A data-based comparison of BN-HRA models in assessing human error probability
An offshore evacuation case study

Abrishami, Shokoufeh; Khakzad, Nima; Hosseini, Seyed Mahmoud

DOI
10.1016/j.ress.2020.107043

Publication date
2020

Document Version
Accepted author manuscript

Published in
Reliability Engineering and System Safety

A data-based comparison of BN-HRA models in assessing human error probability: An offshore evacuation case study

Shokoufeh Abrishami [a,b], Nima Khakzad [c,*], Seyed Mahmoud Hosseini [a]

[a] Industrial Engineering Department, Ferdowsi University of Mashhad, Iran

[b] Faculty of Technology, Policy, and Management, Delft University of Technology, The Netherlands

[c] School of Occupational and Public Health, Ryerson University, Toronto, Canada

[*]Corresponding author: nima.khakzad@ryerson.ca (N. Khakzad)

Address: 288 Church Street, Toronto, Canada M5B 1Z5

Abstract

Bayesian Network (BN) has been increasingly exploited to improve different aspects of Human Reliability Analysis (HRA), resulting in a new generation of HRA techniques, known as BN-HRA models. However, validating and evaluating the accuracy of BN-HRA models is still a challenging task. In this study, we have assessed and compared the performance of some of well-known BN-HRA techniques using human performance data obtained from an offshore evacuation simulation. Based on the role of data in quantifying the BN-HRA models, three categories of BN-HRA models have been considered: (i) BN-CREAM and BN-SPARH, which are based on predefined rules (rule-based methods), (ii) Bayesian Parameter Learning (BPL), which is entirely based on the available data (data-based method), and (iii) BN-SLIM model which is based on both the available data and the predefined rules (hybrid method). The results of the present study show that the data-based methods, i.e., BN-SLIM and BPL, in general outperform the rule-based methods. Cross-validation analysis further demonstrates the superiority of BN-SLIM over BPL, particularly in case of data scarcity.

Keywords

Human reliability assessment; k-fold cross validation; BN-CREAM; BN-SPARH; BN-SLIM; Bayesian parameter learning.

1. Introduction

Human factor is one of the main causes of technological accidents, causing environmental damage, major capital losses, and noticeable death toll [1-3]. Human Reliability Analysis (HRA) methods such as CREAM [4], SLIM [5], and SPAR-H [6] have been developed to identify potential human errors and estimate their occurrence probability in the operation of complex systems and processes. An integral part of HRA methods is assessing the performance shaping factors (PSFs), which characterize the context and human aspects of human failure events [7]. HRA methods provide instructions for calculating the conditional Human Error Probability (HEP) during a task in a particular context [8]. On the other hand, a nominal HEP of a given task is the probability of human error when the impact of different contexts on human performance is not considered [9].

The conventional HRA methods have some limitations such as being highly subjective [8, 10, 11], lacking a causal mechanism to link PSFs to the operator performance [12, 13], ineffective in incorporating multiple data sources [10, 14], being deterministic and thus not fully capable of handling uncertainties [8, 10, 15, 16], and not easily compatible with system safety assessment models [8, 13]. To mitigate these shortcomings some researchers have employed Bayesian network (BN) to enhance and extend the conventional HRA models [10].

BN has been introduced as a significant element in the third generation of HRA methods – a generation with more insight into HRA data [14, 17]. BN can effectively model the causal relationships between PSFs and respective human failure events while considering dependencies among the PSFs. BN's ability in combining different sources of information allows the development of HRA models with a stronger basis in cognitive theory and empirical data [8]. Moreover, BN is able to handle uncertainty primarily by assigning prior probability distributions to the PSFs and by updating these priors as new information becomes available, leading to more objective results [18]. BN has also been employed to assess the PSFs and quantify their joint impact on HEP based on expert judgment and empirical data [12, 19, 20].

The integration of BN with the conventional HRA methods has lead to what are generally known as BN-HRA methods, such as BN-SPARH [8], BN-CREAM [16], and BN-SLIM [15]. The causal framework of BN-HRA methods can provide a proactive approach for preventing human errors under different contextual conditions [15]. Moreover, BN-HRA methods are able to work with perfect, partial or very little information on the PSFs [8]. Both conventional HRA methods [21–24] and BN-HRA methods [8, 10, 13, 15, 16] have been widely used in system safety and risk assessment for assessing and reducing HEPs. However, despite the obvious advantages of BN-HRA methods over their conventional counterparts, studies on the performance and accuracy of BN-HRA methods have been very limited (e.g., [10]), particularly using empirical and simulation data (e.g. [25, 26]). The lack of comparative studies, in turn, may leave the impression that since the BN-HRA methods are built on BN would all result in more or less the same HEP for a given task. Therefore, the present study can be considered as an attempt to provide more insight into the performance of some BN-HRA methods using the simulation data generated in an offshore evacuation virtual environment [27].

For the sake of clarity, in the present study we have considered four BN-HRA methods and categorized them into three groups based on the role of data in developing the required conditional probability tables needed to quantify the BN models. The first group includes the BN-CREAM [16] and BN-SPARH [8] which use predefined relationships and cognitive theories to calculate the probabilities. The second group includes a BN which uses the maximum likelihood estimation [28] for calculating the conditional probabilities merely based on the available data. The third group includes a refined version of the BN-SLIM [15], which can be considered as a hybrid model that uses both the available data and the predefined relationships of the original SLIM to calculate the conditional probabilities. It is also worth noting that to perform a quantitative comparison among the foregoing BN-HRA methods, it was inevitable to make assumptions and adjustments both to the BN-HRA methods and the dataset, resulting in the customized BN-HRA models in the present study (These adjustments will be further discussed in the respective sections.). As such, the results of the present study should not be generalized as the results of the original BN-HRA methods.

The rest of the paper is organized as follows: Section 2 briefly revisits the CREAM, SPAR-H and SLIM methods. Section 3 recapitulates the basics of BN, Bayesian parameter learning, and the BN versions of the foregoing HRA methods. In Section 4, the foregoing methods are applied to the simulation data, and their accuracy is evaluated. Section 5 concludes the study.

2. Human reliability assessment methods

2.1. SPAR-H

The SPAR-H method was developed for the U.S. nuclear regulatory commission to be used in probabilistic safety analysis models [6]. This method considers two nominal HEPs (NHEPs) of 0.001 and 0.0001 for two task types of diagnosis and action, respectively. The model uses eight predefined PSFs to represent the performance context and to estimate the conditional HEPs given a particular context. The PSFs are “available time”, “stressors”, “complexity”, “experience/training”, “procedures”, “ergonomics/HMI”, “fitness for duty” and “work processes”. These PSFs are fixed and should be applied to any context regardless of their relevance. Each PSF has a certain number of states each with a particular assigned multiplier S [6]. For instance, for the PSF “experience/training”, the sets of states and their corresponding multipliers are States = {High, Nominal, Low, Insufficient information} and S = {0.5, 1, 3, 1}. Having the state of each PSF identified, Eq. (1) is used to estimate the HEP if the number of negative PSFs (PSFs with a multiplier greater than 1) is less than three; otherwise Eq. (2) is used. S_i is the multiplier of the i-th PSF (i = 1, …, 8).

$$HEP = \frac{NHEP \prod_1^8 S_i}{NHEP(\prod_1^8 S_i - 1) + 1} \tag{1}$$

$$HEP = NHEP \prod_1^8 S_i \tag{2}$$

2.2. CREAM

CREAM was developed by Hollnagel [4] to be used in the general applications of HRA. This method represents a contextual control model and defines four categories for the control mode, namely: scrambled, opportunistic, tactical and strategic, which are ordered ascendingly with regard to the degree of control. The control modes are related to different HEP intervals as presented in Table 1.

Table 1. Control modes and probability intervals in CREAM [4]

Control Modes	**HEP intervals**
Strategic	5.0 E-06 < HEP < 0.01
Tactical	0.001 < HEP < 0.1
Opportunistic	0.01 < HEP < 0.5
Scramble	0.1 < HEP < 1.0

In the original CREAM, nine Common Performance Conditions (CPCs) or PSFs are defined to describe the context. The nine PSFs are “adequacy of organization”, “working conditions”, “adequacy of man-machine interface and operational support”, “availability of procedures and plans”, “number of simultaneous goals”, “available time”, “time of day”, “adequacy of training and experience”, and “crew collaboration quality”. Each PSF has a number of determined states with the negative, positive or neutral effects on performance probability. For instance, for “Adequacy of training and experience”, the sets of the states and their effects are States = {Adequate with high experience, Adequate with limited experience, Inadequate} and Effect = {Positive, Neutral, Negative}.

According to the number of positive and negative effects of the PSFs and using the basic diagram of CREAM, the likely control mode of an operator is determined. CREAM uses Table 2 to reflect on how the effects of PSFs on human performance would change (from neutral to positive or negative) due to the dependencies among the PSFs [4]. For example, according to Table 2, the ratio (2/3) in the third row indicates that if at least two out of the three PSFs “Working conditions”, “Adequacy of MMI and operational support” and “Availability of procedure and plans” have negative effects, the neutral effect of “Number of simultaneous goals” changes to negative as well.

Table 2. Rules for adjusting the effects of PSFs in CREAM [4].

