpolation function, the sum of total performed flights from 1982 to 2006 is estimated as 184,517,128. With the denominator, we then count the occurrence times of each accident between 1982 and 2016, and then compute the prior probability following Eq. (6).

$$P(e) = \frac{T(e)}{T_{sf}} \tag{6}$$

where T_{sf} and $T(e)$ denote the number of total performed flights and the times that event e occurred during the period, respectively.

For example, to compute the prior probability of fire, we count how many times fire has occurred between 1982 and 2006 in the NTSB aviation data. Based on the calculation, we have $T(e = \text{fire}) = 102$, then we have the prior probability of fire occurrence as $P(e = \text{fire}) = \frac{102}{184517128} \approx 5.52 * 10^{-7}$. Following the same procedures, the prior probabilities of other aviation events can be estimated accordingly. As can be seen, the prior probabilities are extremely small, which could result in computational errors. However, once the target node

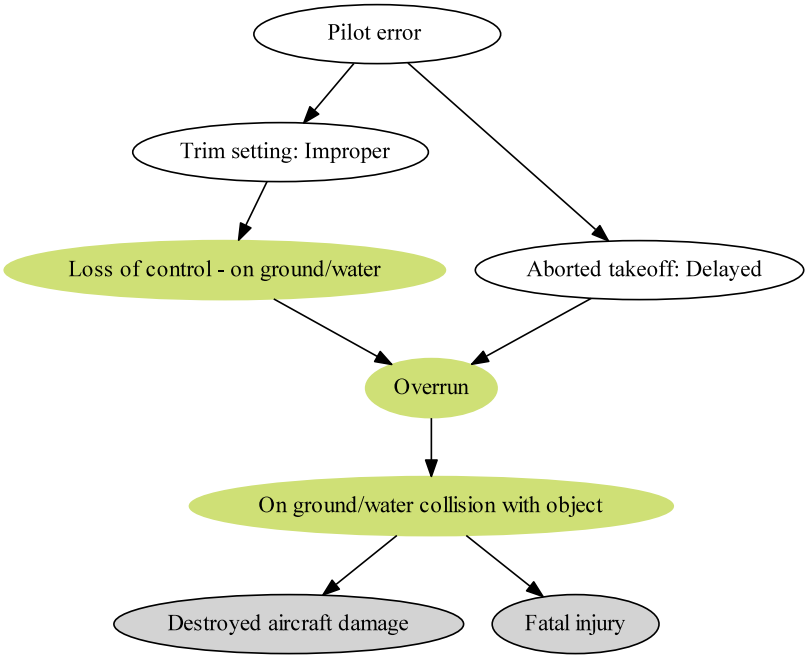


Fig. 4. Graphical representation of the sequences for the sample accident event 20001213X29335.

Table 6
U.S. Air Carrier Aircraft Departures from 1975 to 2018 [45].

Year	Total performed	Total scheduled	Year	Total performed	Total scheduled
1975	4,555,516	4,530,535	2003	8,585,736	8,479,414
1980	5,156,848	5,204,564	2004	9,444,234	9,193,220
1985	5,505,659	5,591,596	2005	9,859,941	9,722,715
1990	6,641,681	6,758,571	2006	9,512,017	9,429,017
1991	6,545,000	7,024,412	2007	10,985,488	10,669,356
1992	6,606,609	6,703,670	2008	10,307,025	9,975,967
1993	7,193,841	7,058,097	2009	9,646,132	9,324,192
1994	7,513,232	7,359,093	2010	9,596,396	9,241,790
1995	8,030,530	7,920,467	2011	9,577,700	9,160,580
1996	8,204,674	8,064,653	2012	9,345,013	8,836,158
1997	8,095,888	7,907,554	2013	9,217,652	8,765,026
1998	8,248,269	8,094,020	2014	8,986,825	8,631,508
1999	8,605,486	8,432,940	2015	8,965,387	8,622,198
2000	8,929,559	8,688,776	2016	9,276,526	8,976,677
2001	8,548,932	8,340,180	2017	9,305,375	9,019,930
2002	8,052,756	7,981,190	2018	9,559,858	9,214,306

* Total performed includes scheduled departures performed minus those scheduled departures that did not occur plus unscheduled service.

is specified in a Bayesian network (i.e., specify parent node in forward propagation and specify child node in backward inference), we propagate the corresponding information through only a few nodes, thus significantly reducing the error contribution. Such an approach to estimate prior probabilities ignores the cases where abnormal events occurred but did not lead to an accident or incident, since such events are not reported in NTSB aviation accident database.

4.3. Conditional probability estimation

Besides prior probabilities, another major element in a Bayesian network is the conditional probability, which measures the probability of an event occurring given that another event has occurred, defined as.

$$P\left(\omega|\, e_1, e_2, \ldots, e_m\right) = \frac{P\left(\omega, e_1, e_2, \ldots, e_m\right)}{P\left(e_1, e_2, \ldots, e_m\right)} \tag{7}$$

where $P\left(\omega|\, e_1, e_2, \ldots, e_m\right)$ is the conditional probability of event ω to happen given the occurrence of events $e_1, e_2, \ldots, e_m$, $P\left(\omega, e_1, e_2, \ldots, e_m\right)$ denotes the joint probability of events ω and $e_1, e_2, \ldots, e_m$ to occur, while $P\left(e_1, e_2, \ldots, e_m\right)$ represents the probability of events $e_1, e_2, \ldots, e_m$ occurring together.

A common challenge in the Bayesian network is how to build the conditional probability table (CPT) from the available data. Since the number of accidents reported in NTSB is limited, the data will not cover all the possible scenarios and possibilities for different types of events and sequences, thereby making the derivation of conditional probability by only relying on the data impossible. Under this circumstance, one issue is how to estimate the conditional probability in the presence of data scarcity. Considering the example illustrated in Fig. 2, both the overheating of electric wiring and landing gear brake system wear-out can lead to fire in the airplane. In the NTSB aviation database, we might only observe one of these causal events (i.e., either electrical system electric wiring overheating causes fire or landing gear brake system worn causes fire), but not both of them occurring together. However, it is possible that both events can happen together during a flight, and in that case, the probability of fire occurrence will increase accordingly. No record about the joint occurrence of overheating of electric wiring and landing gear brake system wear-out in the NTSB accident database does not imply that they cannot happen together in reality, and does

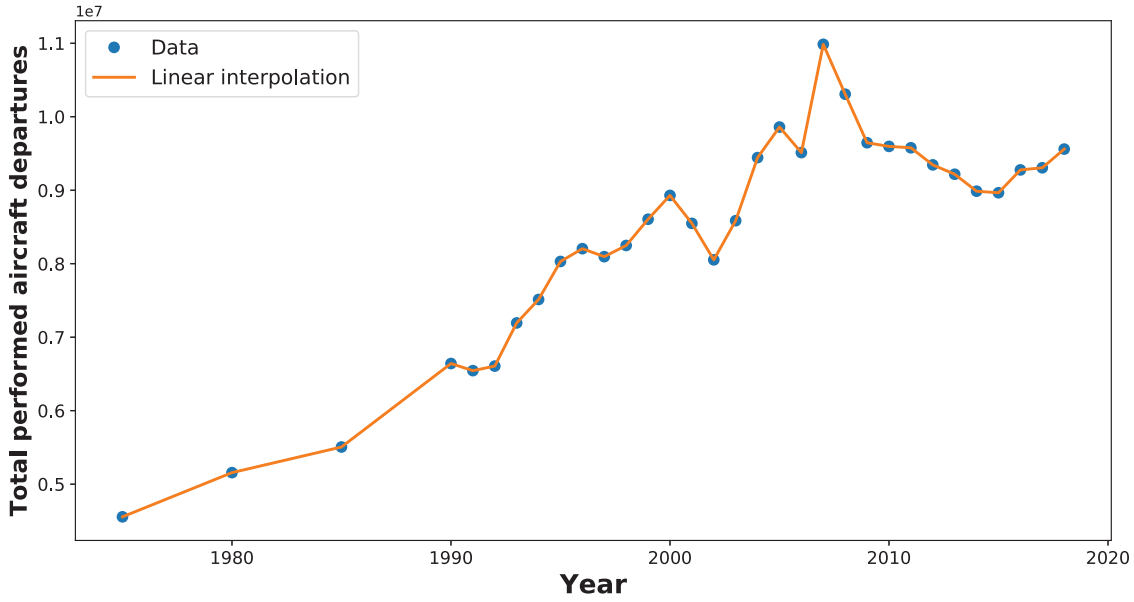


Fig. 5. Linear interpolation of aircraft departures within U.S.

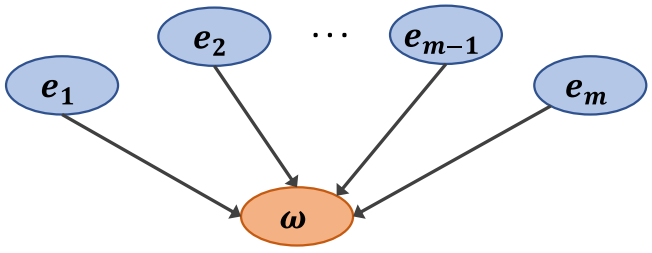


Fig. 6. V-structure common effect in Bayesian network, where events $e_1, e_2, \dots, e_m$ result in the same outcome ω.

not imply that the conditional probability of fire given the occurrence of overheating of electric wiring and landing gear wear-out is zero.

