

# A classification-based framework for trustworthiness assessment of quantitative risk analysis

Zhiguo Zeng, Enrico Zio

► **To cite this version:**

Zhiguo Zeng, Enrico Zio. A classification-based framework for trustworthiness assessment of quantitative risk analysis. *Safety Science*, Elsevier, 2017, 99, pp.215 - 226. <10.1016/j.ssci.2017.04.001>. <hal-01632271>

**HAL Id: hal-01632271**

**<https://hal.archives-ouvertes.fr/hal-01632271>**

Submitted on 9 Nov 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# A classification-based framework for trustworthiness assessment of quantitative risk analysis

Zhiguo Zeng,<sup>†</sup> Enrico Zio (\*),<sup>†‡</sup>

<sup>†</sup>Chair on System Science and the Energy Challenge, Fondation Electricite de France (EDF), CentraleSupélec, Université Paris-Saclay, Grande Voie des Vignes, 92290 Chatenay-Malabry, France

<sup>‡</sup>Energy Department, Politecnico di Milano, Milano, Italy

## Abstract

In this paper, we develop a classification-based method for the assessment of the trustworthiness of Quantitative Risk Analysis (QRA). The QRA trustworthiness is assumed to be determined by the quality of the QRA process. Six quality criteria, i.e., completeness of documentations, understanding of problem settings, coverage of accident scenarios, appropriateness of analysis methods, quality of input data, accuracy of risk calculation, are identified as the factors most influencing the trustworthiness. **The assessment is, then, formulated as a classification problem, solved by a Naive Bayes Classifier (NBC) constructed based on a set of training data, whose trustworthiness is given by experts. NBC learns the expert's assessment from the training data: therefore, once constructed, the NBC can be used to assess the trustworthiness of QRAs other than the training data.** Leave-one-out cross validation is applied to validate the accuracy of the developed classifier. A stochastic hypothesis testing-based approach is also developed to check the consistency of the training data. The performance of the developed methods is tested for ten artificially generated scenarios. The results demonstrate that the developed framework is able to accurately mimic a variety of expert behaviors in assessing the trustworthiness of QRA.

## Index Terms

Quantitative Risk Analysis (QRA), validity, reliability, trustworthiness, Naive Bayes classifier

## Highlights

- An assessment framework is developed for trustworthiness of QRA.
- A naive Bayes classifier is developed to assess the trustworthiness of QRA.
- Consistency of training data is checked by developing a statistical hypothesis testing.

\* Email of the corresponding author: enrico.zio@ecp.fr, enrico.zio@polimi.it

# A classification-based framework for trustworthiness assessment of quantitative risk analysis

## I. INTRODUCTION

1 Since its first application in nuclear power plants in 1975 [1], Quantitative Risk Analysis (QRA) has been  
2 widely applied in various fields to support safety-related decision making [2], e.g., chemical process industry  
3 [3], oil & gas industry [4], maritime transportation [5], nuclear installations [6], etc. Various methods have  
4 been developed for QRA [7]. According to Khan et al. [8], in general, QRA methods are evolving from semi-  
5 qualitative analysis to detailed quantitative analysis. For example, in the 1990s, QRA in the process industries  
6 were primarily based on semi-qualitative methods like hazard operability (HAZOP) analysis [9], while recent  
7 QRAs are mainly based on detailed quantitative analysis methods, such as Bayesian network [10], bow-tie  
8 model [11], etc. How to compare the trustworthiness of different QRA methods, then, becomes an essential  
9 problem in QRA.

10 Trustworthiness of a QRA refers to the degree that a decision maker can trust the results of QRA [12–  
11 14]. This question is of paramount importance in practice: only a trustworthy QRA can be useful to support  
12 decision making. Trustworthiness assessment of QRA has been discussed by many researchers [15–17], although  
13 sometimes using different concepts and terminologies, e.g., evaluation [17], validation [18], verification [19],  
14 quality assurance [15], credibility assessment [20], etc. Goerlandt et al. present a thorough survey on the status  
15 quo of the trustworthiness assessment of QRA [18]. According to their survey, existing methods on QRA  
16 trustworthiness assessment can be broadly classified into four categories: benchmark exercise, reality check,  
17 independent peer review and quality assurance.

18 Benchmark exercise methods rely on the comparisons among several parallel analyses of QRA to determine  
19 its trustworthiness. Usually, two quality characteristics, i.e., reliability and validity (see [18] for a detailed  
20 discussion), are considered in the comparisons. For example, the trustworthiness of the QRA on an ammonia  
21 storage facility is assessed by comparing seven benchmark exercises in terms of their outcomes, methodologies,  
22 data and models [21]. Reality check methods assess the trustworthiness of the QRA by comparing the results  
23 with real data or operating experience of the same system or process [18]. A typical example is presented in [22],  
24 where statistical data of real accidents and incidents are compared to the risk indexes calculated by QRA, to  
25 evaluate its trustworthiness. In independent peer review methods, the process of QRA and its results are reviewed  
26 by independent experts, based on a series of predefined quality requirements, and the trustworthiness of the QRA  
27 is determined by the experts accordingly [18]. Reference [23] presents a typical example of independent review  
28 methods, where QRA is reviewed and its trustworthiness is determined by experts considering the following  
29 factors, i.e., objective and statement of purpose, project plan and scope of work, figures of merit, methodology,  
30 data base, results, implementation and application and verification of selected results. Quality assurance methods  
31

1 apply quality control techniques on each phase of the QRA process, aiming at ensuring the quality of the QRA  
2 process [18]. It is assumed that a high-quality QRA process will yield trustworthy risk assessment results. For  
3 example, in [15], Suokas and Rouhiainen summarized common flaws in each phase of the QRA process and  
4 developed a check-list-based approach to ensure their quality.

5 Among the four methods, benchmark exercise and reality check take a retroactive perspective on trustworthi-  
6 ness assessment, in the sense that their assessments are primarily done by comparing the results of the analysis  
7 with either parallel analyses or field data and experience; independent peer review and quality assurance, on  
8 the other hand, are proactive in the sense that instead of directly assessing the results, these methods evaluate  
9 the capability of the QRA process (in terms of its quality) and predict the trustworthiness of the assessment  
10 results based on the quality of the QRA process. In a sense, the retroactive perspective is preferred as it is  
11 more trustable, since it directly assesses the trustworthiness of the results of the analysis [24]. However, two  
12 major shortcomings might limit its applicability. First, the retroactive methods are normally more difficult and  
13 expensive to implement, due to the requirements on field data or parallel analyses [18]. Second, the retroactive  
14 methods tell us little about the contributing factors to the trustworthiness, which, limit the ability to guide  
15 improvements of the QRA process for improved trustworthiness [25]. Since in practice, the strict requirements  
16 of the retrospective methods on field data or parallel analyses are always hard to fulfill, in this paper, we focus  
17 only on the proactive methods.