PSF	**The effect depends on the following PSFs**				
Working conditions (4/5)	Adequacy of organization	Adequacy of MMI and operational support	Available time	Time of day	Adequacy of training and experience
Number of simultaneous goals (2/3)	Working conditions	Adequacy of MMI and operational support	Availability of procedure and plans		
Available time (4/5)	Working conditions	Adequacy of MMI and operational support	Availability of procedure and plans	Number of simultaneous goals	Time of day
Crew collaboration quality (2/2)	Adequacy of organization	Adequacy of training and experience			

2.3. SLIM

SLIM is a flexible technique to estimate HEP during task execution [5]. It is a decision analysis approach in which the success likelihood index (SLI) of an error is calculated under the combined effects of the PSFs. A wide range of PSFs can be considered in the SLIM, enabling it to be used in different industries and contexts [29–31]. Although SLIM heavily relies on expert judgment, it could be quite practical where data on human error is insufficient. For a given task, the SLI is calculated by Eq. (3). The rate (R_i) shows the extent to which the PSF_i is desirable for executing the task while the weight (W_i) shows the relative importance of the PSF_i to the task.

$$SLI = \sum_{i=1}^{N} W_i R_i \tag{3}$$

To estimate the HEP in executing the task, the logarithmic relationship can be used to calibrate the SLI as:

$$Log(HEP) = aSLI + b \tag{4}$$

where the constant parameters a and b can be determined by two tasks for which the amounts of HEPs and the corresponding SLIs are already known using, for instance, historical data or expert judgment. In the conventional SLIM all the input parameters (the weights, rates, and the constants a and b) are determined by experts, introducing degrees of epistemic uncertainty into the analysis.

3. BN versions of HRA methods

3.1. Bayesian Network and Bayesian Parameter Learning

$BN = (G, \theta)$ is a graphical model for probabilistic inference. G is the graphical structure in which the nodes display the random variables X = {x_1, x_2, …, x_n}, and the directed arcs represent the dependencies among the random variables; θ is the set of network parameters presented as the conditional probability tables (CPTs) of the nodes [32]. BN satisfies the Markov condition in that the variables (nodes) in the graph are independent of their non-descendants given their parents. As such, the joint probability distribution of the random variables can be presented as the product of the conditional probabilities of the nodes given their immediate parents as:

$$P(X) = \prod_{i=1}^{n} P(x_i | Pa(x_i)) \quad (5)$$

where $Pa(x_i)$ is the parent set of nodex_i, and $P(x_i|Pa(x_i)) = \theta_i$ is the network parameter used to populate the CPT of node x_i. These parameters can be elicited from experts or be learned from data. Using the Bayes' theorem, BN is able to update the prior probabilities of the nodes by observing new evidence (E), as presented in Eq. (6). The main application of probability updating is in sensitivity analysis [33]. In the context of HRA, the evidence can be observation of human error in a task, an occurrence of incidents in an operation, or new information about the performance context.

$$P(X|E) = \frac{P(E|X)P(X)}{P(E)} = \frac{P(X,E)}{\sum_X P(X,E)} \quad (6)$$

The BN parameters can be estimated via parameter learning algorithms, e.g., the maximum likelihood estimation. Given a dataset $D = \{X^1, X^2, \ldots, X^m\}$ which contains complete observations of the states of the BN variables $X^j = \{x_1^j, x_2^j, \ldots, x_n^j\}$, the network parameters θ can be estimated by maximizing the likelihood or log-likelihood of the dataset as [28, 34]:

$$Log\_likelihood(D; G, \theta) = Log(P(D|\theta)) = Log \prod_{j=1}^{m} P(x_1^j, x_2^j, \ldots, x_n^j | \theta) =$$
$$Log \prod_{j=1}^{m} \prod_{i=1}^{n} P(x_i^j | Pa(x_i^j)) = Log \prod_{j=1}^{m} \prod_{i=1}^{n} \theta_i^j = \sum_j^m \sum_i^n Log\, \theta_i^j \quad (7)$$

3.2. BN-SPARH

Groth and Sliwer [8] proposed that using BN would make HRA models more compatible with the HRA practitioners' perspective. They illustrated how BN-SPARH can be useful for causal and evidential reasoning with perfect, partial or no information on the PSFs states. The main steps for developing the BN-SPARH can be summarized as:

Building the BN-SPARH structure: BN-SPARH has a simple structure with 9 nodes; eight nodes to represent the eight PSFs and one node to represent the HEP. The states of the PSF nodes are the same as the states defined in the conventional SPAR-H method [6]; however, the “Insufficient information” state is excluded because even in the absence of sufficient information (non-informative) prior probability distributions can still be assigned to the PSF nodes of the BN. The HEP node has two states: human error occurs (HEP = Yes) and human error does not occur (HEP = No). The causal arc between a PSF node and the HEP node illustrates the conditional dependence of the latter on the former.

Quantifying BN-SPARH: Using the predefined mathematical relationships given in Eqs. (1) and (2), the CPT of the HEP node can be populated. However, in case of “Available time = Inadequate” or “Fitness for duty = Unfit” the conditional HEP would be equal to 1 (i.e., we are certain that HEP = Yes). The probability mass function of the states of each PSF is identified using the available data and/or experts’ knowledge.

3.3. BN-CREAM

Kim et al. [16] developed the BN-CREAM so that the uncertainty associated with the states of the PSFs can be modeled using probability distributions. To better handle the uncertainties, Yang et al. [35] and Zou et al. [36] proposed fuzzy BN-CREAM, which are beyond the scope of the present study. The BN-CREAM can be developed through the following steps:

Determining the primary effect of each PSF: For each PSF, there is a node that represents the states of the PSF and is connected to another node for modeling the primary effect of the states of that PSF on the performance reliability. To demonstrate how to relate the states of a PSF to their effects, the CPT of node “Effect of crew collaboration quality” has been presented in Table 3.

Table 3. CPT of node “Effect of crew collaboration quality”.

Expected effect	**States**			
	Very efficient	Efficient	Inefficient	Deficient
Positive	1	0	0	0
Neutral	0	1	1	0
Negative	0	0	0	1

Adjusting the PSFs’ effects: Considering the dependencies among the four PSFs (Table 2), the adjusted effects of the PSFs are considered by assigning four specific nodes. The CPTs of these nodes are filled using the rule presented in Section 2.2. For the sake of clarity, Table 4 reports parts of the CPT of node "Adjusted crew collaboration quality".

Table 4. Parts of the CPT of node "Adjusted crew collaboration quality"

Crew collaboration quality	**Adequacy of organization**	**Adequacy of training and experience**	**Adjusted crew collaboration quality**		
			Positive	Neutral	Negative
Neutral	Positive	Positive	0	1	0
		Neutral	0	1	0
		Negative	0	1	0
	Neutral	Positive	0	1	0
		Neutral	0	1	0
		Negative	0	1	0
	Negative	Positive	0	1	0
		Neutral	0	1	0
		Negative	0	0	1

Determining the control mode: Given the effects of all the 9 PSFs, the CPT of node "control mode" can be determined by employing the rules defined in the conventional CREAM. Due to the massive size of the CPT of this node (size of $3^7 \times 2^2$), in some studies the nine PSFs are divided into 3 groups to reduce the calculation load [16, 36].

***Calculating HEP*:** Although the HEP estimation is not included in the BN-CREAM proposed by Kim et al. [16], adding the HEP node with the two states of "HEP = Yes" and "HEP = No" can facilitate the calculation of the HEP. The CPT of the HEP node can be filled in with the mean values of the HEP intervals.

Using the mean values of probability intervals is a common practice in probabilistic safety assessment [37] although some information may be lost using this approach. Another alternative would be using Dempster-Shafer theory to handle probability intervals [38], which could increase the accuracy of the calculated HEP yet at the expense of a more complicated analysis, which is beyond the scope of the present study.

3.4. BN-SLIM

Abrishami et al. [15] developed BN-SLIM and demonstrated that it outperforms the conventional SLIM by considering the probability distribution of PSFs, by considering the dependencies among the HEPs, and by identifying the critical PSFs and PSF rates using the probability updating feature of the BN. To develop the BN-SLIM the following steps should be taken:

Building the BN-SLIM structure: According to the conventional SLIM, the total effect of contributing PSFs on the HEP is modeled through the SLI variable. Thus, two functions are needed for estimating the HEP: One for calculating the SLI given a set of *N* PSFs, and the other for calculating the HEP given the SLI. Thus, a BN with *N* + 2 nodes would be required, *N* nodes for representing the PSFs and 2 nodes for representing the SLI and the HEP.

Each PSF node has several states to represent its rates. Thus, the number of the states of the SLI node is equal to the number of possible combinations of the rates (states) of the PSFs nodes. For example, consider a case with two PSFs, PSF1 and PSF2, each with two rates of 3 (indicating a poor state) and 7 (indicating a good state) and respective weights of 0.2 and 0.8. As a result, the SLI node would have four states as $SLI = 0.2 \times \{3, 7\} + 0.8 \times \{3, 7\} = \{3.0, 3.8, 6.2, 7.0\}$. The SLI node should be the only parent of the HEP node, which in turn would have two states, human error occurs (HEP = Yes) and human error does not occur (HEP = No).

BN-SLIM quantification: To quantify the effects of the PSFs nodes, CPTs should be assigned to the SLI and HEP nodes. The CPT of the SLI node shows which combination of the PSF rates would result in which state (value) of the SLI. To build the CPT of the HEP node, the conditional error probability is assigned via direct application of the logarithmic formula in Eq. (4). For example, P(HEP = Yes | SLI = 3.8) = $10^{-(3.8a+b)}$ where a and b are determined based on expert knowledge and/or available data.