To address this issue, we develop a simple method to estimate the conditional probability. Suppose events $e_1, e_2, \dots, e_m$ denote a set of *collectively exhaustive* causes contributing to the event ω in the NTSB aviation database, Fig. 6 illustrates a v-structure network-based representation for the relationship between $e_1, e_2, \dots, e_m$ and ω. The contribution of each individual factor e_i to the occurrence of ω can be estimated as:

$$P\left(\omega|\, e_i\right) = \frac{P\left(\omega, e_i\right)}{P\left(e_i\right)} \tag{8}$$

where $P\left(\omega, e_i\right)$ denotes the joint probability of events ω and e_i to occur together, and $P\left(e_i\right)$ indicates the probability of e_i to happen. Both quantities $P\left(\omega, e_i\right)$ and $P\left(e_i\right)$ can be estimated from the NTSB accident database as described in the previous subsection.

With Eq. (8), we quantify the contribution of individual factors to the occurrence of event ω. Consider the occurrence of fire in an airplane as an example; this event can be caused by a wide range of factors, such as fuel control leakage, electrical system wiring overheating, airframe component malfunction etc. Following Eq. (8), we count how many times the aforementioned events and fire occur together in the database. By combining it with the total number of fire occurrences between 1982 and 2006 (which is 102 times according to the NTSB database), we have:

$$\begin{aligned}
&P\left(\omega = \text{fire}|\, e = \text{fuel system fuel control leak}\right) = \tfrac{1}{102} = 9.80 * 10^{-3},\\
&P\left(\omega = \text{fire}|\, e = \text{electrical system electric wiring overheating}\right)\\
&\quad = \tfrac{1}{102} = 9.80 * 10^{-3},\\
&P\left(\omega = \text{fire}|\, e = \text{airframe component malfunction}\right) = \tfrac{32}{102} = 31.37 * 10^{-2}.
\end{aligned} \tag{9}$$

Likewise, we follow the same procedures to estimate the conditional probability of other effect variables if they are conditional on a single cause. In the case that ω is conditional on multiple events, since events $e_1, e_2, \dots, e_m$ are independent from each other, each event e_i contributes to the occurrence of ω to a different degree, and they are not mutually exclusive. Under such circumstance, the more events like e_i to occur, the more likely for event ω to happen. Hence, the lower bound of $P\left(\omega|\, e_1, e_2, \dots, e_m\right)$ can be approximated as:

$$P\left(\omega|\, e_1, e_2, \dots, e_m\right) \geq P\left(\omega|\, e_i\right), \quad \forall\, i = 1, 2, \dots, m. \tag{10}$$

and Eq. (10) can be reformulated as:

$$P\left(\omega|\, e_1, e_2, \dots, e_m\right) \geq \max_{i=1,2,\dots,m} P\left(\omega|\, e_i\right) \tag{11}$$

In a similar way, we have:

$$P\left(\omega|\, e_1, e_2, \dots, e_{m-1}\right) \leq P\left(\omega|\, e_1, e_2, \dots, e_m\right) \tag{12}$$

Since we do not know the exact value of $P\left(\omega|\, e_1, e_2, \dots, e_m\right)$, we approximate this quantity considering the worst scenario. In the worst case, when all the events $e_1, e_2, \dots, e_m$ happen, the following assumption can be made to estimate the conditional probability:

$$P\left(\omega|\, e_1, e_2, \dots, e_m\right) = 1 \tag{13}$$

Since we only have the conditional probabilities when ω is conditioned on single events as shown in Eq. (9), a function f is needed to support the estimation of the conditional probability when ω is conditioned on multiple events, where the following considerations should be included in developing the function f:

1. **Monotonicity**. f should be a monotonically increasing function when the number of events that ω is conditional on increases; this consideration corresponds to the constraint formulated in Eq. (12);

2. **Incorporation of single event contribution to the occurrence of** ω. The contribution of each single event e_i to the occurrence of ω should be incorporated when developing the function f; in other words, $P(\omega|e_i)$ needs to be included in the function f;
3. **Generalization**. The function f should be generalizable to several special cases. When conditional events are $e_1, e_2, \ldots, e_m$, then $P(\omega|e_1, e_2, \ldots, e_m) = 1$; similarly, when there is no conditional event, then the conditional probability should be approximately zero; when ω is conditional on a single event e_i, then the value of the function f should be equal to $P(\omega|e_i)$;

Any function satisfying the above constraints can be utilized to estimate the conditional probabilities. One promising function is beta distribution, which is shaped by four parameters, it is very versatile and a variety of uncertainties can be modeled with beta distribution. Due to its flexibility, beta distribution has been extensively used in a wide range of applications [46], such as fitting the dose–response model [47], modeling oil conversion rate after distillation and fractionation [48]. In this paper, we use the cumulative distribution function (CDF) of beta distribution as a nonlinear function, which is shown in Eq. (14b), to help estimate the conditional probabilities when ω is conditional on multiple events. In Section 5.1, we will show the performance of the CDF of beta distribution in fitting the conditional probabilities of single events as estimated from the data.

$$g(\lambda) = \frac{\lambda^{\alpha-1}(1-\lambda)^{\beta-1}}{B(\alpha,\beta)},\ 0 \le \lambda \le 1. \tag{14a}$$

$$f(\lambda) = \frac{\int_0^{\lambda} \lambda^{\alpha-1}(1-\lambda)^{\beta-1}dt}{B(\alpha,\beta)},\ 0 \le \lambda \le 1; \alpha,\beta > 0 \tag{14b}$$

where Eq. (14a) shows the probability distribution density (PDF) of beta distribution with two shape parameters α and β, $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, and Γ is the Gamma function; $f(\lambda)$ is the cumulative distribution function (CDF) of beta distribution as shown in Eq. (14b), λ is a factor within the range $[0, 1]$ that is used to measure the contribution from all the events that ω is conditional on, and the two unknown shape parameters α and β in beta distribution can be calibrated from the available data. The function f has the following attractive properties:

1. When $\lambda = 0$, $f(\lambda) = 0$. Therefore, it can be used to model the scenario that the conditional probability is zero when ω is not conditional on any event.
2. When $\lambda = 1$, $f(\lambda) = 1$. Hence, when ω is conditional on all the contributory factors $e_1, e_2, \ldots, e_m$, we have $f(1) = 1$.
3. For any two values, λ_1 and λ_2, if $\lambda_1 \ge \lambda_2$, we always have $f(\lambda_1) \ge f(\lambda_2)$. As can be seen, function f is a monotonically increasing function: with the increase of λ, $f(\lambda)$ increases accordingly. Hence, we have $f(\lambda_1) \ge f(\lambda_2)$ when $\lambda_1 \ge \lambda_2$.

The next issue is how to define λ. Since we have the contribution of each cause e_i to the occurrence of ω as $P(\omega|e_i)$, we define λ as below to relate it with the contribution of each event e_i.

$$\lambda = \frac{\sum_{i=1}^{n} P(\omega|e_i)}{\sum_{i=1}^{m} P(\omega|e_i)} \tag{15}$$

The combination of λ and the beta CDF facilitates the estimation of probabilities when events ω is conditional on multiple events. More importantly, the quantity $P(\omega|e_i), (i = 1, \ldots, m)$ can be estimated from the data in a straightforward manner. In consideration of the definition of λ, we can observe that the more events e_i occur in the numerator of Eq. (15), the higher the λ value. The higher λ value eventually results in a higher conditional probability. Such observation complies with common sense as indicated in Eq. (12).

After quantifying the contribution from the subset of causes $e_1, e_2, \ldots, e_n$ to ω, we combine Eq. (15) with Eq. (14b), then the conditional probability $P(\omega|e_1, e_2, \ldots, e_n)$ can be estimated as:

$$P(\omega|e_1, e_2, \ldots, e_n) = \frac{\int_0^{\frac{\sum_{i=1}^{n} f(e_i)}{\sum_{i=1}^{m} f(e_i)}} \left(\frac{\sum_{i=1}^{n} f(e_i)}{\sum_{i=1}^{m} f(e_i)}\right)^{\alpha-1} \left(1 - \frac{\sum_{i=1}^{n} f(e_i)}{\sum_{i=1}^{m} f(e_i)}\right)^{\beta-1} dt}{B(\alpha,\beta)} \tag{16}$$

As shown in Eq. (16), there are two unknown parameters α and β. Since we have the conditional probability $P(\omega|e_i)$ that measures the contribution of individual factor e_i to the occurrence of ω. Thus, they can be used to calibrate the parameters of a and b. By calibrating the values of α and β, we guarantee that the function f is generalizable. The generalizability and the monotonicity features ensure that the constraints formulated in Eqs. (11) and (12) are satisfied.

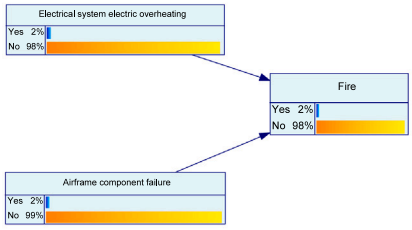
The intuition of approximating the joint probability through the above method is that the contribution of each individual factor e_i to the outcome ω is characterized separately. The effects of parent nodes on the child are essentially independent. In the domain of accidents, the more events (e.g., $e_1, e_2, \ldots, e_i$) occur, the more likely it is for ω to occur. Considering the independence of these events, the occurrence probability of ω increases accordingly when more and more events that ω is conditional on occur. In particular, if all the contributory factors happen, then the event ω will happen with 100% probability. By doing this, the Bayesian network provides a way to perform inference and propagation in the domain of accidents.

4.4. Bayesian network construction

The previous three subsections lay the theoretical foundation for the development of the Bayesian network to quantitatively capture the causal relationships embedded in the NTSB aviation accident database. In this section, we focus on how to build such a large-scale Bayesian network in an effective manner. Consider a Bayesian network with a Boolean node ω that has m Boolean parents $e_1, e_2, \ldots, e_m$. In general, the conditional probability table for ω and $e_1, e_2, \ldots, e_m$ requires us to determine 2^{m+1} parameters, corresponding to the probabilities that each of the parent state combination is true. When $m = 10$, 2048 entries need to be filled in the conditional probability table. Obviously, it is a tedious task to manually input so many values for all the entries in the conditional probability table through a commercial software for Bayesian network analysis. In fact, there are more than 500 nodes and 1000 edges in the Bayesian network for the NTSB accident database, it is thus essential to develop a procedure to automate the Bayesian network construction.