18 In the proactive methods, the trustworthiness is assessed (in fact, predicted) based on the capability (in terms  
19 of quality) of the QRA process. Two issues are essential when developing such methods:

- 20 • how to evaluate the capability (in terms of quality) of a QRA process?
- 21 • how to relate the trustworthiness of QRA to the capability of the QRA process?

22 The first issue has been addressed relatively well in literature: the major influencing factors for the quality of  
23 a QRA process have been widely discussed in literature [26–28] and various methods have been developed  
24 for the assessment [29–31]. The second issue, however, is not so well explored. In most existing researches  
25 ([29–31], for example), the relationship between the trustworthiness and the quality of the QRA process is  
26 treated as a black box and the experts are asked to directly construct a mapping from the process capability to  
27 the trustworthiness of QRA. Usually, this is done by a simple conformance/non-conformance-based framework:  
28 the conclusion of whether the QRA is trustworthy or not is made by comparing the number of the conformed  
29 quality criteria to a predefined threshold value [32]. Such a process is subject to several uncertainties, primarily  
30 due to the opacity in the elicitation process and lack of procedures to assess the accuracy of such assessments  
31 [33]. A formal and quantitative method is, therefore, needed for trustworthiness assessment.

32 In fact, trustworthiness assessment can be viewed as a classification problem (more broadly, supervised  
33 learning, see [34] for details): train a classifier, which is a mapping from the capability of the QRA process  
34 to the trustworthiness of its results, based on a set of training data that are pre-assessed by experts. Hence,  
35 in this paper, we develop a classification-based framework, using Naive Bayes Classifier (NBC), for a formal  
36 and quantitative assessment of the trustworthiness of QRA. NBC is a simple but effective classifier widely  
37 applied in machine learning applications, e.g., text classification [35], tumor diagnosis [36], etc. Although  
38 classification-based frameworks have been developed to assess other qualitative factors, e.g., vulnerability and  
39 safety criticality of nuclear power plants [37], prediction capability of prognostic methods [38], etc., to the best

1 of our knowledge, it is the first time that a classification-based framework is developed for the assessment of  
 2 QRA trustworthiness and that NBC is used for that purpose. **It should be noted that in this paper, the classifier  
 3 is not directly used to assess the trustworthiness. Rather, it is used as a tool for constructing the evaluation  
 4 criteria used for determining the trustworthiness.**

5 The rest of the paper is organized as follows. A general classification-based assessment framework of QRA  
 6 trustworthiness is developed in Section II. NBC is applied in Section III to assess the trustworthiness of QRA.  
 7 In Section IV, we develop a method to check the consistency of the experts that generate the training data. Ten  
 8 numerical case studies are considered and, then, an application is presented regarding a real trustworthiness  
 9 assessment of QRA in Section V. The paper is concluded in Section VI, with a discussion on potential  
 10 future developments.

## 11 II. ASSESSMENT FRAMEWORK

12 In this section, we present a general framework to support classification-based trustworthiness assessment  
 13 of QRA. Let  $T$  represent the trustworthiness of QRA. We take a proactive perspective on trustworthiness  
 14 assessment and assume that  $T$  is determined by the quality of the QRA process. According to Rae et al. [13],  
 15 a typical QRA process involves eight sub-processes, as shown in Figure 1. To ensure the quality of a QRA  
 16 process, all the eight sub-processes should be conducted with high quality [13]. A framework for trustworthiness  
 17 assessment is, then, developed in Figure 2 by considering the quality requirements on the eight sub-processes  
 18 in Figure 1.

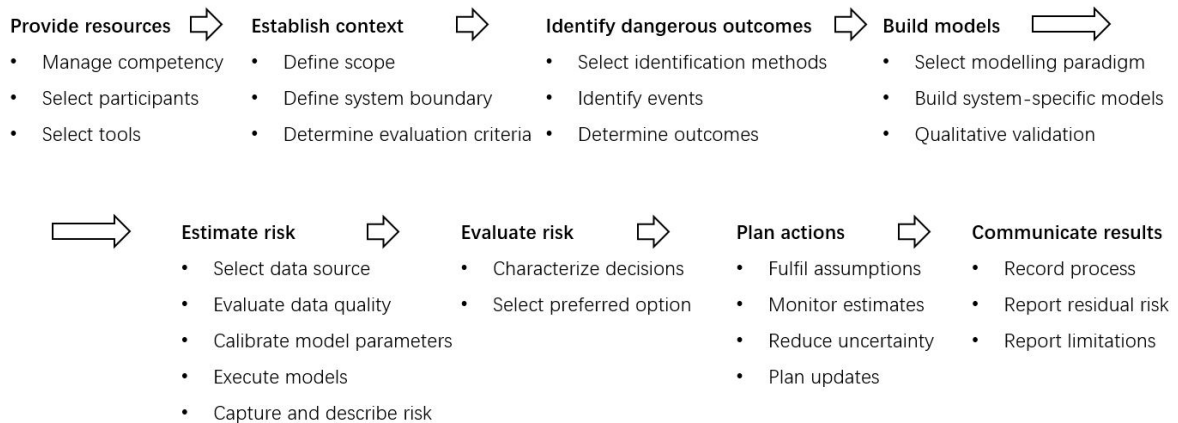


Fig. 1. A typical QRA process [13]

19 In Figure 2, the trustworthiness of QRA is characterized in terms of six criteria, i.e., completeness of docu-  
 20 mentations ( $x_1$ ), understanding of problem settings ( $x_2$ ), coverage of accident scenarios ( $x_3$ ), appropriateness  
 21 of analysis methods ( $x_4$ ), quality of input data ( $x_5$ ), accuracy of risk calculation ( $x_6$ ), which reflect the quality  
 22 requirements on the QRA process. Each criterion is evaluated into three grades, i.e., problematic ( $x_i = 0$ ),  
 23 acceptable ( $x_i = 1$ ) and satisfactory ( $x_i = 2$ ),  $i = 1, 2, \dots, 6$ , based on a set of predefined scaling rules  
 24 in Table A.1-A.6. Three discrete levels of  $T$ , i.e.,  $T \in \{0, 1, 2\}$ , are considered in this paper. The levels are  
 25 distinguished in Table I based on their reliability, which concerns the repeatability of the risk analysis [26]