4. Comparing the performance of BN-HRA models

4.1. Case study

In this study, we use the simulation data of human performance during offshore emergency evacuation generated in a virtual environment [27]. The dataset contains 129 observations with six binary variables. Each record contains three dependent variables associated with three PSFs and three independent variables associated with three possible responses of the test participants (each response is considered as a possible human failure). According to the designed experiment, “Training”, “Visibility”, and “Complexity” are selected as the three PSFs as in Table 5. The three executive tasks in the evacuation process are defined as “Evacuation”, “Backtracking” and “Exposure to hazard” [27]. The definitions of these tasks are presented in Table 6. If the time of “Evacuation” or “Backtracking” takes longer than a benchmark time, or if the “Exposure to hazard” leads to injury, a human failure is supposed to have occurred.

Table 5. Description of the PSFs [27].

PSF	Description	State
Visibility	It refers to the amount of ambient light available while performing a specific task. The amount of light is believed to affect the visibility of the evacuees and hence their performance.	High: performing a task in daytime Low: performing a task at night
Complexity	It refers to how difficult it is to perform the task in a given context. Complexity considers both the task and the environment in which the task is to be performed. The more difficult the task to perform the greater the likelihood of human error.	Low: if there is no hazard or obstacle on the available routes to the lifeboat station. High: if several routes are blocked with hazards such as jet fire, pool fire, and heavy smoke
Training	It refers to the type of training provided to the evacuees (participants in the virtual experiment).	Active: learning to navigate to the lifeboat platform by freely exploring the environment. Active - passive: learning to navigate to the lifeboat platform by watching three training videos hosted by an avatar who described a specific predetermined path. The participant can imitate the routes taken by the avatar after each video.

Table 6. Tasks description [27].

Task	Description
Evacuation	Time to evacuation refers to the time taken by the participant to reach the lifeboat platform from the starting position.
Backtracking	Backtracking time is the time spent by the participant to go back the way they had come. In an ideal case, the participant should not spend time in backtracking unless the route followed is blocked, in which case they might have to backtrack to find an alternative route.
Exposure to hazard	Depending on the type of hazard and time spent close enough to the hazard, the participant could be injured or not.

Tables 7 and 8 present the data-derived relative frequencies of the PSF states and the relative failure frequencies of the tasks. The relative failure frequency of each task has been considered as the objective HEP of that task in the present study.

Table 7. Data-derived relative frequencies of the states of PSFs [27].

Visibility		**Training**		**Complexity**	
State	Frequency	State	Frequency	State	Frequency
High	0.67	Active	0.51	Low	0.67
Low	0.33	Active-Passive	0.49	High	0.33

Table 8. Data-derived relative failure frequencies of the tasks [27].

Evacuation		**Backtracking**		**Exposure to hazard**	
State	Frequency	State	Frequency	State	Frequency
Time of evacuation < benchmark time (HEP = No)	0.37	Time of backtracking < benchmark time (HEP = No)	0.26	No exposure to hazard (HEP = No)	0.83
Time of evacuation > benchmark time (HEP = Yes)	0.63	Time of backtracking > benchmark time (HEP = Yes)	0.74	First or second-degree burn or death (HEP = Yes)	0.17

4.2. Applying BN-HRA models

In the present study, the BN-HRA models are categorized into three groups with regard to the role of data in calculating the conditional dependency of the HEP node on the PSF nodes. It should be noted that in all the three categories the prior probabilities of the root nodes (i.e., PSFs) are identified using the available data.

- **Rule-based models**: BN-SPARH and BN-CREAM estimate the HEP using the predefined rules given in the original SPAR-H and CREAM. For example, the probabilities to populate the CPTs of the BN-SPARH can be calculated using Eqs.(1) and (2) regardless of the available data. In other words, the CPT of the HEP node in a rule-based model remains the same for any task in a specific context since the available data does not play a role in quantifying the relationship between the PSFs and the HEP.
- **Data-based model**: It refers to a BN model in which the CPT of the HEP node given the PSFs are solely estimated based on the available data using parameter learning algorithms.
- **Hybrid model**: As is the case in the BN-SLIM, the relationship between the HEP node and the PSFs is given by Eqs. (3) and (4), i.e., the rule-based part of the modeling. The probability distribution of the rates and weights of the PSFs in Eq.(3) and the constant parameters in Eq.(4)

are determined based on the available data, i.e., the data-based part of modeling. This makes the BN-SLIM a semi-rule-based semi-data-based technique, or a hybrid technique.

The main features of the three categories are summarized in Table 9.

Table 9. Main features of rule-based, data-based, and hybrid BN-HRA methods in the present study.

Model	**Examples**	**Flexible set of PSFs?**	**Ability to calculate distinct HEPs?**	**How to populate CPTs?**
Rule-based	BN-SPARH; BN-CREAM	No	No	Using predefined rules; available data do not play a role
Data-based	BN	Yes	Yes	Using Bayesian parameter learning algorithms
Hybrid	BN-SLIM	Yes	Yes	Using predefined rules and available data

It should be noted that BN-SPARH has the potential to be upgraded to a hybrid model if the weights of its PSFs can be evaluated using the data and then be accommodated in the mathematical relationship between PSFs and HEP (i.e., Eqs. (1) and (2)). However, this topic is beyond the scope of the present study and can be investigated in a separate work. To evaluate the validity and accuracy of the foregoing models, the observed relative frequency of the HEP of each task, i.e., the objective HEP, is compared with the corresponding HEPs estimated by the BN-HRA methods.

4.2.1. Rule-based models: BN-SPARH and BN-CREAM

The PSFs defined in the dataset of Musharraf et al. [27] – herein, dataset PSFs – are different from the PSFs defined in the original SPAR-H and CREAM – herein, model PSFs. As such, the model PSFs which are the closest in meaning and context to the dataset PSFs should first be identified. For instance, “Visibility” (Table 5), which is a dataset PSF, has been related to “Work condition” and “Ergonomic”, which are the model PSFs in CREAM and SPAR-H, respectively.

The corresponding PSFs to “Training”, “Visibility” and “Complexity” are listed in Tables 10 and 11 for BN-CREAM and BN-SPARH, respectively [4, 6]. Using the data, the probabilities (relative frequencies) of the states of these three PSFs are calculated. However, due to the lack of simulation data about the rest of the PSFs, equal probabilities have been assigned to their states in both BN-SPARH and BN-CREAM.

Table 10. Probability distribution of the rates of the PSFs in BN-CREAM. Corresponding dataset PSFs are mentioned in the brackets.

PSF	State	Probability
Adequacy of training and experience (Training)	Inadequate	0
	Adequate with low experience	0.49
	Adequate with high experience	0.51
Working condition (Visibility)	Incompatible	0.33
	Compatible	0.67
	Advantageous	0
Number of simultaneous goals (Complexity)	Fewer than the actual capacity	0
	Matching current capacity	0.67
	More than the actual capacity	0.33

Table 11. Probability distribution of the rates of PSFs in BN-SPARH. Corresponding dataset PSFs are mentioned in the brackets.

PSF	State	Probability
Experience /Training	Low	0.00
	Nominal	0.49
	High	0.51
Ergonomic (Visibility)	Missing	0.00
	Poor	0.33
	Nominal	0.00
	Good	0.67
Complexity	Nominal	0.67
	Moderate	0.00
	High	0.33

It is worth noting that if the available information is not enough, the conventional SPAR-H considers the nominal states of the PSFs; it is also able to assign a probability distribution to the states [8], which is the case in the present study. The resulting BN-CREAM and BN-SPARH for the backtracking task are displayed in Figures 1 and 2, respectively. The models have been generated using AgenaRisk software [39]. Since the context of the three tasks is the same, and all the tasks are of action type, the BN-CREAM and BN-SPARH both result in the identical HEPs for all the three tasks. That is why the modeling has been performed only for “Backtracking”.

It should be noted that both SPAR-H and CREAM (and their BN versions) are built on the predefined sets of PSFs which cannot be changed regardless of their relevance to the context of interest. Therefore, if some PSFs are eliminated, the defined rules in CREAM and SPAR-H become futile. The BN-SPARH and BN-CREAM also inherit this limitation in which all the predefined PSFs, whether relevant or irrelevant to the dataset, would be required to calculate the CPTs of the models.

One way to minimize the impact of irrelevant PSFs on the calculated HEP is to keep all the model PSFs but assign equal probabilities to the states of the PSFs which are deemed irrelevant to the dataset PSFs. This modeling technique is expected to reduce the impact of irrelevant PSFs because equal state probabilities of a PSF node would result in the minimum amount of mutual information between the PSF node and the HEP node [40].