To address this challenge, we utilize the syntax of the GeNIe modeler to represent a Bayesian network. GeNIe modeler is a graphical user interface (GUI) to the Structural Modeling, Inference, and Learning Engine (SMILE) from BayesFusion, LLC. It allows for interactive model building and learning, where SMILE is a reasoning and learning/causal discovery engine for graphical models, such as Bayesian networks, influence diagrams, and structural equation models [49]. When the Bayesian network is at a small scale (e.g., 2–20 nodes), we can manually create the network in the GeNIe modeler as shown in Fig. 7(a). The GeNIe modeler models the created Bayesian network in Fig. 7(a) as a XML file as shown in Fig. 7(b).

Fig. 7(a) illustrates a very simple Bayesian network comprised of three nodes with each node having two states: Yes or No. Since the node *fire* has two parent nodes, its conditional probability table has $2^3 = 8$ entries. Fig. 7(b) demonstrates the corresponding representation with the extensible markup language (XML) for the same Bayesian network. As can be observed, the XML representation has two principal tags: nodes and extensions. In the tag nodes, the prior probability of each node is reported consisting of node ID, the states of each node, and the probability of each state. For example, the node with the ID Node3 has two states: Yes and No, and the prior probabilities for the two states are reported following the order of the states, which are



(a) An illustrative Bayesian network

(b) XML representation of the Bayesian network shown in Fig. 7(a)

Fig. 7. Bayesian network visualization and the corresponding XML representation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

0.02 and 0.98, respectively. Another major component in the tag nodes is the conditional probability table. As shown in the second yellow box, the conditional probability table has one additional attribute — parent nodes. The parent nodes defines the connections among the nodes. For example, it can be observed that Node5 is the child node of Node3 and Node6. Right below the attribute parents, the conditional probability for Node5 is reported following the column-major order. The second tag extensions primarily define the additional decorative attributes pertaining to each node, e.g., node name, node color, node position etc. For example, the name of node with ID node3 is *Electrical system wiring overheating*. The extension tag define the layout of each node in the figure, and facilitates the visual inspection of the causal relationships in the network quickly.

The two objects shown in Fig. 7 are interchangeable in the GeNIe modeler. The file organized in XML format shown in Fig. 7(b) can be directly fed into the GeNIe modeler for visualization and decision analytics, while the basic information of node and edges in the Bayesian network shown in Fig. 7(a) is represented in the XML file properly (e.g., the position and color of each node, the parent nodes of each child node, and the marginal & conditional probabilities of each node). Since the two objects are interchangeable, we develop an approach to generate the large-scale Bayesian network for the NTSB aviation accidents as an XML file in an automatic manner. The Bayesian network will characterize the causal relationships in NTSB aviation accidents in a form that is in compliance with the syntax used in the GeNIe modeler. Two challenges need to be addressed in the automatic generation of the Bayesian network for NTSB accidents:

1. Pruning parent nodes in the conditional probability table. In the NTSB accident database, it is common that many child nodes have more than 20 parent nodes. To keep the network at a manageable size, we prune the network by keeping the top 12 parent nodes that have the highest contribution to the occurrence of ω in the conditional probability table. In other words, we rank all the nodes based on their conditional probabilities $P(\omega|e)$, and only generate the conditional probabilities for the retained 12 parent nodes. Keeping 12 parent nodes in the Bayesian network is already time-consuming to perform analysis with the GeNIe modeler. If we further increase the number of parent nodes, the Bayesian inference in such a large scale might become computationally unaffordable.
2. Order of nodes. As shown in Fig. 7(b), the parent nodes corresponding to each child node must appear before the child node itself. In other words, prior to building the conditional probability table for the child node, we need to check whether all of its parent nodes are already present in the XML file. All the nodes must be generated in this order such that it can be understood by the GeNIe modeler.

When the generated XML file complies with the syntax used in GeNIe modeler, it can then be directly fed into GeNIe modeler for analytics. With respect to the NTSB aviation accident database, the prior probability and the conditional probability can be derived following the equations developed in the previous two sections. In the Bayesian network, the marginal and conditional probabilities are assigned to corresponding nodes and edges. Once the network is loaded into GeNIe modeler, we can manually adjust the position of each node in the network for the purpose of better visualization, thereby supporting accident analysis. The developed program allows us to generate the Bayesian network in a timely and flexible manner when the accident data changes over time.

4.5. Summary

In this section, an end-to-end Bayesian network construction methodology is developed to quantitatively represent the event sequences embedded in the NTSB aviation database. The proposed methodology follows a four-step procedure:

1. For each individual accident, an end-to-end graphical representation is developed to model the escalation pathways of initiating events into accidents in the system, and the graphical visualization greatly facilitates the understanding of all the contributory factors involved in the occurrence of an accident.
2. Using the historical U.S. air carrier aircraft departure data from the Bureau of Transportation Statistics, we estimate the prior probabilities of individual events as the ratio of the number of occurrences of accidents to the number of aircraft departures (total performed) over the time period.

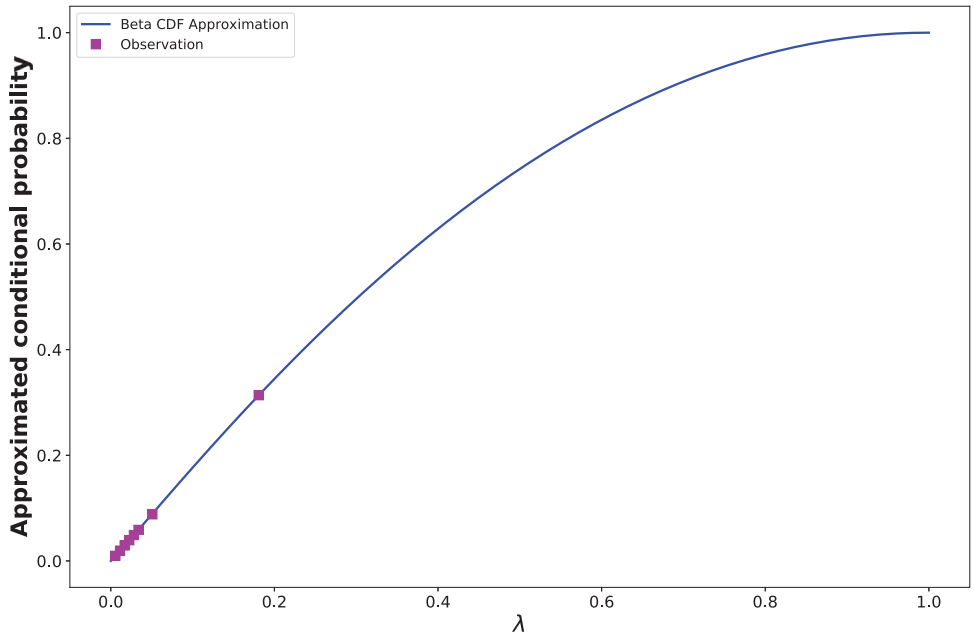


Fig. 8. Conditional probability function estimated from the observed data.

3. A function is proposed to approximate the conditional probabilities by incorporating several important considerations: monotonicity, individual factor contribution, and generalization. The parameters in this function can be calibrated with the conditional probabilities $P(\omega|e_i)$ measuring individual factor e_i to the occurrence of ω.
4. A computer program is developed to automate the generation of Bayesian network following the structure derived from the NTSB database in XML format. The generated Bayesian network can be directly fed into commercial software GeNIe modeler for forward propagation and backward inference.

5. Computational results

In this section, we first illustrate how the unknown parameters a and b in the function f can be calibrated with the available data, then we validate the Bayesian network constructed in the previous section by performing sensitivity analysis to analyze the effect of varying the probability of parent nodes on child nodes. Finally, the validated Bayesian network is used for a case study, and we demonstrate how the developed Bayesian network can be used for forward propagation and backward inference.

5.1. Parameter calibration

Take the fire occurrence as an example; in the NTSB aviation database, fire can be caused by a wide variety of factors ranging from fuel system malfunction to procedures/directives not followed. The specific information of the contributory factors to fire occurrence and their corresponding conditional probabilities are shown in Table 7. As shown in Table 7, all the cells in blue color represent a set of records with unique conditional probabilities. These records can then be used for estimating unknown parameters a and b. By summing the contributions of all the factors as shown in Table 7 together, we have the contribution of all the factors to fire occurrence as 1.735, then the λ corresponding to each case can be derived following Eq. (15) in a straightforward manner. For example, the λ corresponding to fuel system, drain is $\frac{0.0196}{1.735} = 0.0113$. By substituting the values of λ into Eq. (14b), we can formulate the calibration of parameters α and β as an optimization problem. Let the actual conditional probability denoted by y_i for the record i, the value of λ corresponding to record i is represented as λ_i, then we can minimize the following function to optimize the parameters of α and β for the developed nonlinear function to approximate conditional probabilities provided that we observe the conditional probabilities for k records:

$$\min \quad h = \sum_{i=1}^{k} \left[f(\lambda_i) - y_i \right]^2 = \sum_{i=1}^{k} \left[\frac{\int_0^{\lambda_i} {\lambda_i}^{\alpha-1} (1-\lambda_i)^{\beta-1} dt}{B(\alpha,\beta)} - y_i \right]^2 \tag{17}$$

Afterwards, we can use the conditional probabilities in the observed data to optimize the two shape parameters α and β. Thus, we use the unique conditional probabilities (0.00980, 0.01960, 0.02941, 0.03921, 0.04902, 0.05882, 0.08823, 0.31372) as shown in Table 7 (see the blue shaded cells in Table 7) to optimize the values of α and β. Using the Nelder–Mead optimization algorithm [50], we derive the values of α and β as 1.04645 and 2.02591, respectively. The mean squared error between the predictions of the fitted beta CDF and the actual values for fire occurrence is $3.42608 * 10^{-7}$, which indicates that the beta CDF fits the observed conditional probabilities very well. Fig. 8 shows the performance of beta CDF to approximate the conditional probabilities estimated from the observed data. As can be observed, the function fits all the data points perfectly, and the fitted function can then be used to approximate the conditional probabilities if ω is conditional on multiple events because its λ is readily available following Eq. (15). Likewise, we build predictive models for other variables following the same procedure, the average mean squared error between the predictions of fitted models and the actual probabilities for all the variables is $1.048 * 10^{-4}$, and the average standard deviation is $3.6663 * 10^{-4}$. The low prediction errors reveal that the developed method shows generalized performance when approximating the conditional probabilities of a large number of variables.