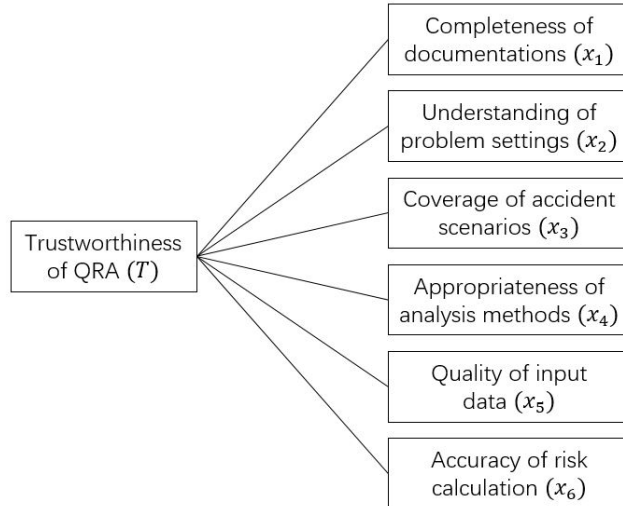


Fig. 2. Trustworthiness assessment framework

1 and validity, which concerns whether the risk analysis addresses the “right problem” [26]. The problem of  
 2 trustworthiness assessment is, then, formulated as a classification problem: given the states of the six criteria  
 3  $x_1, x_2, \dots, x_6$ , determine an appropriate category for the trustworthiness  $T$ . It should be noted that both the  
 4 assessment framework in Figure 2 and the scaling rules in Table A.1-A.6 are constructed for illustrative purposes.  
 5 They are defined in a general form that allows them to be adapted for capturing the problem-specific features  
 6 in practical applications.

TABLE I  
 THREE LEVELS FOR  $T$

Levels of trustworthiness	Descriptions
$T = 0$ : Unreliable	<ul style="list-style-type: none"> <li>• The result of the QRA is unrepeatable.</li> <li>• No further judgements can be made on the trustworthiness of the QRA.</li> <li>• Such QRA should not be used to support any decision making.</li> </ul>
$T = 1$ : Reliable but invalid	<ul style="list-style-type: none"> <li>• The result of QRA is repeatable but</li> <li>• some critical hazards are not identified and analyzed by the QRA or</li> <li>• some important risks (and their uncertainties) are not accurately quantified by the QRA.</li> <li>• Such QRA can be used to support decision making, but not for safety-critical decisions.</li> </ul>
$T = 2$ : Reliable and valid	<ul style="list-style-type: none"> <li>• The result of the QRA is repeatable and</li> <li>• all critical hazards are identified and analyzed by the QRA;</li> <li>• all important risks (and their uncertainties) are accurately quantified by the QRA.</li> <li>• Such QRA can be used to support critical decision making.</li> </ul>

### 7 III. TRUSTWORTHINESS ASSESSMENT BASED ON NAIVE BAYES CLASSIFIER

8 In this section, we first review some preliminaries on NBC-based classification in Subsection III-A and, then,  
 9 develop a NBC-based method to assess the trustworthiness of QRA in Subsection III-B.

1 *A. Naive Bayes classifier*

2 Let us define  $\mathbf{x} = [x_1, x_2, \dots, x_n] \in \mathbb{X}$  to be the input feature vector of the classification problem, where  
 3  $\mathbb{X}$  is the feature space. A NBC is a function  $f_{\text{NBC}}$  that maps input feature vectors  $\mathbf{x} \in \mathbb{X}$  to output class  
 4 labels  $T \in \{0, 1, \dots, C\}$  [39]. Usually, the feature vector also takes discrete values, so that we have  $x_i \in$   
 5  $\{0, 1, \dots, n_i\}, i = 1, 2, \dots, n$ . Given a feature vector  $\mathbf{x}$ , a NBC classifies it into the class with the maximum  
 6 posterior probability [39]:

$$T = \arg \max_T Pr(T | \mathbf{x}). \quad (1)$$

7 The posterior probability in (1) is calculated using Bayes rule [39]:

$$Pr(T | \mathbf{x}) = \frac{Pr(\mathbf{x}, T)}{Pr(\mathbf{x})} = \frac{Pr(\mathbf{x} | T)Pr(T)}{\sum_{T=0}^C Pr(\mathbf{x} | T)Pr(T)}. \quad (2)$$

8 If we further assume that the elements  $x_i, i = 1, 2, \dots, n$  of the input feature vector  $\mathbf{x}$  are independent, the  
 9 nominator of (2) becomes:

$$Pr(\mathbf{x} | T)Pr(T) = Pr(T) \prod_{i=1}^n Pr(x_i | T). \quad (3)$$

10 Note that the denominator in (2) is the same for all possible values of  $T$ . Therefore, (1) can be simplified:

$$T = \arg \max_T Pr(T) \prod_{i=1}^n Pr(x_i | T). \quad (4)$$

11 In order to apply the NBC, the  $Pr(T)$  and  $Pr(x_i | T)$  in (4) should be estimated from training data. Training  
 12 data are a set of samples whose correct classes are already known. Suppose we have  $N_{\text{training}}$  training data,  
 13 denoted by  $(\mathbf{x}^{(q)}, T^{(q)}), q = 1, 2, \dots, N_{\text{training}}$ . Then, the required probabilities are estimated by:

$$Pr(T = k) = \frac{\sum_{q=1}^{N_{\text{training}}} \mathbb{1}(T^{(q)} = k)}{N_{\text{training}}}, \quad (5)$$

14

$$Pr(x_i = j | T = k) = \frac{\sum_{q=1}^{N_{\text{training}}} \mathbb{1}(x_i^{(q)} = j, T^{(q)} = k)}{\sum_{q=1}^{N_{\text{training}}} \mathbb{1}(T^{(q)} = k)}, \quad (6)$$

15 where  $\mathbb{1}(\cdot)$  is the indicator function and  $i = 1, 2, \dots, n, j = 0, 1, \dots, n_i, k = 0, 1, \dots, C$ .

16 There is one potential problem for (5) and (6). Suppose that due to statistical variations, for some specific  
 17 values of  $j$  and  $k$ , we have  $\sum_{q=1}^{N_{\text{training}}} \mathbb{1}(x_i^{(q)} = j, T^{(q)} = k) = 0$ . In this case,  $Pr(x_i = j | T = k) = 0$ ,  
 18 which, according to (3), results in  $Pr(\mathbf{x} | T) = 0$ , regardless of the posterior probabilities for other features.  
 19 Misclassification often happens in such situations. To avoid such a problem, a technique called Laplacian  
 20 correction is often applied when estimating  $Pr(T = k)$  and  $Pr(x_i = j | T = k)$  [39]:

$$Pr(T = k) = \frac{\sum_{q=1}^{N_{\text{training}}} \mathbb{1}(T^{(q)} = k) + \gamma}{N_{\text{training}} + (C + 1) \cdot \gamma}, \quad (7)$$

21

$$Pr(x_i = j | T = k) = \frac{\sum_{q=1}^{N_{\text{training}}} \mathbb{1}(x_i^{(q)} = j, T^{(q)} = k) + \gamma}{\sum_{q=1}^{N_{\text{training}}} \mathbb{1}(T^{(q)} = k) + (n_i + 1) \cdot \gamma}, \quad (8)$$

22 where  $\gamma \in (0, 1]$  is an adjustment factor introduced to compensate for the possible zero probabilities;  $C + 1$   
 23 and  $n_i + 1$  are the number of possible values for  $T$  and  $x_i$ , respectively.