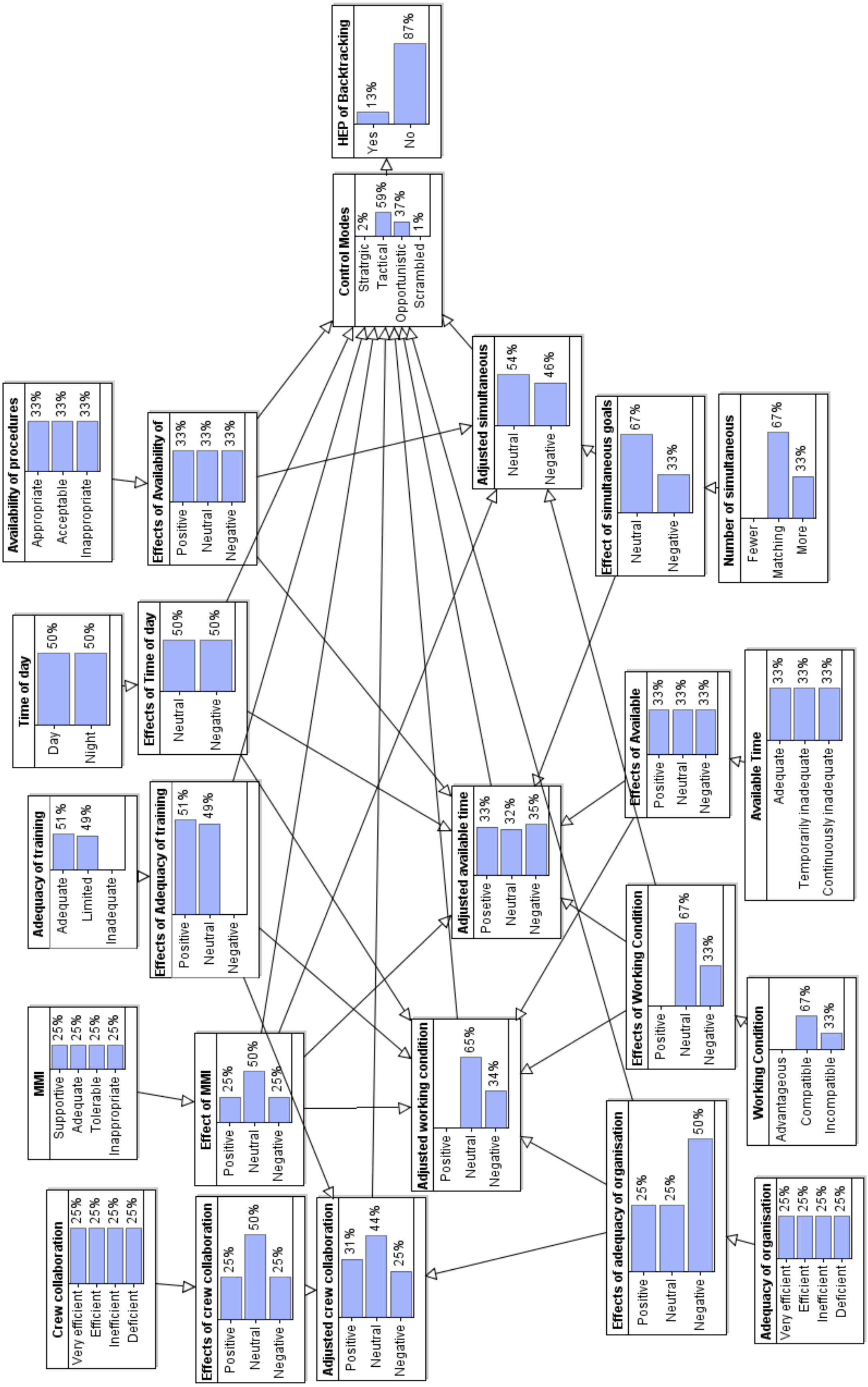

Figure 1. BN-CREAM model for predicting the HEP of "Backtracking". The HEPs of "Evacuation" and "Exposure to hazard" would be the same.

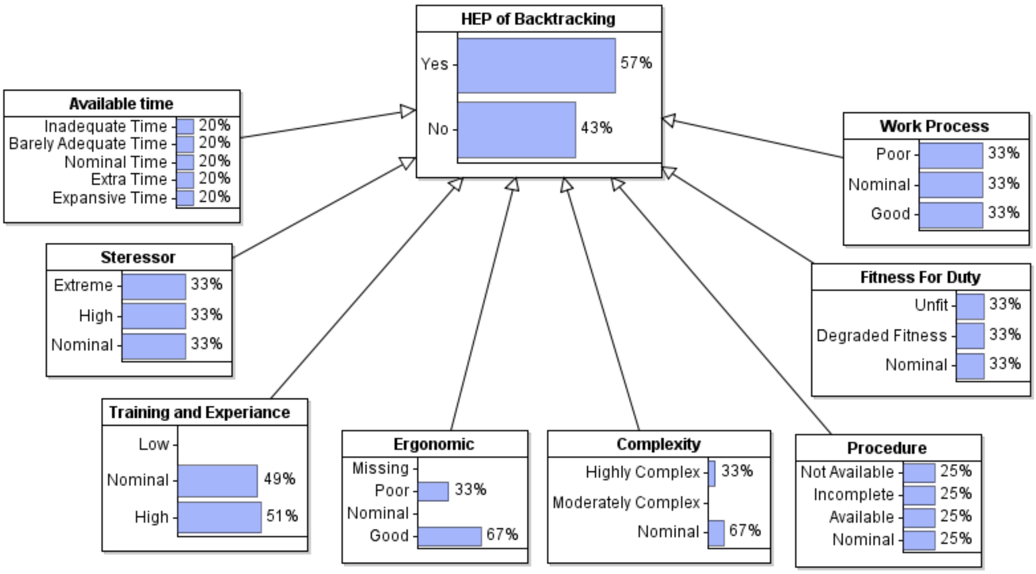

Figure 2. BN-SPARH model for predicting the HEP of “Backtracking”. The HEPs of “Evacuation” and “Exposure to hazard” would be the same.

4.2.2. Hybrid model: BN-SLIM

For building the BN-SLIM in the present study, we used the simulation data to calculate the probability of the rates of the PSFs, the weights of the PSFs with respect to each task, and also the parameters *a* and *b* in Eq (4). Due to the binary nature of the variables in the simulation data, two rates of 3 and 7 are considered as the worst and the best states of the PSFs. Table 12 presents the data-derived probabilities (relative frequencies) of the rates of the PSFs. To measure the strength of the causal relationship between a PSF and a task failure, Jaccard coefficient [41] in Eq. (8) can be used:

Table 12. Probability distribution of the PSFs rates in BN-SLIM.

PSF	**Rate**	**Probability**
Training	7	0.51
	3	0.49
Visibility	7	0.67
	3	0.33
Complexity	7	0.67
	3	0.33

$$J(y,z) = \frac{e+h}{e+f+g+h} \tag{8}$$

where for the binary variables *y* (e.g., a PSF) and *z* (e.g., the task), *e* represents the number of observations where *y* and *z* are equal to 1; *f* represents the number of observations where *y* is 0 and *z* is 1; *g* represents the number of observations where *y* is 1 and *z* is 0; *h* represents the number of observations where both *y* and *z* are 0. The calculated Jaccard coefficient and the normalized weights of the PSFs are listed in Table 13.

Table 13. Jaccard coefficient and normalized weights of the PSFs derived from the data.

PSFs	**Jaccard coefficient**			**Normalized weight**		
	Evacuation	Backtracking	Exposure to hazard	Evacuation	Backtracking	Exposure to hazard
Training	0.55	0.42	0.58	0.34	0.33	0.30
Visibility	0.53	0.35	0.52	0.32	0.28	0.27
Complexity	0.55	0.49	0.84	0.34	0.39	0.43

The two constant parameters in Eq. (4) are calculated considering the highest and the lowest SLI values and their corresponding HEP (frequency) for each task. The SLI values and their corresponding HEPs are presented in Table 14. Due to no observed error for the “Exposure to a hazard” in the dataset, the lowest HEP of this task is assumed to be as 1.0 E-06. Unlike the BN-SPARH and BN-CREAM, the BN-SLIM does not result in the same HEPs for all the tasks as, despite the same PSFs, the weights of the PSFs differ from task to task. The developed BN-SLIM is depicted in Figure 3.

Table 14. The Lowest and highest SLI values and their corresponding relative error frequencies (objective HEPs) estimated directly from the simulation data.

SLI values	**Relative error frequencies**		
	Evacuation	Backtracking	Exposure to hazard
7	0.55	0.59	1.0 E -06
4.30	0.91	-	-
4.11	-	0.95	-
4.07	-	-	0.67

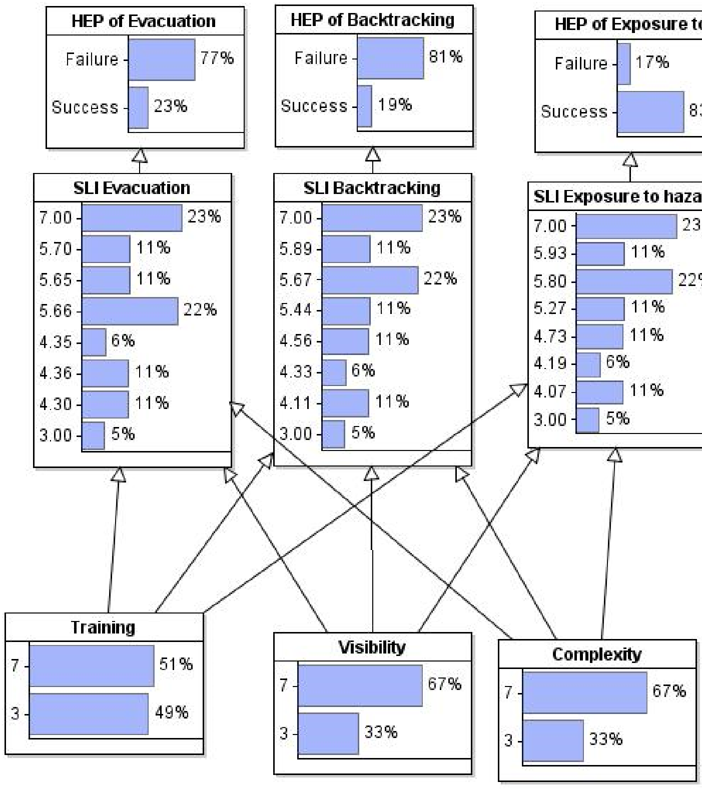

Figure3. BN-SLIM model for predicting the HEP of “Backtracking”, “Evacuation” and “Exposure to hazard”.