5.2. Sensitivity analysis

In this section, we illustrate the developed Bayesian network, and examine its performance in accident investigation and analysis. With

Table 7
The contributory factors to fire occurrence and the corresponding conditional probabilities.

Cause	Conditional probability	Cause	Conditional probability
Fuel system, nozzle	0.01960	Exhaust system, stack	0.00980
Fuel system, drain	0.01960	Fuel system, fuel control	0.01960
Maintenance, service of aircraft/equipment	0.01960	Miscellaneous, bolt/nut/fastener/clamp/spring	0.00980
Electrical system, fuse	0.00980	Evacuation	0.00980
Ignition system, exciter	0.00980	Aircraft/equipment, inadequate design	0.00980
Electrical system, circuit breaker	0.01960	Maintenance, service bulletin/letter	0.02941
Aircraft/equipment inadequate, aircraft component	0.00980	Engine accessories, engine starter	0.01960
Condition(s)/step(s) insufficiently defined	0.00980	Panic	0.00980
Unknown quantity	0.01960	Engine compartment	0.02941
Passenger compartment light(s)	0.00980	Aircraft/equipment, inadequate standard/requirement	0.00980
Fluid, hydraulic	0.00980	Overrun	0.00980
Anti-ice/deice system, windshield	0.00980	Inadequate substantiation process, Insufficient review	0.00980
Loss of engine power (total) - mechanical failure/malfunction	0.08823	Landing gear	0.00980
Portable electrical equipment	0.00980	Hazardous material (HAZMAT)	0.00980
Miscellaneous/other	0.01960	Engine assembly, other	0.00980
Smoke detector(s)	0.00980	Fire extinguisher, portable	0.00980
Ignition system, ignition harness	0.01960	Maintenance, installation	0.03921
Emergency procedure	0.00980	Airframe/component/system failure/malfunction	0.31372
Fire warning system, lavatory	0.00980	Electrical system, generator	0.00980
Fluid, fuel	0.05882	Engine assembly	0.00980
Auxiliary power unit (APU)	0.04901	Landing gear, tire	0.01960
Fire extinguisher, powerplant	0.01960	Starting procedure	0.00980
Electrical system	0.00980	Compressor assembly, blade	0.00980
Fuselage, cabin	0.00980	Overheat warning system	0.00980
Procedure inadequate	0.03921	Fire extinguisher, cargo	0.01960
Miscellaneous equipment/furnishings, lavatories	0.00980	Ignition system, igniter plug	0.00980
Maintenance	0.02941	Maintenance, overhaul	0.00980
Insufficient standards/requirements, Aircraft	0.00980	Combustion assembly, combustion liner	0.00980
Fire/explosion	0.00980	Fuel system, line fitting	0.01960
Powerplant	0.00980	Insufficient standards/requirements	0.00980
Loss of engine power (partial) - mechanical failure/malfunction	0.03921	Window, flight compartment window/windshield	0.00980
Hydraulic system, line	0.00980	Hazardous materials leak/spill	0.01960
Brakes (normal)	0.01960	Maintenance, approved airworthiness inspection program (AAIP)/progressive program	0.00980
Fuel system, tank	0.00980	Maintenance, alignment	0.00980
Miscellaneous	0.00980	Wing	0.00980
Reason for occurrence undetermined	0.01960	Maintenance, compliance with airworthiness directive (AD)	0.00980
On ground/water collision with object	0.00980	Maintenance, inspection	0.00980
Maintenance, modification	0.01960	Cargo/baggage	0.02941
Electrical system, auxiliary power unit (APU)	0.00980	Electrical system, electric wiring	0.08823
Loss of engine power	0.01960	Lubricating system, oil line	0.00980
Fuel system, fuel flow divider/distributor	0.00980	Procedures/directives	0.01960
Loss of engine power (total) - nonmechanical	0.00980	Fuel system, primer system	0.00980
Weather condition	0.00980		

the developed automation program mentioned before, we generate the XML representation for the large-scale Bayesian network. Fig. 9 shows the visualization of the Bayesian network after the XML file is fed into the software GeNIe modeler from BayesFusion, LLC. The Bayesian network is used to capture the causal relationships embedded in the aviation accidents that happened from 1982 to 2006 in the NTSB aviation accident database. There are 740 nodes and 1300 edges in the Bayesian network, and each node has two states: occur or not occur. Each parent node has a prior probability. For example, there is a probability of $5.4 * 10^{-9}$ for the node *improper training*. Similarly, each child node has a conditional probability table that is used to model the occurrence probability if other events in the parent nodes happen. The developed Bayesian network covers a wide range of accidents and a wide range of factors contributing to the occurrence of such accidents, ranging from human factors (e.g., improper ATC supervision, poor airplane maintenance) to mechanical or electrical component malfunction or failure (e.g., propeller failure, landing gear failure).

Once the Bayesian network is constructed, an important next step is model verification to assess the confidence in the results produced. In this study, we use sensitivity analysis to partially validate the model based on the following two axioms [6,51].

Axiom 1. A slight increase/decrease in the prior subjective probabilities of each parent node should result in a relative increase/decrease in the posterior probability of the child nodes.

Axiom 2. The variation of subjective probability distributions of each parent node, its influence magnitude to the child node values should be consistent.

From this perspective, a valid Bayesian model should at least satisfy the above two axioms. Hence, we perform sensitivity analysis of the influence of parent nodes on child nodes to examine the consistency of the developed model. Fig. 10 shows a small Bayesian network for the sake of demonstration. In this Bayesian network, the parent node *landing main gear strut failure* is the root cause of *main gear collapse* and *gear collapse*. To perform sensitivity analysis, we increase the probability of *landing main gear strut failure* from 0 to 1 gradually, and examine how the probabilities associated with the child nodes change. The *landing main gear strut failure* has two child nodes, namely: *main gear collapse* and *gear collapse*. The collapse of main gear can be caused by a wide variety of factors, such as material inadequate, improper for the landing gear, or the gear locking mechanism malfunctions, etc, while the occurrence *gear collapse* also results from many factors, such as forced landing, insufficient standards, or inappropriate maintenance, to name a few.

Table 8 shows the computational results pertaining to the probability of *main gear collapse* and *gear collapse* when we vary the probability of *landing main gear strut failure* (e.g., fatigue, overload, total failure, cracked) and fix the probabilities of other contributory factors. As can be seen from Table 8, when the probability of *landing main gear strut failure* is set to 6.5×10^{-8}, there is a very low probability of 1.21×10^{-7} for the main landing gear to collapse and a value of 9.51×10^{-8} for the gear to collapse. After we increase the probability of *landing main gear strut failure* by ten times to 6.5×10^{-7}, the probabilities of *main gear collapse* and *gear collapse* increase to a different scale, which are 2.67×10^{-7} and 2.42×10^{-7}, respectively. When the prior probability is lifted to 6.5×10^{-4}, the probabilities of *main gear collapse* and *gear collapse* increase to the same value 1.63×10^{4}. As can be seen, the probabilities of child nodes *main gear collapse* and *gear collapse* increase to another magnitude in comparison with the original values 1.21×10^{-7} and 9.51×10^{-8}.

The same trend can be observed from other cases when the probability of *landing main gear strut failure* increases to 0.1, 0.2, 0.3, 0.5, 0.8, 0.9, and 1.0, respectively. Each time when the probability of *landing main gear strut failure* gets increased, there is an corresponding increase in the probability of *main landing gear collapse* and *gear collapse*. In the worst case, when the probability of *landing main gear strut failure* increases to 1, the probabilities of *main gear collapse* and *gear collapse* increase to the same value 25.00×10^{-2}. Since the probabilities of other contributory factors to *main gear collapse* and *gear collapse* are at the level of 10^{-8}, when the probability of *landing main gear strut failure* increases to 6.5×10^{-6}, it dominates in terms of contributing to the occurrence of *main gear collapse* and *gear collapse*, which leads the occurrence probabilities of *main gear collapse* and *gear collapse* to converge to the same value. The above analysis indicate that there is a consistent relationship between the parent node *landing main gear strut failure* and the subsequent child nodes *main gear collapse* and *gear collapse*. An increase/decrease in the prior probabilities of the parent node result in a relative increase/decrease in the posterior probabilities of child nodes.

Table 8
Sensitivity analysis for Bayesian network: verification of a single parent node.