## 1 B. Trustworthiness assessment

2 In this section, we apply the NBC to develop a classifier for the trustworthiness assessment problem in Figure  
 3 2. In this case, we have six features, i.e.,  $\mathbf{x} = [x_1, x_2, \dots, x_6]^T$ . Each feature has three discrete levels, i.e.,  
 4  $x_i \in \{0, 1, 2\}, i = 1, 2, \dots, 6$ . Hence,  $\mathbb{X} = \{0, 1, 2\} \times \dots \times \{0, 1, 2\} = \{0, 1, 2\}^6$ . The trustworthiness also  
 5 takes three values, i.e.,  $T \in \{0, 1, 2\}$ . In general, three steps are involved in the development of the classifier,  
 6 as shown in Figure 3.

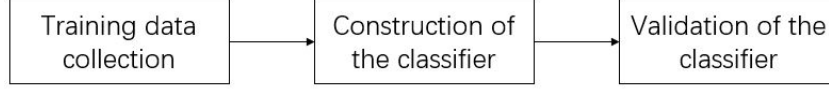


Fig. 3. Procedures of constructing the NBC for trustworthiness assessment

7 1) *Training data collection*: Since  $\mathbb{X} = \{0, 1, 2\}^6$ , the feature vector  $\mathbf{x}$  can take  $3^6 = 729$  different values. A  
 8 fraction of them, denoted by  $\mathbf{x}^{(q)}, q = 1, 2, \dots, N_{training}$ , are selected as training samples. The trustworthiness  
 9 of these training samples, denoted by  $T^{(q)}, q = 1, 2, \dots, N_{training}$ , are evaluated by experts, based on the  
 10 descriptions in Table I. The training data are, then, used to construct the NBC and once constructed, the NBC  
 11 is exploited to replace the expert for the assessment of trustworthiness.

12 Since the NBC learns the expert's evaluation rationale from the training data, it is essential that the training  
 13 data are a reasonable representation of the whole feature space. On the other hand, we want to reduce the number  
 14 of training data as much as possible, since collecting training data is often expensive and time-consuming. For  
 15 this, in this paper, we use an experiment design technique, i.e., the row-exchange algorithm in Matlab R2015b,  
 16 to design the training data collection scheme. The response model in the row-exchange algorithm is assumed  
 17 to be a linear model and the resulted D-optimal design matrix is used for the collection of training data. This  
 18 approximates an orthogonal design on the  $\mathbf{x}^{(q)}, q = 1, 2, \dots, N_{training}$ , where the collected training data are  
 19 equally distributed and can equally "represent" the entire space of  $\mathbb{X}$ .

20 Another issue that needs to be considered when designing the training data collection scheme is the sample  
 21 size  $N_{training}$ . Apparently, a large value of  $N_{training}$  would enhance the performance of the developed classifier  
 22 in terms of its accuracy. On the other hand, large values of  $N_{training}$  also create more difficulties in collecting  
 23 the data (experts easily get impatient when asked to judge too many scenarios). Hence, a trade-off needs to be  
 24 made in determining the value of  $N_{training}$ .

25 2) *Construction of the classifier*: The procedures for constructing the NBC is summarized in Figure 4. In the  
 26 preparation phase, the sample size of the training set and the training data collection scheme are determined using  
 27 the methods discussed previously. The training data  $(\mathbf{x}^{(q)}, T^{(q)}), q = 1, 2, \dots, N_{training}$  are, then, collected by  
 28 expert judgements following the scaling rules in Table A.1-A.6. In the training phase, the NBC is constructed  
 29 by estimating  $Pr(T)$  and  $Pr(x_i|T)$  from the training data, using (7) and (8), respectively. In the evaluation  
 30 phase, the constructed NBC is applied to replace the role of the experts and determine the trustworthiness of  
 31 a new QRA. By reviewing the related documents, the value for the feature vector  $\mathbf{x}$  of the QRA is determined  
 32 first, based on the scaling rules defined in Table A.1-A.6. Its trustworthiness is, then, determined based on the  
 33 constructed NBC using (4).



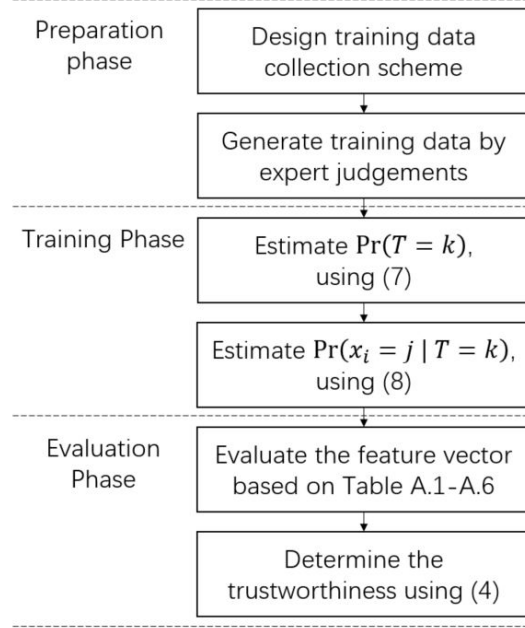


Fig. 4. NBC construction procedure for QRA trustworthiness assessment

1 3) *Validation of the classifier*: In practice, the sample size of the training data available for the construction  
 2 of the NBC is always small, which might impair the accuracy of the classifier. If the NBC is not accurate  
 3 enough, it might not be able to correctly “mimic” the expert’s behavior in assessing the trustworthiness. It is,  
 4 then, necessary for us to consider the validation of the developed NBC, i.e., to determine our confidence that the  
 5 classifier can correctly represent the experts’ judgement behaviors. As [37], such confidence is measured by the  
 6 probability that the classifier correctly determines the trustworthiness of a QRA, denoted by  $CR$ . Leave-One-  
 7 Out Cross Validation (LOOCV) is exploited in this paper to estimate  $CR$ , where one sample from the training  
 8 data is left to test the model while the remaining training data are used to train the classifier [33, 34]. The  
 9 procedures of implementing LOOCV is summarized in Figure 5, where  $(\mathbf{x}^{(q)}, T^{(q)})$ ,  $q = 1, 2, \dots, N_{training}$   
 10 are training data and  $CR_{CV}$  is the correctness rate estimated by LOOCV. The initial values for  $i$  and  $sum$  are  
 11  $i = 0$ ,  $sum = 0$ , respectively.