4.2.3. Data-based model: Bayesian parameter learning

To develop the data-based model for estimating the HEPs, the structure of the BN (Figure 4) is built with six nodes associated with the three PSFs and the three tasks. Having the structure of the BN determined, the network’s conditional probabilities can be calculated from the dataset using the parameter learning algorithms embedded in AgenaRisk software [39].

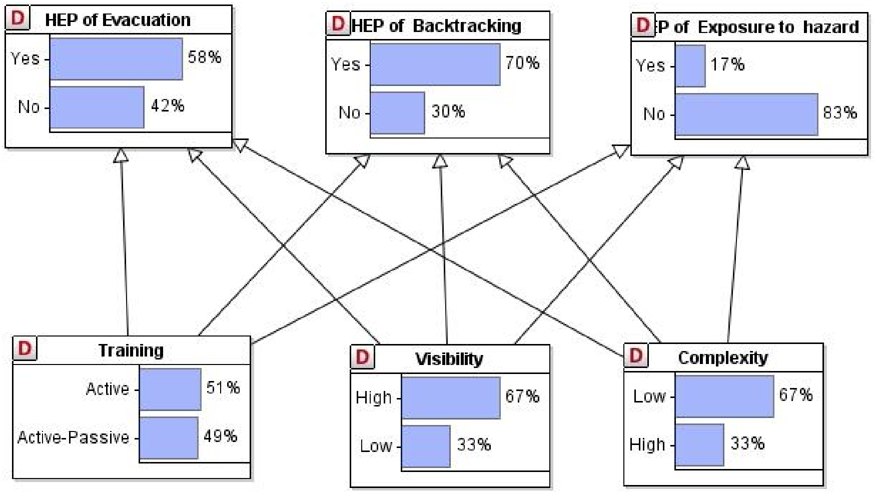

Figure 4. Developed BN via the learning parameter algorithm (BPL model).

4.3. Results

To evaluate the validity and accuracy of the models in the present study, in Figures 5-7 the HEPs estimated by the models are compared with the corresponding objective HEPs (data-derived relative error frequencies).

As can be seen in Figure 5, the BPL model and BN-SPARH predict the HEP of “Evacuation” as 0.58 and 0.57, respectively, which are close to the objective HEP of 0.63. The BN-SLIM with the HEP of 0.77 seems to have slightly overestimated the HEP of “Evacuation” while the HEP of 0.13 estimated by the BN-CREAM is too far from the objective HEP. As can be seen in Figure 6, with an objective HEP of 0.74 for the “Backtracking”, the BPL model provides a relatively more accurate estimation (HEP = 0.7) than the BN-SLIM (HEP = 0.81). However, the estimations of the BN-SPARH (HEP = 0.57) and BN-CREAM (HEP = 0.13) remarkably differ from the objective HEP.

As illustrated in Figure 7, with the objective HEP of 0.18 for “Exposure to hazard”, the BPL model and the BN-SLIM both result in a very close HEP of 0.17. The BN-CREAM results in the most accurate HEP (0.13) for this task than the other two tasks, while there is a huge gap between the result of the BN-SPARH (HEP = 0.57) and the objective HEP of 0.18 for this task.

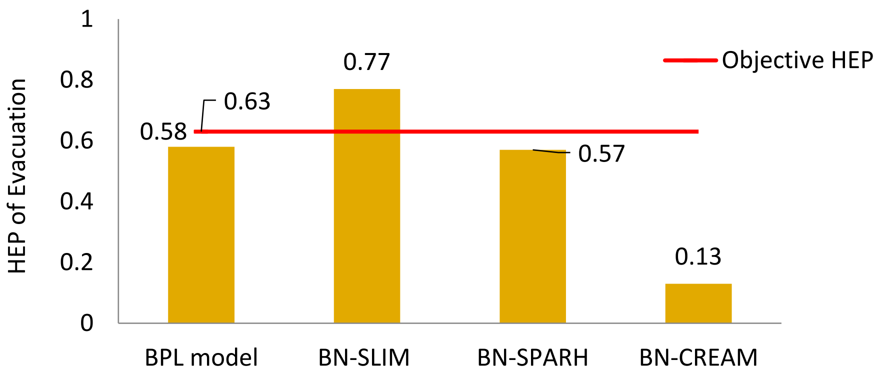

Figure 5. Comparison between the model HEPs and the objective HEP for “Evacuation”.

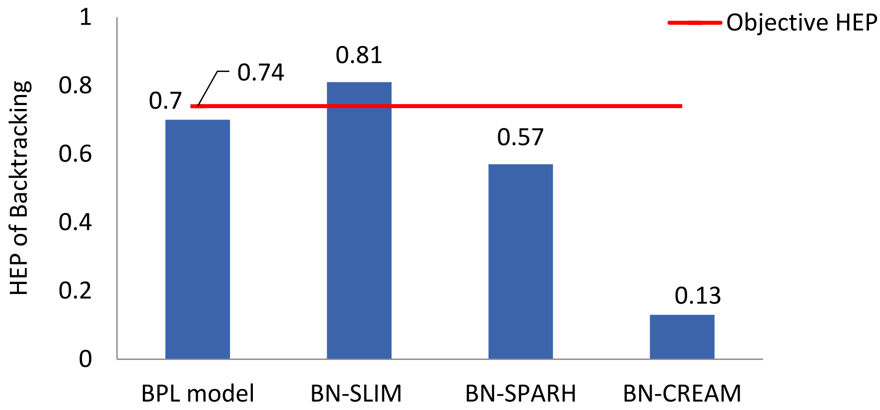

Figure 6. Comparison between the model HEPs and the objective HEP for “Backtracking”.

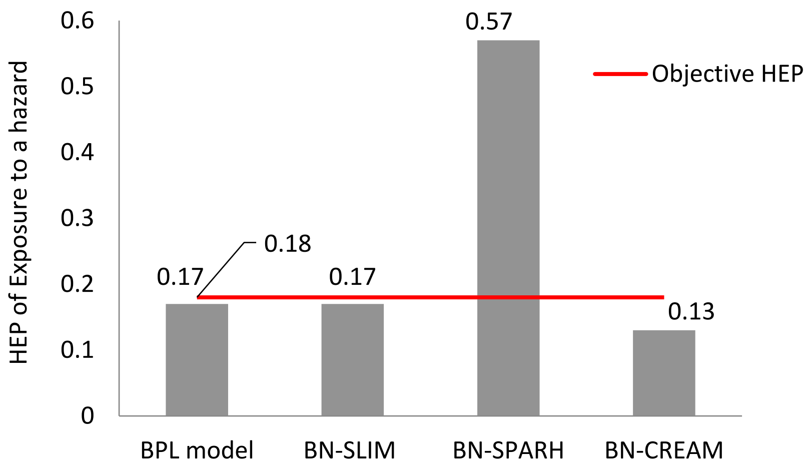

Figure 7. Comparison between the model HEPs and the objective HEP for “Exposure to hazard”.

To make a better view of the models' accuracy and validity, we have introduced the Overall Performance Accuracy (OPA) as a performance indicator of the models by measuring the Euclidean distance between the model HEPs and the objective HEPs. Considering the foregoing three tasks, the distance between the objective $HEP = (HEP_1, HEP_2, HEP_3)$ and the model $\widehat{HEP} = \left(\widehat{HEP}_1, \widehat{HEP}_2, \widehat{HEP}_3\right)$ can be calculated for each BN-HRA model as:

$$OPA_{model} = \sqrt{\sum_{i=1}^{3}(HEP_i - \widehat{HEP}_i)^2} \tag{9}$$

where i = 1, 2, 3 denotes the three tasks of “Evacuation”, “Backtracking”, and “Exposure to hazard”. A lower value of OPA represents a more accurate model estimation. For instance, using the number in Figures 5-7, the OPA of the BN-SLIM can be calculated as:

$$OPA_{BN-SLIM} = \sqrt{\underbrace{(0.63 - 0.77)^2}_{Evacuation} + \underbrace{(0.74 - 0.81)^2}_{Backtracking} + \underbrace{(0.18 - 0.17)^2}_{Exposure\ to\ hazard}} = 0.157$$

The OPAs of the models are presented in Table 15. The comparison between the OPA values shows that BPL model with an OPA of 0.065 has a better performance in predicting the HEPs than other BN-HRA models. The BN-SLIM stands in the second place which would demonstrate the higher performance of the data-based models in general (BPL model, and to a lesser degree the BN-SLIM) in estimating the HEPs.

Table 15. Comparing the models performance based on their OPA.

BN-HRA models	BPL model	BN-SLIM	BN-SPARH	BN-CREAM
OPA	0.065	0.157	0.430	0.790

4.4. Evaluation of models’ generalizability

Although the accuracy of the BPL model, given a sufficiently large dataset, is better than the other BN-HRA models, it is important to evaluate the models accuracy in a more practical condition where the models need to be extended to cases with no or insufficient data.