Probability of landing main gear strut failure	Probability of main gear collapse	Probability of gear collapse
Prior value (6.5×10^{-8})	1.21×10^{-7}	9.51×10^{-8}
6.5×10^{-7}	2.67×10^{-7}	2.42×10^{-7}
6.5×10^{-6}	1.73×10^{-6}	1.70×10^{-6}
6.5×10^{-4}	1.63×10^{-4}	1.63×10^{-4}
6.5×10^{-2}	1.62×10^{-2}	1.62×10^{-2}
0.1	2.50×10^{-2}	2.50×10^{-2}
0.2	5.00×10^{-2}	5.00×10^{-2}
0.3	7.50×10^{-2}	7.50×10^{-2}
0.5	12.50×10^{-2}	12.50×10^{-2}
0.8	20.00×10^{-2}	20.00×10^{-2}
0.9	22.50×10^{-2}	22.50×10^{-2}
1.0	25.00×10^{-2}	25.00×10^{-2}

We perform more comprehensive analysis to further verify the model. To be specific, we calculate the occurrence probability of a child node with multiple parents when more and more evidence on parent nodes becomes available. In particular, we analyze the effect of multiple factors contributing to the occurrence of *main gear collapse*. As mentioned before, many factors contribute to *main gear collapse*, such as *landing main gear attachment failure* (e.g., separation, overload, loose, undertorqued, total failure), *landing main gear failure* (e.g., leak, unlocked, stress corrosion, failure), material inadequate or improper, and *landing gear emergency extension assembly failure* (gusts, jammed, not engaged) etc. The *main gear collapse* leads to three principal consequences: *destroyed aircraft*, *minor aircraft damage*, and *minor personnel injury*. Along with our previous analysis, when we observe the occurrence of *landing main gear strut failure*, the probability of *main gear collapse* is 25×10^{-2}. When we set the probability of *landing gear emergency extension assembly failure* as 1, this results in further increase in the probability of *main gear collapse* to 68.2×10^{-2}. When the *landing gear locking mechanism* fails, the probability of *main gear collapse* further increases to 77.7×10^{-2}. Likewise, if the landing main gear attachment fails, the probability of *main gear collapse* gets further increased to 89.4×10^{-2}. From our analysis, we observe that the *landing gear emergency extension assembly failure* contributes the most to the occurrence of *main gear collapse*, followed by *landing main gear strut failure* and the *landing main gear attachment failure*. Such analysis will offer guidance in the resource prioritization for improving the system safety. Regarding the outcome of *main gear collapse*, Fig. 11 shows the probabilities of aircraft damage and personnel injury as more and more evidences regarding landing gear failure are observed. As can be observed in Fig. 11, the probabilities of *destroyed aircraft*, *minor aircraft damage*, and *minor personnel injury* have all increased in a steady manner as more and more hazardous events accumulate. With the accumulation of hazardous events, the probabilities of aircraft damage at the destroyed and minor levels, and minor personnel injury get increased accordingly as more and more components related to the landing gear malfunction.

The analysis of increasing the probability of each parent node shows that its effect on the child node is in compliance with the axioms stated

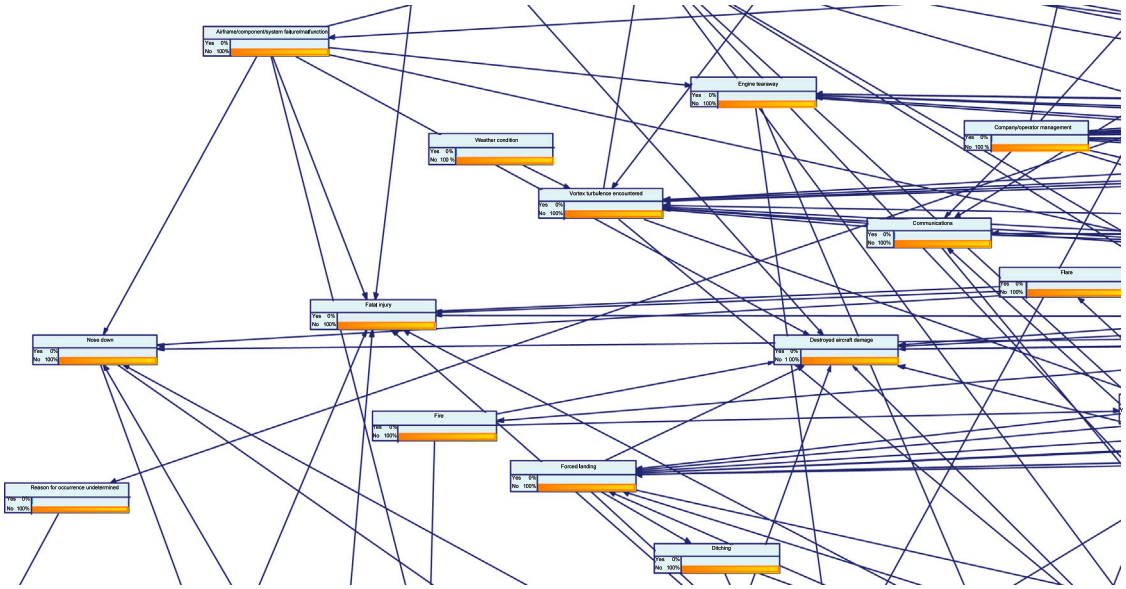


Fig. 9. A partial view of Bayesian network-based representation for NTSB aviation accidents. A full version of the Bayesian network is available at https://github.com/zxgcqupt/NTSB_Bayesian_Network/blob/master/NTSB.pdf.

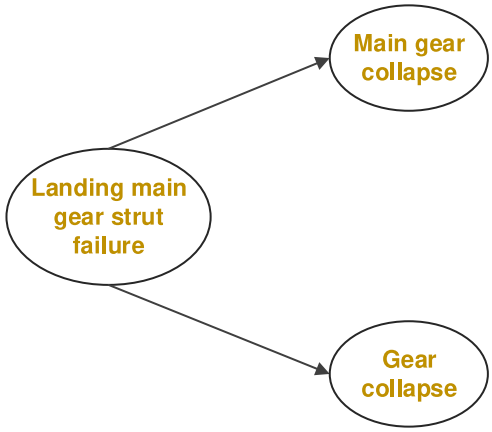


Fig. 10. A Bayesian network for landing gear main gear strut failure.

before, thus providing a partial verification to the model. Validation of the model can be performed when the data from the aircraft operations becomes available. Real-world data can be used to further validate the probabilities approximated in this paper, but a common challenge is that aviation accidents are rare events, therefore it is difficult to obtain the actual occurrence probability of each event even with the real-world data.

5.3. Scenario analysis

The proposed approach enables the risk analysis of flight safety in a variety of contexts ranging from procedure non-compliance to equipment impairment or malfunction. In this section, we analyze two scenarios with the developed Bayesian network for the sake of demonstrating its usefulness in accident analysis and investigation.

5.3.1. Scenario one: Pilot error and unstable approach

An approach is considered as unstable if at least one of the following variables are not maintained stable during approach: speed, descent rate, vertical/lateral flight path and in landing configuration [52]. In the first scenario, we analyze the impact of the pilot-in-command's erroneous action on the safety of the flight. In the developed Bayesian network, by setting the probability of *pilot error* as 1, we observe that the probability of *unstable approach* increases from 2.71×10^{-8} to 4.84×10^{-3}. As a result of the increase in the probability of *unstable approach*, the probability of *wing, tail or other components dragged on the runway* increases from 1.14×10^{-7} to 2.30×10^{-2}. Likewise, the probability of *no injury* reduces from 99.99×10^{-2} to 97.0×10^{-2}, and the probability of the aircraft having substantial damage increases from 2.22×10^{-7} to 4.58×10^{-2}.

Fig. 12 shows the propagation of influence as a result of the observed evidence regarding pilot error. As can be observed, the increase in the probability of substantial aircraft damage not only results from *unstable approach*, but also from the *hard landing* as a result of *improper flare* (e.g., misjudged flare, improper flare, inadequate flare) due to pilot error. Suppose now we observe the occurrence of *unstable approach*, as shown in Fig. 12, the probabilities of events following unstable approach are updated accordingly again. In particular, the probability of *wing, tail or other components dragged on the runway* increases remarkably to 41.72×10^{-2} compared to its initial state 2.30×10^{-2}. Meanwhile, the chance for *substantial aircraft damage* increases to 24.64×10^{-2}, and the probability of *no personnel injury* reduces significantly from 97.0×10^{-2} to 61.30×10^{-2}. This analysis reveals that the developed Bayesian network can dynamically update our understanding on the flight safety, and the probabilities reported from the Bayesian network can be used as safety indicators in flight safety assessment. In this case, with the analysis result from the Bayesian network, the airport can prepare the necessary resources, such as ambulance for the passengers or flight crews that might get injured, and response teams to limit the damage to the aircraft; and this could be one of the possible future benefits of the proposed analysis approach.

5.4. Scenario two: Loss of engine power

In this scenario, we analyze the impact of loss of engine power on flight safety. Loss of engine power has been reported to cause forced landing and fatality [53,54]. A variety of factors can result in the loss of engine power, for example, improper fluid, oil grade, or contaminated oil; failed, seized, or burned engine assembly, piston; fractured, overload, or fatigue of engine assembly, connecting rod; misread, nor unverified or improper use of engine instrument, etc. In this section, we analyze the impact of three possible causes to the loss of engine power