#### 12 IV. CHECKING THE CONSISTENCY OF TRAINING DATA

13 Since the training data are empirically assessed by the experts, the consistency of the expert, therefore, is  
 14 essential in the NBC-based trustworthiness assessment. Training data generated by an inconsistent expert might  
 15 be self-contradicting and therefore misleading. In this section, we develop a statistical hypothesis testing to  
 16 check the consistency of the training data.

17 Motivated by the methods for consistency checking used in the Analytical Hierarchy Process (AHP) [40],  
 18 we assume that if an expert is inconsistent, he/she would classify a feature vector  $\mathbf{x}$  in the training data set to  
 19 a random trustworthiness level, regardless of the value of  $\mathbf{x}$ . Suppose we have  $T \in \{0, 1, \dots, C\}$ . Then, for  
 20 any feature vector  $\mathbf{x}$  in the training data set, an inconsistent expert would judge it to be  $T = i$ ,  $i = 0, 2, \dots, C$   
 21 with probability  $1/(C + 1)$ . Hence, we develop the following hypothesis testing to check the consistency of

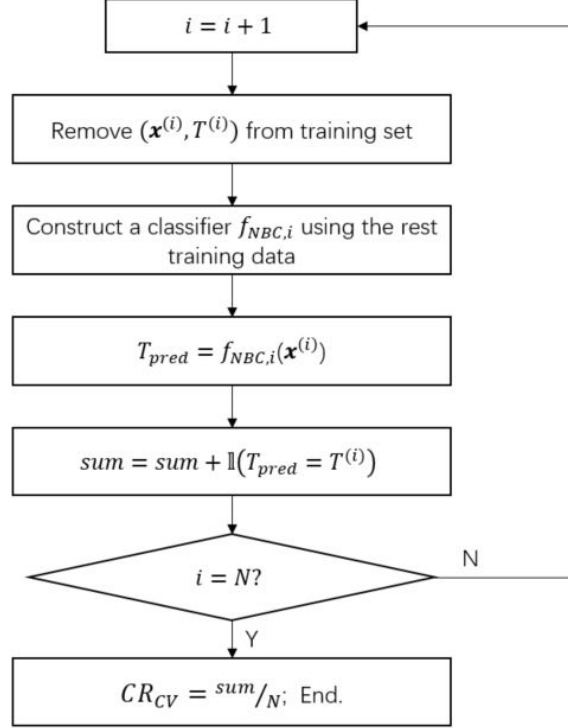


Fig. 5. Procedures of implementing LOOCV

1 the training data:

$$H_0 : \text{The expert is inconsistent.} \quad H_1 : \text{The expert is consistent.} \quad (9)$$

2 In (9),  $H_0$  and  $H_1$  are the null hypothesis and alternate hypothesis, respectively. If the observed data can support  
 3 us to reject the null hypothesis, the training data is believed to be consistent and can be used to train the NBC;  
 4 otherwise, we cannot trust the consistency of the training data and a reevaluation of the training data is required.

5 The  $CR_{CV}$  estimated by LOOCV is chosen as the test statistic:

$$CR_{CV} = \frac{\sum_{i=1}^{N_{training}} \mathbb{1}(T_{pred}^{(i)} = T^{(i)})}{N_{training}}, \quad (10)$$

6 where  $T_{pred}^{(i)}$  and  $T^{(i)}$  are the predicted and true classes of the  $i$ th cross validations, respectively. The empirical  
 7 distribution of the test statistic under null hypothesis can be approximated using randomly generated training  
 8 data. Suppose for a given significance level  $\alpha$ ,  $p_{1-\alpha}$  denotes the  $(1 - \alpha)$  percentile of the empirical distribution  
 9 of the  $CR_{CV}$  calculated using the randomly generated training data. Then, decisions on the consistency of  
 10 the training data can be made by comparing  $p_{1-\alpha}$  and  $CR_{data}$ , which is calculated using the real training  
 11 data [41, 42]:

- 12 • If  $CR_{data} > p_{1-\alpha}$ , reject  $H_0$ , the expert is consistent;
- 13 • otherwise, cannot reject  $H_0$ , the expert is inconsistent.

14 The physical meaning of the significance level  $\alpha$  is the probability that an inconsistent expert is mistakenly  
 15 judged consistent by the test [42]. Depending on the confidence requirements, two values of  $\alpha$  are commonly  
 16 used:  $\alpha = 0.01$  and  $\alpha = 0.05$ .

1 Algorithm 1 below presents a pseudo-code for the developed consistency-checking method. In Algorithm 1,  
 2  $n_{\text{random}}$  is the sample size used to approximate the distribution of the test statistic. In principle, we prefer a  
 3 large value of  $n_{\text{random}}$  since it helps to reduce the uncertainties in the decision caused by the approximation  
 4 errors in estimating  $p_{1-\alpha}$ . However, balance is needed between the accuracy of the estimation and the required  
 5 computational costs. If Algorithm 1 returns  $\text{IsConsist} = 1$ , we conclude that the training data under evaluation  
 6 is consistent under the significance level  $\alpha$ ; otherwise, we cannot reach the conclusion that the training data  
 7 are consistent and a re-evaluation of the training data is required.

---

**Algorithm 1** Consistency verification based on hypothesis testing

---

**Inputs:**  $\alpha, n_{\text{random}}, (\mathbf{x}^{(q)}, T^{(q)}), q = 1, 2, \dots, N_{\text{training}}$

**Onputs:**  $\text{IsConsist}$

```

1: for  $i = 1 : n_{\text{random}}$  do
2:    $T_{\text{random}}^{(q)} \leftarrow \text{GENRNDSAMPLE}, q = 1, 2, \dots, N_{\text{training}};$ 
3:    $CR_{\text{random}}^{(i)} \leftarrow \text{Do LOOCV using the randomly generated training data } (\mathbf{x}^{(q)}, T_{\text{random}}^{(q)}), q =$ 
    $1, 2, \dots, N_{\text{training}};$ 
4: end for
5:  $CR_{\text{sort}} \leftarrow \text{Sort } CR_{\text{random}}$  in ascending order;
6:  $p_{1-\alpha} \leftarrow CR_{\text{sort}}^{(n_{\text{random}} \cdot (1-\alpha))};$ 
7:  $CR_{\text{data}}^{(i)} \leftarrow \text{Do LOOCV using real training data } (\mathbf{x}^{(q)}, T^{(q)}), q = 1, 2, \dots, N_{\text{training}};$ 
8: if  $CR_{\text{data}} > p_{1-\alpha}$  then
9:   return  $\text{IsConsist} = 1;$ 
10: else
11:   return  $\text{IsConsist} = 0;$ 
12: end if