Cross-validation is a technique used for evaluating the performance of machine learning models. The goal of cross-validation is to test the model's ability in predicting data that was not used in the development of the model so that problems like overfitting [42] can be marked. It also helps gain insight into how reliably the model could be generalized to an independent dataset. K-fold is a popular cross-validation technique when there is limited input data [43]. For example, if 4-fold cross-validation is used, the data set is split into four subsets of equal size; then in each iteration, the model is trained on the three data subsets (train folds) and tested on the remaining fourth subset (test fold) (Figure 8). Repeating this operation for all the subsets, the averaged result may give an estimate of the model's predictive performance.

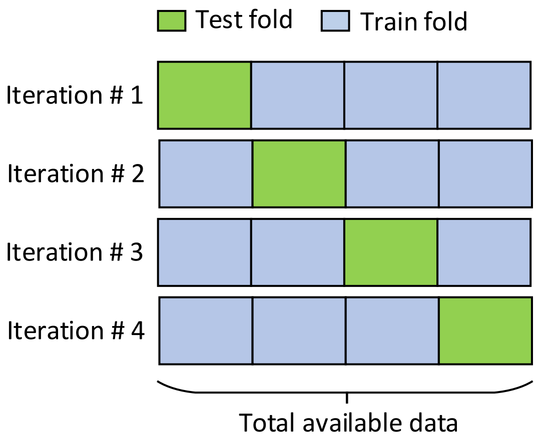

Figure 8. Four-fold cross-validation.

In the present study, we use the four-fold cross-validation to assess the generalizability of the models. For this purpose, the train and test errors in each iteration can be calculated for a task as:

$$E_j^{TR} = |\widehat{HEP}_j^{TR} - HEP_j^{TR}| \tag{10}$$

$$E_j^{TE} = |\widehat{HEP}_j^{TE} - HEP_j^{TE}| \tag{11}$$

when E_j^{TR} and E_j^{TE} are the train error and the test error of the j-th iteration (given a 4-fold validation, j = 1, 2, 3, 4), respectively. For a given task, $\widehat{HEP}^{TR}$ and $\widehat{HEP}^{TE}$ are the model HEPs of the train and test datasets, respectively, while HEP^{TR} and HEP^{TE} are the relative human error frequencies (objective HEPs) calculated using the train and the test datasets, respectively. So, after four iterations, four pairs of train and test errors are calculated, and the average train error (E^{TR}) and the average test error (E^{TE}) of a model are calculated as:

$$E^{TR} = \frac{\sum_{j=1}^{4} E_j^{TR}}{4} \tag{12}$$

$$E^{TE} = \frac{\sum_{j=1}^{4} E_j^{TE}}{4} \tag{13}$$

Train error is used to identify the extent to which a model fits the train dataset, while the test error is used to ensure that the model is not overfitting [44]. In other words, a large train error illustrates that the model is underfitting and thus unable to predict the HEP accurately. Nevertheless, a small train error may not guarantee the model accuracy unless there is a small difference between the test and the train errors.

It should be noted that the CPTs of the BN-SPARH and the BN-CREAM are constants in all the iterations as these two models are rule-based, and their CPTs are thus defined based on predefined rules not the train or test data. However, the probabilities of the PSFs, as the root nodes of the BN models, would change in each iteration.

To obtain a better insight into the models' accuracy, the test and train errors of the models for the three tasks are depicted in Figures 9-11. As can be seen in Figure 9, for the "Evacuation", the BN-CREAM has the highest train error (0.48) and thus the lowest accuracy among the models. (It is worth noting that since the train error of the BN-CREAM is already large, there is no point in considering its test error). The large differences between the train and the test errors of the BPL model and the BN-SPARH indicate that these models are susceptible to overfitting (i.e., a small train error but a large test error). On the other hand, the BN-SLIM has a small train error (0.09), and there is a small difference between its train and test errors, ruling out the possibility of overfitting. This shows a better performance of the BN-SLIM in predicting the HEP of "Evacuation" compared to the other models.

Considering the HEP of the "Backtracking", Figure 10 illustrates that the BN-CREAM may not be an accurate model since it has the highest train error (0.56) among the models. There is a notable difference between the train and test errors of the BPL model while the difference between the train and test errors of both the BN-SPARH and the BN-SLIM is negligible. This may imply the BN-SPARH and BN-SLIM are more accurate than the BPL model. Furthermore, the smaller train error of the BN-SLIM (0.1) indicates that it is more accurate than the BN-SPARH in estimating the HEP of "Backtracking".

Considering the "Exposure to hazard", as can be seen in Figure 11, there are no noticeable differences between the train and the test errors of the models. The train error of the BN-SPARH is the highest (0.31) and that of the BN-SLIM is the lowest (0.01), indicating that BN-SLIM is able to calculate the HEP of this task more accurately than the other models.

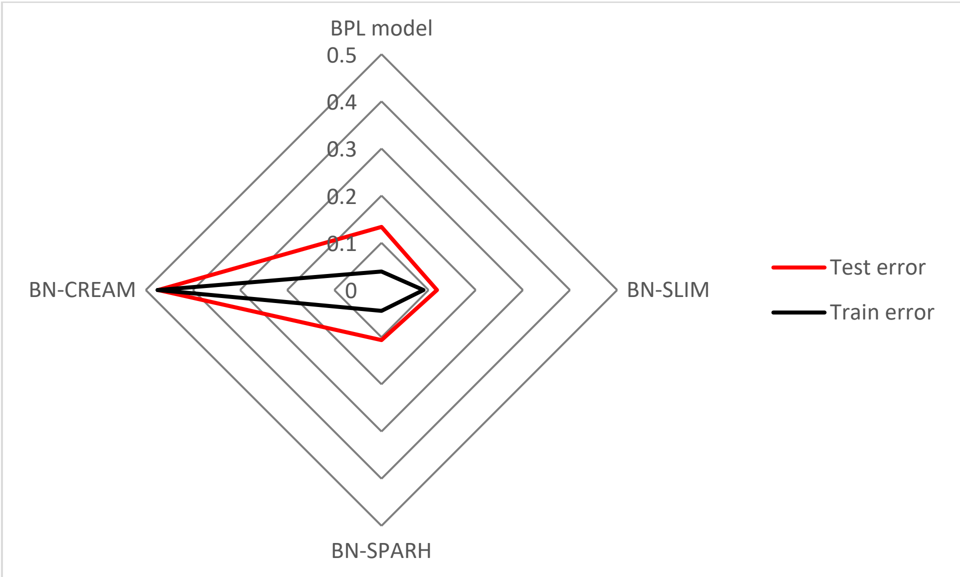

Figure 9. Test and train errors of the BN-HRA models for the “Evacuation”.

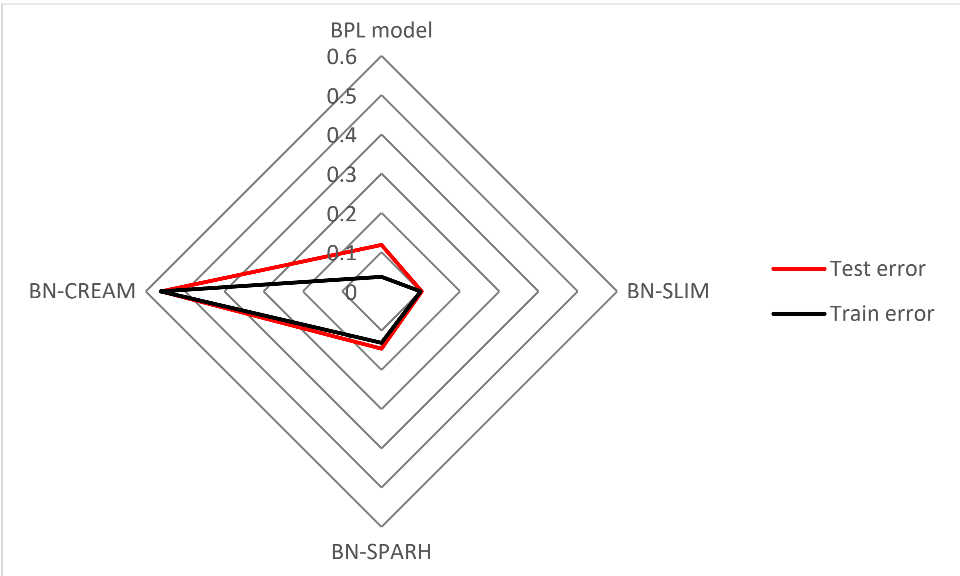

Figure 10. Test and train errors of the BN-HRA models for the “Backtracking”.

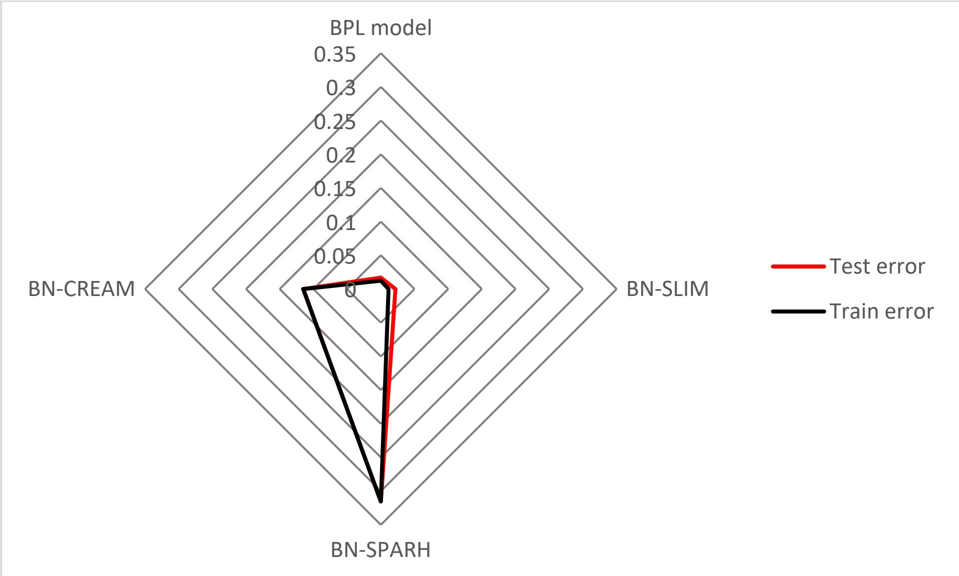

Figure 11. Test and train errors of the BN-HRA models for the “Exposure to a hazard”.