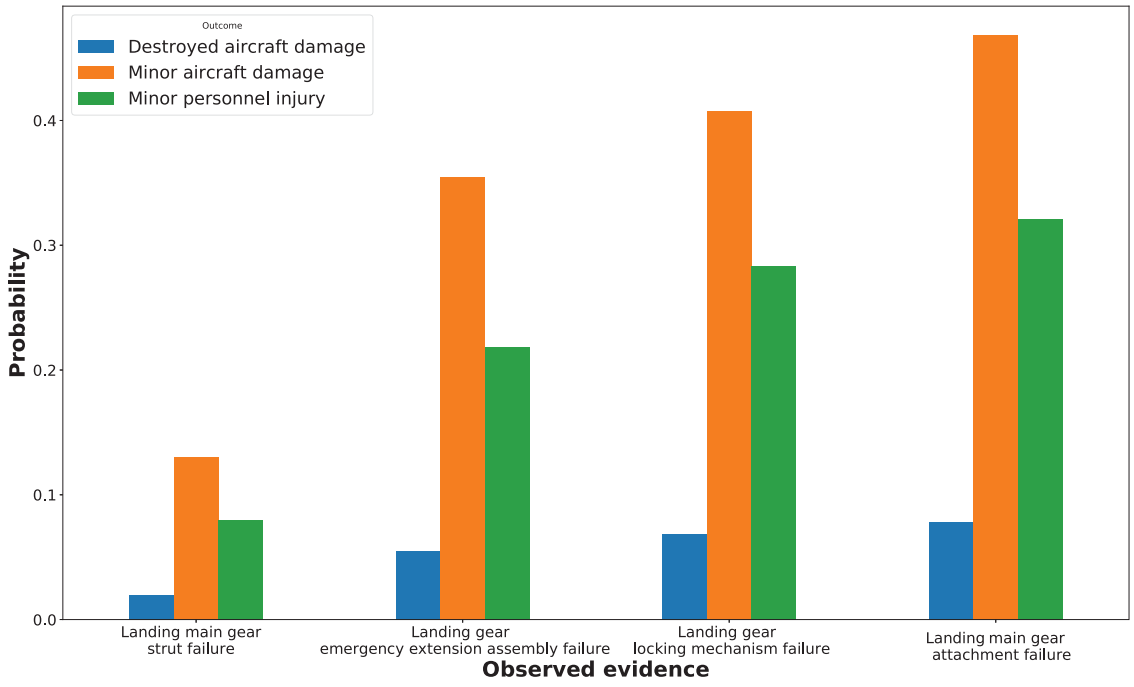


Fig. 11. The probability of aircraft damage and personnel injury as a result of main gear collapse when more and more evidences are observed.

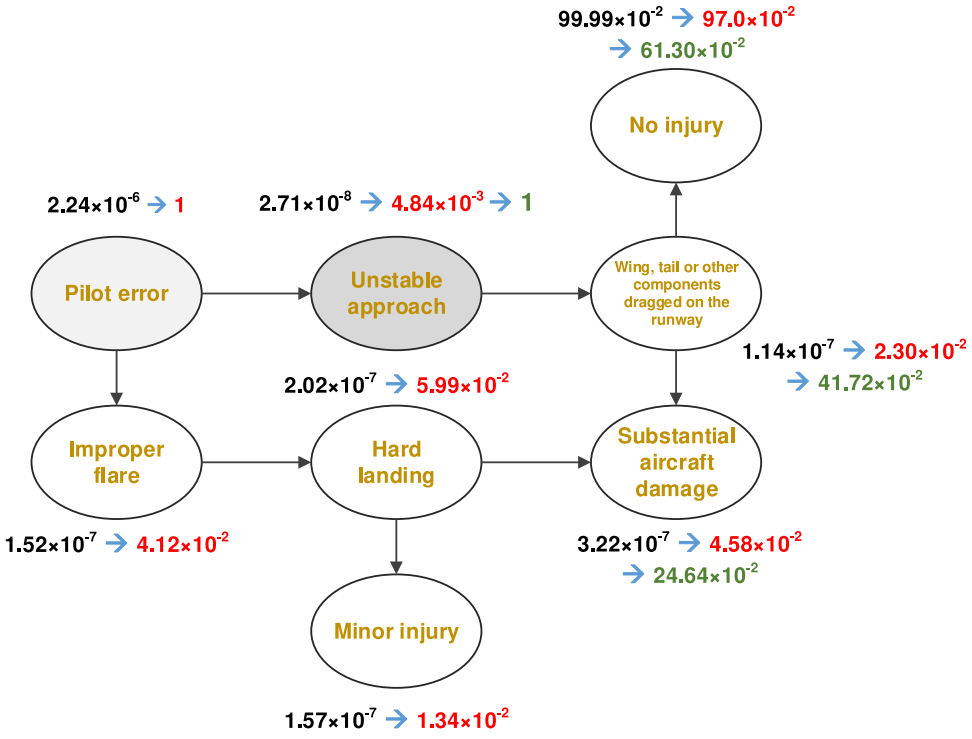


Fig. 12. The influence propagation in the Bayesian network as a result of observing pilot-in-command's erroneous action, where black fonts represent the prior probability of each event, red fonts denote the updated posterior probability after observing the first evidence – pilot-in-command, and green fonts denote the posterior probability after observing the second evidence – unstabilized approach. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and flight safety, namely: engine instrument failure (e.g., misread, not verified, inaccurate, improper use), combustion assembly/liner failure (e.g., distorted, fatigue), and fluid, oil grade failure (e.g., improper, contamination).

Fig. 13 demonstrates the influence propagation of observed evidence through the Bayesian network, where the solid lines denote the forward propagation of observed evidence in the constructed Bayesian network, and the dashed arrows represent the backward inference of the observed evidence (inoperative engine instruments) as well as the propagation of the inferred information through the network. As shown in the figure, the *loss of engine power* leads to *forced landing*, and the forced landing results in the collapse of main gear and other gear, as well as aircraft damage (substantial, minor, and destroyed) and personnel injury (serious and no injury). Since it is important to characterize the relative contributions of different upstream events leading to *loss of engine power*, we perform multiple studies by checking

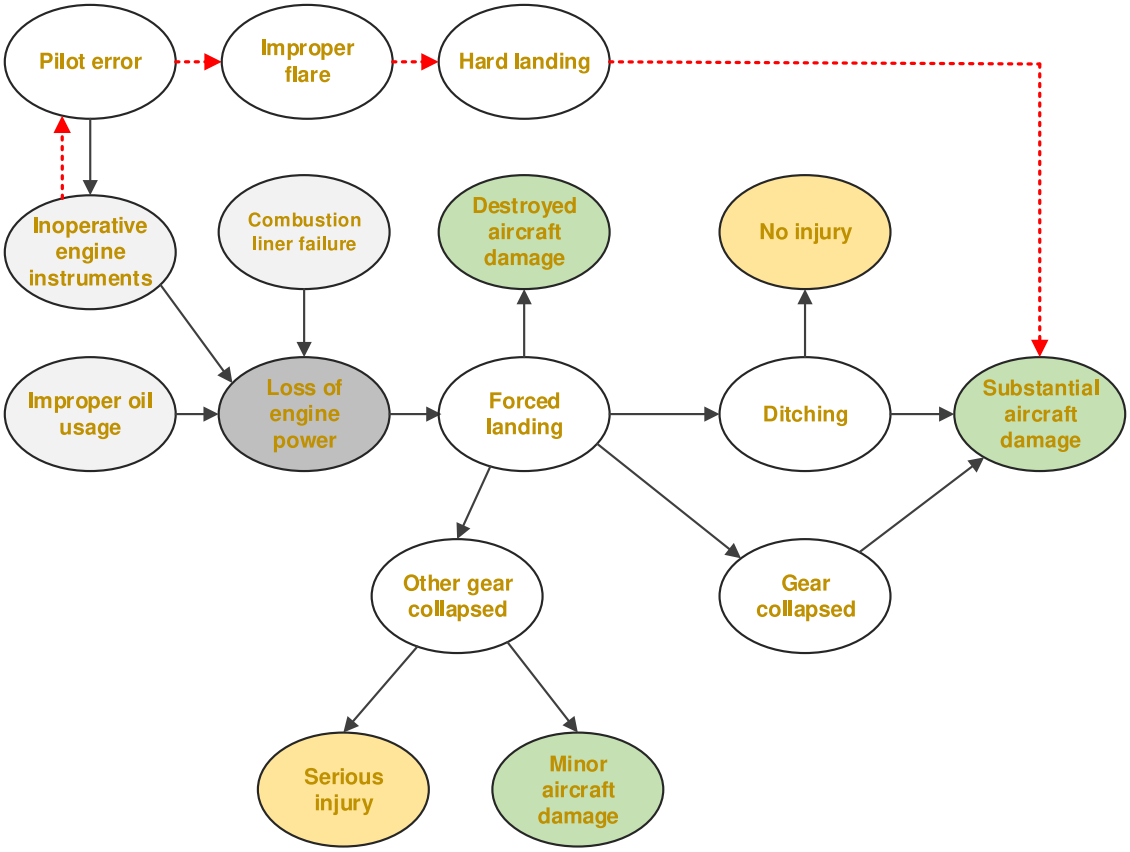


Fig. 13. Influence propagation as a result of the loss of engine power.

the effect of observing the occurrence of different causes leading to the loss of engine power. In this study, we achieve this goal by analyzing the contributions of different upstream events to the occurrence probability of downstream events that eventually converge to *loss of engine power*. Table 9 provides an end-to-end contributions of different upstream events to the occurrence probabilities of various downstream events, both direct and indirect. As shown in Table 9, among the three causes we analyzed, *inoperative engine instruments* and *improper oil usage* contribute to *loss of engine power* at the same level. When something is wrong with the engine instruments or oil grade, there is a probability of 0.95 for the *loss of engine power*. Whereas, if the *combustion liner* fails, the loss of engine power has a probability of 0.50 to occur. When there are failures w.r.t. both engine instruments and oil usage, the probability of *loss of engine power* increases to 0.99. Among all the evidences, *inoperative engine instruments* and *improper oil usage* have the highest combined effect on aircraft damage across all the levels and personnel injury, followed by *inoperative engine instruments* and *loss of engine power*. Regarding ditching, *loss of engine power* leads to the highest probability of ditching with a value of 4.61×10^{-3}, followed by *inoperative engine instruments*.