```

**Inputs:** NULL

**Onputs:**  $y_{\text{random}}$

```

13: function GENRNDSAMPLE
14:   Generate a random number  $r$ ;
15:   if  $r \leq 1/3$  then
16:     return  $y_{\text{random}} = 0;$ 
17:   else if  $r \leq 2/3$  then
18:     return  $y_{\text{random}} = 1;$ 
19:   else
20:     return  $y_{\text{random}} = 2;$ 
21:   end if
22: end function

```

---

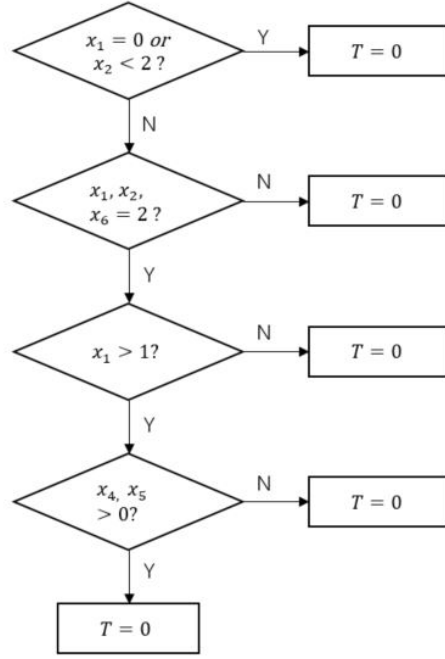


Fig. 6. Illustration of the assumed expert rationale

1

## V. APPLICATIONS

2 In this section, we first test the performance of the developed methods under ten artificially generated scenarios  
 3 in Subsection V-A. Then, the methods are applied in Subsection V-B to assess the trustworthiness of a real-world  
 4 methanol QRA.

### 5 A. A numerical case study

6 Ten scenarios are artificially generated to test the performance of the developed algorithms. For each scenario,  
 7 we assume that the behavior of the experts when assessing the trustworthiness is consistent and known to us.  
 8 Ten different expert behaviors are used to generate the testing scenarios. We present one of the assumed expert  
 9 behaviors in Figure 6 for illustration purposes.

10 Let  $\mathbf{x}_{full}^{(i)}$  represent all the possible states in the feature space  $\mathbb{X}$ , where in this case, we have  $i = 1, 2, \dots, 729$ .  
 11 The true trustworthiness for each  $\mathbf{x}_{full}^{(i)}$ , denoted by  $T_{full}^{(i)}$ , can, then, be determined by the same known  
 12 assessment behavior. According to the designed training data collection scheme, a fraction of the  $(\mathbf{x}_{full}^{(i)}, T_{full}^{(i)})$   
 13 is selected as the training data and the remaining elements are used as test data to assess the performance of  
 14 the developed NBC. Let the test data be  $(\mathbf{x}_{test}^{(i)}, T_{test}^{(i)})$ ,  $i = 1, 2, \dots, N_{test}$ . For each artificially generated  
 15 scenario, two numerical metrics are calculated:

- 16 •  $CR_{CV}$ , calculated by LOOCV using the training data set, based on (10);
- 17 •  $CR_{test}$ , the classification correctness rate, calculated using the test data set:

$$CR_{test} = \frac{\sum_{i=1}^{N_{test}} \mathbb{1}(T_{pred}^{(i)} = T_{test}^{(i)})}{N_{test}}, \quad (11)$$

1 The influence of the size of the training data set on the performance of trustworthiness assessment is also  
 2 investigated. For this, nine levels of  $N_{training}$ , i.e.,  $N_{training} = 9, 18, 27, 36, 45, 54, 63, 72, 81$  are considered  
 3 and the row-exchange algorithm in Matlab R2015b is used to design the training data collection scheme for  
 4 each value of  $N_{training}$ .

5 Figure 7 and Figure 8 present the values of  $CR_{CV}$  and  $CR_{test}$ , evaluated under different scenarios, using  
 6 different numbers of training data. It can be seen from the two Figures that both  $CR_{CV}$  and  $CR_{full}$  are  
 7 improved as the number of training data  $N_{training}$  increases. Hence, we can improve the accuracy of the  
 8 developed classifier by choosing a larger  $N_{training}$ .

9 Although the classification accuracies are affected by uncertainties arising from the difference in the expert  
 10 judgement behaviors adopted in different testing scenarios, the developed NBC can, in general, achieve satis-  
 11 factory accuracies even for small values of  $N_{training}$ . As demonstrated in Figure 7: the average  $CR_{test}$  exceeds  
 12 0.9 for  $N_{training}$  greater than 18. However, when  $N_{training} = 9$ , the classification accuracy is relatively poor.  
 13 This is because when  $N_{training} = 9$ , the training data set is too small for accurate estimation of the posterior  
 14 probabilities using (7) and (8). To avoid such a problem, it is suggested to ensure that  $N_{training} \geq 18$  in  
 15 practical applications.

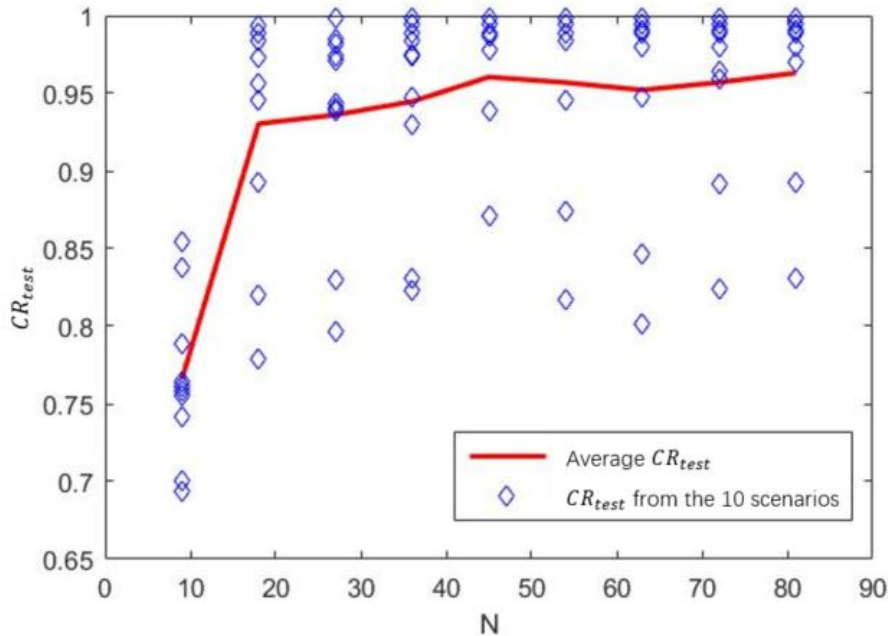


Fig. 7.  $CR_{test}$  evaluated under different scenarios, with different numbers of training data

16 A comparison is made between  $CR_{CV}$  and  $CR_{test}$  in Figure 9. It can be seen that although  $CR_{CV}$  is estimated  
 17 by cross-validation using only the training data, it shows the same tendency as  $CR_{full}$ , which is estimated  
 18 using the true data outside the training data set. Therefore, although sometimes in practical applications, we  
 19 can only do cross-validation using the training data, a reasonable estimate of the classification accuracy can  
 20 still be achieved.