To identify a model with the best performance with regard to all the three tasks, the OPAs of each model for both the train and the test datasets are computed. The train OPA of a model measures the Euclidean distance between the average HEPs estimated by the model using the train dataset and the average objective HEPs derived from the same train dataset. The test OPA can be calculated in the same way yet using the test dataset instead of the train datasets. By comparing the OPAs of the models and also by comparing the train and test OPAs of a single model, an analyst may get some idea about the performance of the models. For instance, between two models:

- the model with a smaller train OPA generally outperforms the one with a larger train OPA. In other words, the former model better fits the data whereas the latter model relatively underfits the data.
- the model with a smaller difference between its train and test OPAs is preferred over the model with a larger difference. This is because a model with a small train OPA and a large test OPA (i.e., a larger difference between its train and test OPAs) may suffer from overfitting.

As can be seen in Figure 12, the train OPAs of the BN-CREAM (0.74) and the BN-SPARH (0.34) are higher than the train OPAs of the other two models, indicating that the BN-CREAM and the BN-SPARH are not sufficiently accurate for estimating the HEPs using the train data (let alone using the test data which is one-fourth the size of the train data.) The least amount of train OPA for the BPL model may give the impression that it is the most accurate model given a sufficiently large dataset. However, the large difference between its train and test OPAs shows that it is overfitting the train data.

Figure 12 depicts that the BN-SLIM has relatively a small train OPA (0.13), and there is no considerable difference between its train and test OPAs, indicating a generally better performance of the BN-SLIM. Therefore, considering the performance of the models with regard to the individual tasks (Figures 9-11) and the three tasks altogether (Figure 12), the BN-SLIM can be identified as the model with the best performance.

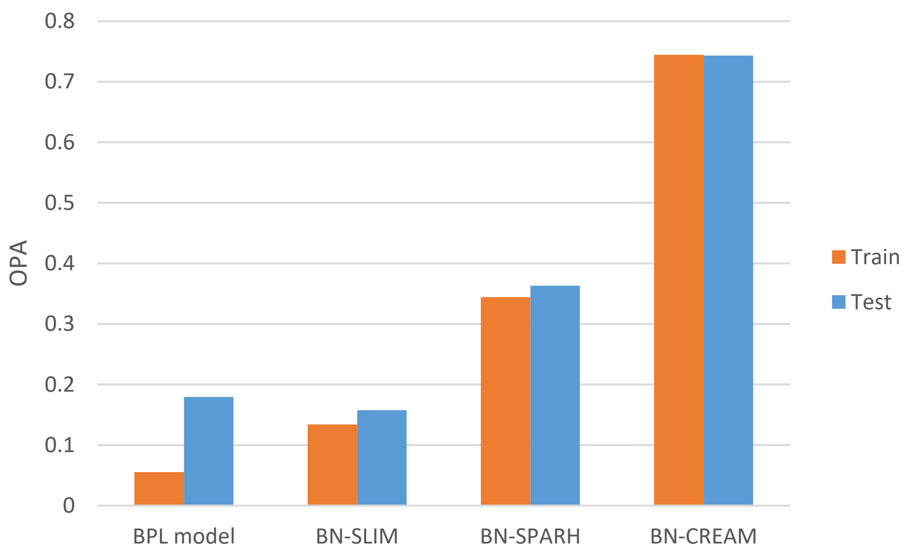

Figure 12. Models' OPAs calculated using the train and test data. The BN-CREAM and BN-SPARH have the highest train and test OPAs, indicating their lower performance in estimating the HEP. The BPL model has the lowest train OPA, but the notable difference between its train and test OPAs may imply overfitting. The BN-SLIM has relatively low train and test OPAs, and the slight difference between its train and test OPAs indicates its better performance than the BPL model.

4.5. Final remarks

As discussed before, the predetermined sets of PSFs in the BN-CREAM and the BN-SPARH may include some PSFs irrelevant to the context or dataset of interest. To reduce the impact of irrelevant (or redundant) PSFs on the estimated HEP, in Section 4.2.1 we assigned equal probabilities to the states of such PSFs. However, the inclusion of irrelevant PSFs may to some extent affect the accuracy of the HEPs estimated by the BN-SPARH and BN-CREAM. To illustrate this better, we added a redundant PSF

– the “Available time” – with equal state probabilities as P(rate =7, rate =3) = (0.5, 0.5) to the BN-SLIM[1] which resulted in the OPA of the BN-SLIM to increase from 0.157 to 0.373. This experiment may further demonstrate the advantage of the BN-SLIM and the BPL model as the choice of PSFs are more intuitive in these two models (compared to the forced PSFs in the BN-CREAM and BN-SPARH) in accordance with the context of interest.

Furthermore, the BN-CREAM and the BN-SPARH, unlike the BN-SLIM and the BPL model, are not able to differentiate among the HEPs of the tasks within the same context, resulting in the same HEPs for all the tasks. This limitation could result in an overestimation or underestimation of the total HEP depending on whether the tasks are performed sequentially or simultaneously. The BN-SLIM would have also resulted in the same HEPs had it not been able to assign different weights to the PSFs for different tasks.

The foregoing restrictions, i.e., being developed on predefined and unchangeable sets of PSFs and being incapable of considering different weights for the PSFs in different tasks, are in our perspective two of the main reasons for the lower performance of the BN-SPARH and the BN-CREAM in the present study. Nevertheless, before a verdict can be announced on the performance of the BN-HRA methods, further research must be carried out using data of different size and context, especially with the development of data collection systems such as SACADA [45] and HERA [46], and under different assumptions and model modifications.

5. Conclusions

In the present study we compared the performance of some selected BN-HRA models using the simulation data of human performance generated in an offshore evacuation virtual experiment. Considering the role of data in establishing the causal links between the PSFs and the HEP, three types of BN-HRA methods were investigated: (i) the rule-based methods of BN-CREAM and BN-SPARH, (ii) the data-based method of Bayesian parameter learning (BPL model), and (iii) the semi-rule-based (or semi-data-based) method of BN-SLIM. The BN-CREAM, the BN-SPARH and to some extent the BN-SLIM use fixed rules (mathematical relationships) to estimate the HEP from the PSFs. The BPL model, on the other hand, relies solely on the available data to derive the correlation between the PSFs and the HEP without any restrictive presumptions.

The comparison of the models' overall performance illustrated that data-based methods – the BPL model and the BN-SLIM – are more accurate than the rule-based methods. Furthermore, the k-fold

[1] Note that neither the BN-SLIM nor the BPL model forces the analyst to use a predefined set of PSFs, and can consider only the PSFs which are deemed relevant to the context.

validation of the methods demonstrated that the BN-SLIM may outperform the BPL model particularly in the absence of complete and sufficiently large databases, which is usually the case. (BPL model is more data sensitive than the BN-SLIM and is thus less accurate under data scarcity).

However, it should be noted that the performance of the BN-HRA methods in the present study was compared using a limited dataset and under assumptions and model adjustments. Such assumptions and model modifications (e.g., the selection of PSFs, the use of mean values instead of the probability intervals) were necessary to make the BN-HRA methods applicable to the dataset. Therefore, the performance of the customized BN-HRA methods employed in the current study may not exactly reflect the performance of the original BN-HRA methods. That being said, the outcomes of the present study cannot fully be extended to other contexts and domains unless further studies are conducted using different datasts and assumptions.

References

[1] L. Högberg, "Root causes and impacts of severe accidents at large nuclear power plants", AMBIO 42, 267–284 (2013). https://doi.org/10.1007/s13280-013-0382-x.

[2] G. Simpson, T. Horberry, and T. Horberry, Understanding Human Error in Mine Safety. CRC Press, 2018.

[3] R. P. E. Gordon, “The contribution of human factors to accidents and near misses in the offshore oil and gas industry : development of a human factors investigation tool,” PhD thesis, University of Aberdeen, 2002.

[4] E. Hollnagel, Cognitive Reliability and Error Analysis Method (CREAM). Elsevier, 1998.

[5] D. Embrey, P. Humphreys, E. Rosa, B. Kirwan, and K. Rea, “SLIM-MAUD: an approach to assessing human error probabilities using structured expert judgment. Volume II. Detailed analysis of the technical issues,” Brookhaven National Lab., 1984.

[6] D. Gertman, H. Blackman, J. Marble, J. Byers, and C. Smith, “The SPAR-H human reliability analysis method,” US Nuclear Regulatory Commission, 2005.

[7] K. M. Groth and A. Mosleh, “A data-informed PIF hierarchy for model-based Human Reliability Analysis,” Reliability Engineering & System Safety, vol. 108, pp. 154–174, Dec. 2012, doi: 10.1016/j.ress.2012.08.006.