Next, we analyze event outcomes from two separate individual perspectives: aircraft damage and personnel injury. An interesting finding is that *inoperative engine instruments* causes the highest probability among all the individual contributors, and it is even higher than the case that we observe *loss of engine power*. As shown in Fig. 13, when we observe engine instrument failure, the Bayesian network updates the probability of *pilot error* because it is the only contributory factor to *inoperative engine instruments*. The pilot error might lead to *improper flare*, which eventually results in *hard landing*. The increase in the probability of *hard landing* further increases the probability of *substantial aircraft damage*. The aforementioned information propagation is represented as red dashed arrows in Fig. 13. In this discussion, we only list this alternative path for the sake of illustration. In the actual Bayesian network, there are many other paths leading to the same outcome of aircraft damage and personnel injury.

The ability to aggregate information in both forward and backward propagation makes the Bayesian network an ideal tool to fuse information and dynamically update our understanding on the system safety. The Bayesian network construction approach developed in this paper can be utilized to support decision making activities regarding flight safety assessment, and also accident investigation of future accidents by providing systematically organized quantitative information on event sequences in past accidents.

6. Discussion

In this paper, we develop a Bayesian network to represent the causal relationships among a wide variety of factors contributing to passenger airliner accidents and incidents. To achieve this goal, we propose a four-step methodology to construct the Bayesian network representation of accident investigation data. In the first step, for each individual aviation accident, an end-to-end graphical representation is developed to depict the propagation from initiating events to accidents, where all the causal events (e.g., personnel, equipment, procedures, management) identified in the NTSB investigation reports are included. Secondly, all the individual graphs are aggregated together to form an aggregated Bayesian network, where marginal probabilities are used for root nodes, and conditional probabilities are used to describe the probability relationships between parent and child nodes. Second, the marginal probabilities are estimated as the ratio of the occurrence times of each accident to the number of aircraft departures from the Bureau of Transportation Statistics (BTS). Third, to estimate conditional probabilities, the CDF of a beta distribution with two unknown parameters α and β is used to approximate the conditional probabilities if the event

Table 9
The posterior probabilities related to each event when observing different evidences.

Evidence	Inoperative engine instruments	Combustion liner failure	Improper oil usage	Inoperative engine instruments & Improper oil usage	Loss of engine power
Loss of engine power	0.95	0.50	0.95	0.99	1
Forced landing	13.57×10^{-2}	7.14×10^{-2}	13.57×10^{-2}	14.71×10^{-2}	14.29×10^{-2}
Ditching	4.37×10^{-3}	2.30×10^{-3}	4.37×10^{-3}	4.57×10^{-3}	4.61×10^{-3}
Gear collapsed	9.60×10^{-3}	2.30×10^{-3}	4.37×10^{-3}	9.82×10^{-3}	5.18×10^{-3}
Other gear collapsed	4.80×10^{-3}	2.30×10^{-3}	4.37×10^{-3}	5.00×10^{-3}	4.66×10^{-3}
Destroyed aircraft	1.33×10^{-2}	2.30×10^{-3}	4.37×10^{-3}	1.35×10^{-2}	5.59×10^{-3}
Substantial aircraft damage	4.60×10^{-2}	3.63×10^{-3}	6.09×10^{-3}	4.63×10^{-2}	1.66×10^{-2}
Minor aircraft damage	9.34×10^{-3}	1.54×10^{-3}	2.92×10^{-3}	9.47×10^{-3}	3.78×10^{-3}
Serious injury	6.23×10^{-2}	7.68×10^{-4}	1.46×10^{-3}	6.23×10^{-2}	8.22×10^{-3}
No injury	94.31×10^{-2}	99.78×10^{-3}	99.58×10^{-3}	94.29×10^{-2}	98.99×10^{-2}

ω is conditional on multiple events, based on the quantified conditional probabilities on individual events. The two unknown parameters α and β are then calibrated with the probabilities $P(\omega|e)$ derived from the available data. Finally, we develop a procedure to automate the generation of the aggregated Bayesian network consisting of more than 500 nodes and 1000 links in the XML format. The generated Bayesian network is fed into the commercial software GeNIe modeler for accident analysis and investigation in a straightforward manner.

6.1. Contributions

The principal merit of the proposed methodology is in developing an end-to-end framework to construct a large Bayesian network to capture the causal-effect relationship across a large number of aviation accident events in an automatic manner. The methodology developed in this paper benefits the assessment of aviation safety in several aspects. First of all, we develop an end-to-end graphical representation to capture the sequences and outcomes of each event as reported in the NTSB accident database. The graphical representation greatly facilitates our understanding on the escalation path of initiating events as an accident in the system.

Secondly, we aggregate event-wise graphical representations to form an aggregated Bayesian network to capture the causal relationships across all the events in the NTSB accident database. By doing this, we create the structure of Bayesian network to represent all the event sequences and their outcomes. To automate the construction of the entire Bayesian network, we generate an XML file in compliance with the syntax used in GeNIe modeler. The XML-based representation is directly consumable by GeNIe modeler. The automation program greatly facilitates the update of the Bayesian network if extra data is collected. In the future, when more data is available in the NTSB aviation database, the program can load all the data and generate a new XML-based representation of Bayesian network automatically, which can then be fed into GeNIe modeler for event cause and outcome analysis.

Thirdly, the marginal probabilities of root nodes are estimated with the aircraft departure data from the Bureau of Transportation Statistics (BTS), and the conditional probabilities are approximated using a flexible beta distribution. The two unknown shape parameters α and β in the beta distribution are calibrated with the probabilities of events conditional on single event that are estimated from the data directly. By casting the parameter calibration as an optimization problem, we achieve good prediction performance as reflected by the low mean squared error and standard deviation of prediction discrepancy.

Last but not the least, we verify the performance of developed Bayesian network through what-if and sensitivity analysis on the prior distribution. The usefulness of the developed Bayesian network is demonstrated in two scenarios, namely unstabilized approach and loss of engine power, which illustrates the powerful capability of the Bayesian network as a promising information fusion model for accident analysis. Given the publicly accessible source codes, more comprehensive and customized analysis can be performed with the Bayesian network developed in this paper to explore its extensive applications for aviation accident cause and outcome analysis.

6.2. Limitations

Several limitations need to be addressed in future work. First of all, in the NTSB database, injury data is available from 1982, whereas damage data is only available starting in 2000. This affects the prior and conditional probabilities of some events, as well as the structure of the network. Right now, comparison of our results with the earlier Bayesian network (CATS) [3,33] could not be done because our current model is less detailed and therefore has insufficient granularity. It is important to investigate other data sources, e.g., ADERP database from ICAO, to overcome the missing data issue in NTSB database, enrich the accident dataset, and increase the resolution of the event sequences given in the accident reports based on the detailed narrative available in the reports, thereby improving the quality of the developed Bayesian network and facilitate comparison with the CATS Bayesian network. Second, the levels of personnel injury and aircraft damage are different. Specifically, personnel injury has four levels of ratings: fatal, serious, minor, and none, while aircraft damage only has three levels of ratings: substantial, minor, and none. It will be valuable to investigate approaches to convert the different levels of personnel injury and aircraft damage to a common risk metric in order to support decision making activities in the future. Third, due to the limited data, we assume the parent nodes of ω $(e_1, e_2, \ldots, e_m)$ are independent from each other, and this is an important assumption about priors in Bayesian analysis. In future work, simulation data can be leveraged to enrich the limited dataset, and help examine the relationships between the events $e_1, e_2, \ldots, e_m$. This will help to improve the structure and probabilities of the Bayesian network, and improve its accuracy for aviation accident analysis accordingly. Fourth, as mentioned earlier, the current software uses a maximum of 12 parent nodes; in future, scalable inference algorithms need to be developed to speed up the Bayesian updating of large-size networks. The development of such algorithms will facilitate the construction of a full-resolution Bayesian network for aviation accident analysis. Fifth, the prior and conditional probabilities are estimated by summing over the entire time period (from 1982 to 2006) whereas these quantities actually vary over time due to the introduction of new air traffic management provisions and rules as well as upgrades in commercial aviation technologies. It is worthwhile to develop an approach that represents the probabilities as time-dependent quantities in support of more granular aviation accident analysis.

6.3. Future work

Future work can be carried out in the following directions. First of all, the Bayesian network developed in this paper can be further refined by incorporating additional valuable information pertaining to aircraft characteristics (e.g., aircraft type, engine running time, engine type), weather conditions (e.g., wind direction, wind speed), and pilot information (e.g., pilot flight hours, pilot experience), provided that these data can be collected from the various sources. The incorporation

of these data will further increase the accuracy of the Bayesian network when performing forward propagation and backward inference in accident analysis and investigation. Secondly, the software GeNIe modeler is computationally time-consuming when performing analysis of such a large-scale Bayesian network. More efficient sampling algorithms or parallelization mechanisms need to be investigated to speed up the analysis in large scale Bayesian networks in order to make it ready to be used for real-time analysis. Thirdly, simulation data can be collected and utilized to estimate some of the conditional probabilities in the cases when ω is conditional on multiple events. By doing this, more informative data can be used to calibrate the unknown parameters α and β in the function used to approximate the conditional probabilities, thereby enhancing the accuracy of the aggregated Bayesian network.

CRediT authorship contribution statement

Xiaoge Zhang: Conceived of the presented idea, Developed the program, Performed the computations, Writing - original draft. **Sankaran Mahadevan:** Discussed the conceived idea, Supervison, Writing - review & editing.

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.

Acknowledgments

The research reported in this paper was supported by funds from NASA University Leadership Initiative program, USA (Grant No. NNX17AJ86A, Technical Monitor: Dr. Anupa Bajwa) through subcontract to Arizona State University (Principal Investigator: Dr. Yongming Liu). All the Bayesian network models in this paper were created and tested using SMILE, an inference engine, and GeNIe, a development environment for reasoning in graphical probabilistic models, both from BayesFusion, LLC and available at https://www.bayesfusion.com/.