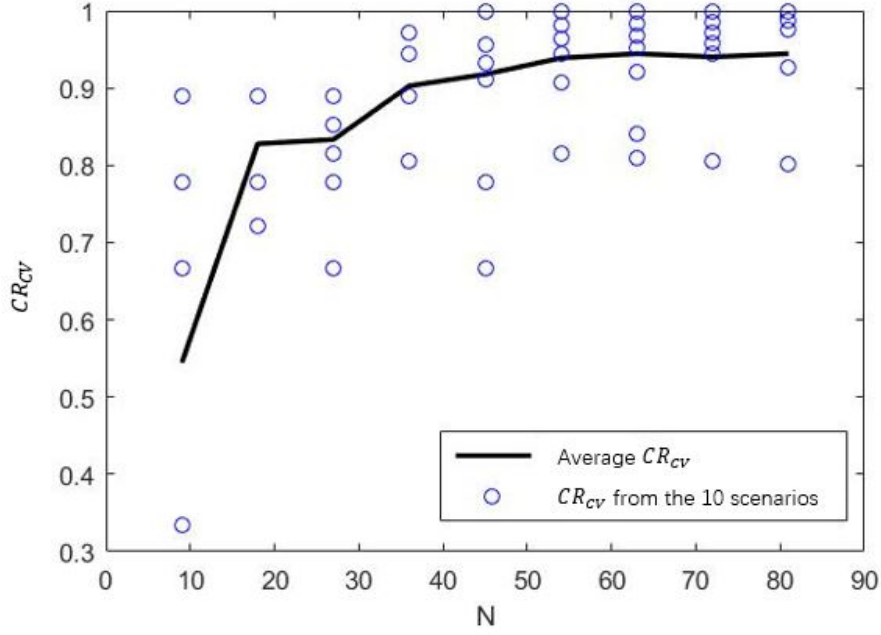


Fig. 8.  $CR_{CV}$  evaluated under different scenarios, with different numbers of training data

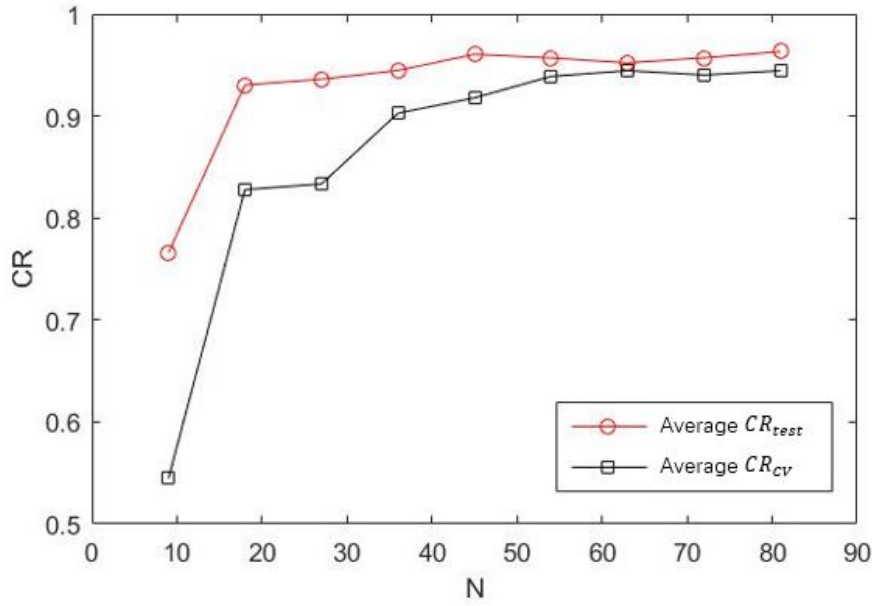


Fig. 9. A comparison of  $CR_{CV}$  and  $CR_{test}$

## 1 B. Application

2 In this section, we show how to apply the developed framework to assess the trustworthiness of a real-world  
 3 methanol plant, wherein the associated individual and social risks are assessed by a systematic QRA process,  
 4 in terms of risk contours and F-N curve, respectively [43]. The training data used for the construction of the  
 5 NBC are generated by asking an expert to assess the trustworthiness of a set of artificially generated “pseudo”

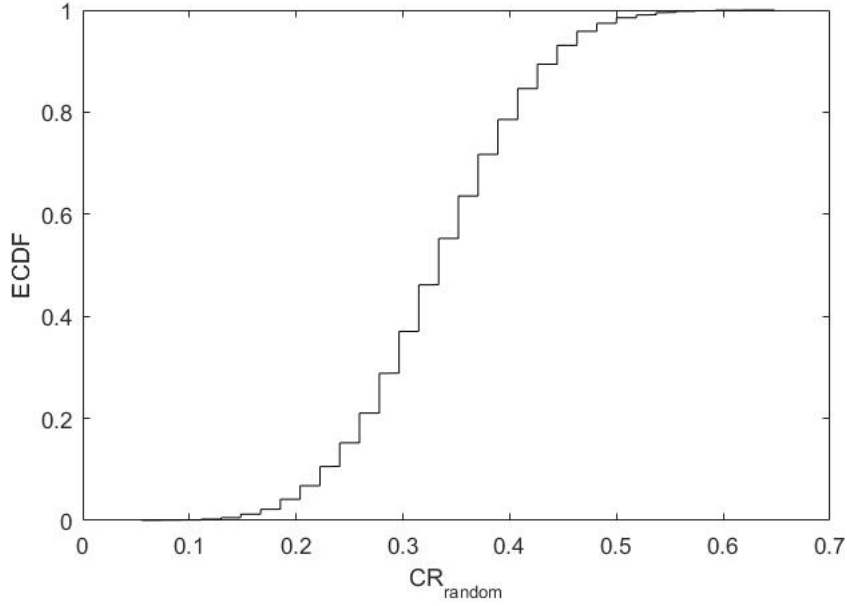


Fig. 10. ECDF of  $CR_{\text{random}}$

1 QRAs. The quality criteria of the methanol QRA is evaluated by reviewing its final report, which is available  
 2 online from [43]. The application follows the procedures in Figure 4 and the main results are summarized as  
 3 follows.