[8] K. M. Groth and L. P. Swiler, “Bridging the gap between HRA research and HRA practice: A Bayesian network version of SPAR-H,” Reliability Engineering & System Safety, vol. 115, pp. 33–42, 2013, doi: 10.1016/j.ress.2013.02.015.

[9] J. Park, Y. Kim, and W. Jung, “Calculating nominal human error probabilities from the operation experience of domestic nuclear power plants,” Reliability Engineering & System Safety, vol. 170, pp. 215–225, Feb. 2018, doi: 10.1016/j.ress.2017.10.011.

[10] L. Mkrtchyan, L. Podofillini, and V. N. Dang, “Bayesian belief networks for human reliability analysis: A review of applications and gaps,” Reliability Engineering & System Safety, vol. 139, pp. 1–16, 2015, doi: 10.1016/j.ress.2015.02.006.

[11] L. Podofillini and V. N. Dang, “A Bayesian approach to treat expert-elicited probabilities in human reliability analysis model construction,” Reliability Engineering & System Safety, vol. 117, pp. 52–64, 2013.

[12] K. M. Groth and A. Mosleh, “Deriving causal Bayesian networks from human reliability analysis data: A methodology and example model,” Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, vol. 226, no. 4, pp. 361–379, 2012.

[13] N. J. Ekanem, A. Mosleh, and S.-H. Shen, "Phoenix–a model-based human reliability analysis methodology: qualitative analysis procedure," Reliability Engineering & System Safety, vol. 145, pp. 301–315, 2016.
[14] K. M. Groth, R. Smith, and R. Moradi, "A hybrid algorithm for developing third generation HRA methods using simulator data, causal models, and cognitive science," Reliability Engineering & System Safety, p. 106507, Jun. 2019, doi: 10.1016/j.ress.2019.106507.
[15] S. Abrishami, N. Khakzad, S. M. Hosseini, and P. van Gelder, "BN-SLIM: A Bayesian Network methodology for human reliability assessment based on Success Likelihood Index Method (SLIM)," Reliability Engineering & System Safety, vol. 193, p. 106647, Jan. 2020, doi: 10.1016/j.ress.2019.106647.
[16] M. C. Kim, P. H. Seong, and E. Hollnagel, "A probabilistic approach for determining the control mode in CREAM," Reliability Engineering & System Safety, vol. 91, no. 2, pp. 191–199, 2006, doi: 10.1016/j.ress.2004.12.003.
[17] E. Calixto, "Chapter 5 - Human Reliability Analysis," in Gas and Oil Reliability Engineering (Second Edition), E. Calixto, Ed. Boston: Gulf Professional Publishing, 2016, pp. 471–552.
[18] N. Khakzad, F. Khan, and P. Amyotte, "Safety analysis in process facilities: Comparison of fault tree and Bayesian network approaches," Reliability Engineering & System Safety, vol. 96, no. 8, pp. 925–932, Aug. 2011, doi: 10.1016/j.ress.2011.03.012.
[19] P. Trucco, E. Cagno, F. Ruggeri, and O. Grande, "A Bayesian Belief Network modelling of organisational factors in risk analysis: A case study in maritime transportation," Reliability Engineering & System Safety, vol. 93, no. 6, pp. 845–856, 2008, doi: 10.1016/j.ress.2007.03.035.
[20] Z. Mohaghegh and A. Mosleh, "Incorporating organizational factors into probabilistic risk assessment of complex socio-technical systems: Principles and theoretical foundations," Safety Science, vol. 47, no. 8, pp. 1139–1158, 2009, doi: 10.1016/j.ssci.2008.12.008.
[21] B. Wu, X. Yan, Y. Wang, and C. G. Soares, "An Evidential Reasoning-Based CREAM to Human Reliability Analysis in Maritime Accident Process," Risk analysis, vol. 37, no. 10, pp. 1936–1957, 2017.
[22] M. Giardina, P. Buffa, V. Dang, S. F. Greco, L. Podofillini, and G. Prete, "Early-design improvement of human reliability in an experimental facility: A combined approach and application on SPES," Safety Science, vol. 119, pp. 300–314, Nov. 2019, doi: 10.1016/j.ssci.2018.08.008.
[23] E. Akyuz, "Quantitative human error assessment during abandon ship procedures in maritime transportation," Ocean engineering, vol. 120, pp. 21–29, 2016.
[24] M. Bevilacqua and F. E. Ciarapica, "Human factor risk management in the process industry: A case study," Reliability Engineering & System Safety, vol. 169, pp. 149–159, Jan. 2018, doi: 10.1016/j.ress.2017.08.013.
[25] R. Boring, J. Forester, A. Bye, V. Dang, E. Lois. Lessons learned on benchmarking from the international human reliability analysis empirical study. Proceedings of the 10th International Probabilistic Safety Assessment and Management Conference, Seattle, Washington, USA, 2010.
[26] J. Forester, V. Dang, A. Bye, E. Lois, et al. The International HRA Empirical Study: Lessons learned from comparing HRA methods predictions to HAMMLAB simulator data. NUREG-2127, Office of Nuclear Regulatory Research, Washington, DC, August 2014. Available from: www.nrc.gov/docs/ML1422/ML14227A197.pdf.
[27] M. Musharraf, D. Bradbury-Squires, F. Khan, B. Veitch, S. MacKinnon, and S. Imtiaz, "A virtual experimental technique for data collection for a Bayesian network approach to human reliability analysis," Reliability Engineering & System Safety, vol. 132, no. Supplement C, pp. 1–8, 2014, doi: 10.1016/j.ress.2014.06.016.
[28] R. Sundaramurthi and C. Smidts, "Human reliability modeling for the next generation system code," Annals of Nuclear Energy, vol. 52, pp. 137–156, 2013.
[29] A. Noroozi, N. Khakzad, F. Khan, S. MacKinnon, and R. Abbassi, "The role of human error in risk analysis: Application to pre-and post-maintenance procedures of process facilities," Reliability Engineering & System Safety, vol. 119, pp. 251–258, 2013.

[30] R. Islam, F. Khan, R. Abbassi, and V. Garaniya, “Human Error Probability Assessment During Maintenance Activities of Marine Systems,” Safety and Health at Work, vol. 9, no. 1, pp. 42–52, Mar. 2018, doi: 10.1016/j.shaw.2017.06.008.

[31] M. Aalipour, Y. Z. Ayele, and A. Barabadi, “Human reliability assessment (HRA) in maintenance of production process: a case study,” International Journal of System Assurance Engineering and Management, vol. 7, no. 2, pp. 229–238, 2016.

[32] J. Pearl, “Fusion, propagation, and structuring in belief networks,” Artificial intelligence, vol. 29, no. 3, pp. 241–288, 1986.

[33] N. Khakzad, “(mis)Using Bayesian networks for dynamic risk assessment,” in Methods in Chemical Process Safety, Elsevier, 2020. https://doi.org/10.1016/bs.mcps.2020.03.001.

[34] N. Khakzad and P. Van Gelder, “Vulnerability of industrial plants to flood-induced natechs: A Bayesian network approach,” Reliability Engineering & System Safety, vol. 169, pp. 403–411, Jan. 2018, doi: 10.1016/j.ress.2017.09.016.

[35] Z. L. Yang, S. Bonsall, A. Wall, J. Wang, and M. Usman, “A modified CREAM to human reliability quantification in marine engineering,” Ocean Engineering, vol. 58, no. Supplement C, pp. 293–303, Jan. 2013, doi: 10.1016/j.oceaneng.2012.11.003.

[36] Q. Zhou, Y. D. Wong, H. S. Loh, and K. F. Yuen, “A fuzzy and Bayesian network CREAM model for human reliability analysis – The case of tanker shipping,” Safety Science, vol. 105, pp. 149–157, Jun. 2018, doi: 10.1016/j.ssci.2018.02.011.

[37] G. Zhang, V. V. Thai, K. F. Yuen, H. S. Loh, and Q. Zhou, “Addressing the epistemic uncertainty in maritime accidents modelling using Bayesian network with interval probabilities,” Safety Science, vol. 102, pp. 211–225, Feb. 2018, doi: 10.1016/j.ssci.2017.10.016.

[38] N. Khakzad, “System safety assessment under epistemic uncertainty: Using imprecise probabilities in Bayesian network,” Safety Science, vol. 116, pp. 149–160, Jul. 2019, doi: 10.1016/j.ssci.2019.03.008.

[39] AgenaRisk, version 10. 2019. Available from: https://www.agenarisk.com/

[40] T. Cover and J. Thomas, The Elements of Information Theory, 2nd ed. New Jersey: John Wiley & Sons, 2006.

[41] P. Jaccard, “The Distribution of the Flora in the Alpine Zone.1,” New Phytologist, vol. 11, no. 2, pp. 37–50, 1912, doi: 10.1111/j.1469-8137.1912.tb05611.x.

[42] G. C. Cawley and N. L. C. Talbot, “On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation,” Journal of Machine Learning Research, vol. 11, no. Jul, pp. 2079–2107, 2010.

[43] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning. New York: Springer, 2013.

[44] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques. Elsevier, 2011.

[45] Y. J. Chang et al., “The SACADA database for human reliability and human performance,” Reliability Engineering & System Safety, vol. 125, pp. 117–133, 2014.

[46] R. Boring et al., “Capturing control room simulator data with the HERA System,” presented at the 2007 IEEE 8th Human Factors and Power Plants and HPRCT 13th Annual Meeting, Aug. 2007, pp. 210–217, doi: 10.1109/HFPP.2007.4413208.