References

[1] IATA forecast predicts 8.2 billion air travelers in 2037. 2019, https://www.iata.org/pressroom/pr/Pages/2018-10-24-02.aspx, Accessed 25 September 2019.
[2] Zhang X, Mahadevan S. Ensemble machine learning models for aviation incident risk prediction. Decis Support Syst 2019;116:48–63.
[3] Ale BJ, Bellamy L, Van der Boom R, Cooper J, Cooke RM, Goossens LH, Hale A, Kurowicka D, Morales O, Roelen A, et al. Further development of a causal model for air transport safety (CATS): Building the mathematical heart. Reliab Eng Syst Saf 2009;94(9):1433–41.
[4] Zhang X, Mahadevan S, Deng X. Reliability analysis with linguistic data: An evidential network approach. Reliab Eng Syst Saf 2017;162:111–21.
[5] Boudali H, Dugan JB. A discrete-time Bayesian network reliability modeling and analysis framework. Reliab Eng Syst Saf 2005;87(3):337–49.
[6] Jones B, Jenkinson I, Yang Z, Wang J. The use of Bayesian network modelling for maintenance planning in a manufacturing industry. Reliab Eng Syst Saf 2010;95(3):267–77.
[7] Dai J, Deng Y. A new method to predict the interference effect in quantum-like Bayesian networks. Soft Comput 2020;24:10287–94.
[8] Liang Y, Lee JD. A hybrid Bayesian network approach to detect driver cognitive distraction. Transp Res C 2014;38:146–55.
[9] Baraldi P, Podofillini L, Mkrtchyan L, Zio E, Dang VN. Comparing the treatment of uncertainty in Bayesian networks and fuzzy expert systems used for a human reliability analysis application. Reliab Eng Syst Saf 2015;138:176–93.
[10] Khakzad N, Khan F, Amyotte P. Dynamic safety analysis of process systems by mapping bow-tie into Bayesian network. Process Saf Environ Prot 2013;91(1–2):46–53.
[11] Dejaeger K, Verbraken T, Baesens B. Toward comprehensible software fault prediction models using Bayesian network classifiers. IEEE Trans Softw Eng 2012;39(2):237–57.
[12] Wang H, Fang Y-P, Zio E. Risk assessment of an electrical power system considering the influence of traffic congestion on a hypothetical scenario of electrified transportation system in new york state. IEEE Trans Intell Transp Syst 2019.
[13] Zhang X, Mahadevan S. Bayesian neural networks for flight trajectory prediction and safety assessment. Decis Support Syst 2020;113246.
[14] Weber P, Medina-Oliva G, Simon C, Iung B. Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. Eng Appl Artif Intell 2012;25(4):671–82.
[15] Trucco P, Cagno E, Ruggeri F, Grande O. A Bayesian belief network modelling of organisational factors in risk analysis: A case study in maritime transportation. Reliab Eng Syst Saf 2008;93(6):845–56.
[16] Khakzad N, Khan F, Amyotte P, Cozzani V. Domino effect analysis using Bayesian networks. Risk Anal Int J 2013;33(2):292–306.
[17] Zhang J, Shields MD. Efficient Monte Carlo resampling for probability measure changes from Bayesian updating. Probab Eng Mech 2019;55:54–66.
[18] Zhang D, Yan X, Yang ZL, Wall A, Wang J. Incorporation of formal safety assessment and Bayesian network in navigational risk estimation of the yangtze river. Reliab Eng Syst Saf 2013;118:93–105.
[19] Wang L, Yang Z. Bayesian network modelling and analysis of accident severity in waterborne transportation: A case study in China. Reliab Eng Syst Saf 2018;180:277–89.
[20] Chen C, Liu X, Chen H-H, Li M, Zhao L. A rear-end collision risk evaluation and control scheme using a Bayesian network model. IEEE Trans Intell Transp Syst 2018;20(1):264–84.
[21] Luxhoj JT. Probabilistic causal analysis for system safety risk assessments in commercial air transport. In: Second Workshop on the Investigation and Reporting of Incidents and Accidents, IRIA. 2003.
[22] Ale BJ, Bellamy L, Cooper J, Ababei D, Kurowicka D, Morales O, Spouge J. Analysis of the crash of TK 1951 using CATS. Reliab Eng Syst Saf 2010;95(5):469–77.
[23] Papazoglou IA, Aneziris ON, Bellamy L, Ale BJ, Oh J. Quantitative occupational risk model: Single hazard. Reliab Eng Syst Saf 2017;160:162–73.
[24] Luxhøj JT, Choopavang A, Arendt DN. Risk assessment of organizational factors in aviation systems. Air Traff Control Quart 2001;9(3):135–74.
[25] Ale B, Bellamy L, Cooke R, Duyvis M, Kurowicka D, Lin C, Morales O, Roelen A, Spouge J. Causal model for air transport safety. Final Rep July 2008;31.
[26] Ale B, Van Gulijk C, Hanea A, Hanea D, Hudson P, Lin P-H, Sillem S. Towards BBN based risk modelling of process plants. Saf Sci 2014;69:48–56.
[27] Greenberg R, Cook S, Harris D. A civil aviation safety assessment model using a Bayesian belief network (BBN). Aeronaut J 2005;109(1101):557–68.
[28] Stamatelatos M, Dezfuli H, Apostolakis G, Everline C, Guarro S, Mathias D, Mosleh A, Paulos T, Riha D, Smith C, et al. Probabilistic risk assessment procedures guide for NASA managers and practitioners. 2011.
[29] Ancel E, Shih AT, Jones SM, Reveley MS, Luxhøj JT, Evans JK. Predictive safety analytics: inferring aviation accident shaping factors and causation. J Risk Res 2015;18(4):428–51.
[30] National transportation safety board aviation database. 2019, https://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx, Accessed 24 June 2019.
[31] Regularly scheduled air carriers (part 121). 2019, https://www.faa.gov/hazmat/air_carriers/operations/part_121/, Accessed 24 June 2019.
[32] Genie modeler: Complete modeling freedom. 2019, https://www.bayesfusion.com/genie/, Accessed 24 June 2019.
[33] Ale B, Bellamy L, BV WQ, Roelen IA, Cooke R, Goossens L, Hale A, Kurowicka D, Smith E. 2005 ASME International Mechanical Engineering Congress and Exposition November 5-11, 2005, Orlando, Florida USA: 2005.
[34] Ale B, Bellamy L, Van der Boom R, Cooke R, Goossens L, Hale A, Kurowicka D, Lin P, Roelen A, Cooper H, et al. Further development of a causal model for air transport safety (CATS): the complete model. In: Ninth International Probabilistic Safety Assessment and Management Conference. 2008, p. 18–23.
[35] NTSB aviation coding manual. 2019, https://www.ntsb.gov/GILS/Documents/codman.pdf, Accessed 24 June 2019.
[36] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier; 2014.
[37] Sun S, Zhang C, Yu G. A Bayesian network approach to traffic flow forecasting. IEEE Trans Intell Transp Syst 2006;7(1):124–32.
[38] Cano A, Masegosa AR, Moral S. A method for integrating expert knowledge when learning Bayesian networks from data. IEEE Trans Syst Man Cybern B 2011;41(5):1382–94.
[39] Jensen FV, et al. An introduction to Bayesian networks, Vol. 210. UCL Press London; 1996.
[40] Zhang L, Wu X, Skibniewski MJ, Zhong J, Lu Y. Bayesian-network-based safety risk analysis in construction projects. Reliab Eng Syst Saf 2014;131:29–39.
[41] D'Addabbo A, Refice A, Pasquariello G, Lovergine FP, Capolongo D, Manfreda S. A Bayesian network for flood detection combining SAR imagery and ancillary data. IEEE Trans Geosci Remote Sens 2016;54(6):3612–25.
[42] Khakzad N, Khan F, Amyotte P. Safety analysis in process facilities: Comparison of fault tree and Bayesian network approaches. Reliab Eng Syst Saf 2011;96(8):925–32.
[43] Washington A, Clothier R, Neogi N, Silva J, Hayhurst K, Williams B. Adoption of a Bayesian belief network for the system safety assessment of remotely piloted aircraft systems. Saf Sci 2019;118:654–73.
[44] Zhang X, Mahadevan S, Lau N, Weinger MB. Multi-source information fusion to assess control room operator performance. Reliab Eng Syst Saf 2020;194:106287.

[45] U.S. air carrier aircraft departures, enplaned revenue passengers, and enplaned revenue tons. 2019, https://www.bts.gov/us-air-carrier-aircraft-departures-enplaned-revenue-passengers-and-enplaned-revenue-tons, Accessed 24 November 2019.
[46] Bury K. Statistical distributions in engineering. Cambridge University Press; 1999.
[47] Teunis P, Havelaar A. The beta Poisson dose-response model is not a single-hit model. Risk Anal 2000;20(4):513–20.
[48] Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. J Appl Stat 2004;31(7):799–815.
[49] SMILE: Structural modeling, inference, and learning engine. 2019, https://www.bayesfusion.com/smile/, Accessed 24 November 2019.
[50] Luersen MA, Le Riche R. Globalized nelder–mead method for engineering optimization. Comput Struct 2004;82(23–26):2251–60.
[51] Yang Z, Bonsall S, Wang J. Fuzzy rule-based Bayesian reasoning approach for prioritization of failures in FMEA. IEEE Trans Reliab 2008;57(3):517–28.
[52] Rao AH, Puranik TG. Retrospective analysis of approach stability in general aviation operations. In: 2018 Aviation Technology, Integration, and Operations Conference. 2018, p. 3049.
[53] Loss of engine power fatal for passenger. 2019, https://generalaviationnews.com/2019/06/18/loss-of-engine-power-fatal-for-passenger/, Accessed 10 December 2019.
[54] 'Total loss of engine power' caused plane to land in susquehanna river, NTSB report says. 2019, http://pressandjournal.com/stories/total-loss-of-engine-power-caused-plane-to-land-in-susquehanna-river-ntsb-report-says,69433, Accessed 10 December 2019.