4 1) *Training data collection scheme*: In this step, we design the training data collection scheme. From the  
 5 discussion in Subsection V-A, we can see that  $N_{\text{training}} = 54$  can, in general, yield good classification accuracy.  
 6 Therefore, we choose  $N_{\text{training}} = 54$ . The row-exchange algorithm in Matlab R2015b is used to design the  
 7 training data collection scheme. The resulting  $\mathbf{x}^{(q)}, q = 1, 2, \dots, N_{\text{training}}$  are listed in Table II. It can be  
 8 verified that the training data collection scheme in Table II is an orthogonal design. The values of  $\mathbf{x}^{(q)}, q =$   
 9  $1, 2, \dots, N_{\text{training}}$  correspond to the levels of the quality criteria in Table A.1-A.6.

10 2) *Training data collection*: Each row in Table II represents a pseudo QRA, characterized by specific quality  
 11 criteria. An expert is asked to assess the trustworthiness for these pseudo QRAs, for generating the training  
 12 data. Take the first row in Table II as an example. To generate the training data, the expert is asked the following  
 13 question: if the quality of a QRA process is as depicted in Table III, which level of trustworthiness in Table  
 14 I do you think the QRA has? Table III is generated by relating the values of  $\mathbf{x}^{(q)}, q = 1, 2, \dots, N_{\text{training}}$  to  
 15 the corresponding quality criteria in Table A.1-A.6. The procedures are repeated for the other rows in Table II.  
 16 The training data generated by the expert are also listed in Table II.

17 3) *Consistency verification*: The consistency of the expert is checked using Algorithm 1. In this case, we  
 18 choose  $\alpha = 0.01$  and  $n_{\text{random}} = 10^4$ . The Empirical Cumulative Distributive Function (ECDF) of  $CR_{\text{random}}$  is  
 19 presented in Figure 10, where  $p_{1-\alpha} = 0.519$ . The procedures in Figure 5 are used to calculate  $CR_{\text{data}}$  using  
 20 the real training data. Since  $CR_{\text{data}} = 0.852 > p_{1-\alpha}$ , according to Algorithm 1, we can conclude that the  
 21 expert provided the training data consistently under the significance level  $\alpha = 0.01$ .

TABLE II  
TRAINING DATA

Runs	$x_1^{(q)}$	$x_2^{(q)}$	$x_3^{(q)}$	$x_4^{(q)}$	$x_5^{(q)}$	$x_6^{(q)}$	$T^{(q)}$	Runs	$x_1^{(q)}$	$x_2^{(q)}$	$x_3^{(q)}$	$x_4^{(q)}$	$x_5^{(q)}$	$x_6^{(q)}$	$T^{(q)}$
1	0	0	0	0	1	0	0	28	1	1	1	2	1	1	1
2	0	0	0	0	2	2	0	29	1	1	2	1	0	1	1
3	0	0	1	2	1	1	0	30	1	1	2	2	1	2	1
4	0	0	1	2	2	2	0	31	1	2	0	1	2	0	1
5	0	0	2	1	0	1	0	32	1	2	0	2	0	1	1
6	0	0	2	1	1	1	0	33	1	2	1	0	1	0	1
7	0	1	0	1	0	0	0	34	1	2	1	0	1	2	1
8	0	1	1	0	2	1	0	35	1	2	2	0	2	0	1
9	0	1	1	2	0	0	0	36	1	2	2	2	0	2	2
10	0	1	2	0	0	0	0	37	2	0	0	0	0	2	0
11	0	1	2	0	1	2	0	38	2	0	0	2	1	0	0
12	0	1	2	2	2	0	0	39	2	0	1	1	0	2	1
13	0	2	0	1	1	1	0	40	2	0	1	2	2	0	1
14	0	2	0	1	1	2	0	41	2	0	2	0	0	0	0
15	0	2	0	2	2	0	0	42	2	0	2	2	2	1	2
16	0	2	1	0	2	1	0	43	2	1	0	0	0	1	0
17	0	2	1	2	0	2	0	44	2	1	0	1	2	2	1
18	0	2	2	1	0	2	0	45	2	1	0	2	1	0	1
19	1	0	0	0	1	2	0	46	2	1	1	1	1	1	1
20	1	0	0	2	0	1	0	47	2	1	1	1	2	2	1
21	1	0	1	1	0	2	1	48	2	1	2	0	1	2	1
22	1	0	1	1	1	0	1	49	2	2	0	0	0	1	0
23	1	0	2	0	2	1	1	50	2	2	1	0	2	1	1
24	1	0	2	1	2	0	1	51	2	2	1	1	0	0	1
25	1	1	0	1	2	1	1	52	2	2	2	1	1	0	1
26	1	1	0	2	2	2	1	53	2	2	2	2	1	1	2
27	1	1	1	0	0	0	0	54	2	2	2	2	2	2	2

1 4) *Classifier construction*: The training data are used to construct the NBC, following the procedures  
2 in Figure 4. The estimated  $Pr(T = k)$  and  $Pr(x_i = j | T = k)$  are presented in Table IV and Figure 11,  
3 respectively. The accuracy of the constructed classifier is evaluated by the correct classification rate and we  
4 have  $CR = 0.944$ . Therefore, the constructed NBC can be used to represent the expert judgements and provide  
5 reasonable assessment of the trustworthiness of QRA.

6 The constructed NBC can also help to explain the expert's behavior in assessing the trustworthiness. For  
7 example, from Figure 11, we notice that  $Pr(x_1 = 0 | T = 0) = 0.6882$ ,  $Pr(x_1 = 0 | T = 1) = 0.0041$ ,  
8  $Pr(x_1 = 2 | T = 0) = 0.0233$ . From Bayes theorem,



TABLE III  
QUALITY OF THE FIRST PSEUDO QRA

Criteria	Level
Completeness of documentation	Some the following elements are missing in the documentations: <ul style="list-style-type: none"> <li>• scopes and objectives of the QRA;</li> <li>• descriptions of the system under investigation and related references;</li> <li>• accounts of the adopted analysis methods;</li> <li>• presentation of source data needed for the analysis;</li> <li>• report of the analysis results.</li> </ul>
Understanding of problem settings	The analysts are unaware of the problem settings of the QRA due to the presence of all the following flaws: <ul style="list-style-type: none"> <li>• the purposes of the QRA are not clearly understood;</li> <li>• the systems of interests are not well defined;</li> <li>• the resources constraints (e.g., time, computational resources, etc) are not clearly defined.</li> </ul>
Coverage of accident scenarios	Some critical accident scenarios are highly likely to be missed by the identification process: <ul style="list-style-type: none"> <li>• the coverage of the identified accident scenarios is not validated;</li> <li>• the validation shows that some critical accident scenarios might be missing.</li> </ul>
Appropriateness of analysis methods	<ul style="list-style-type: none"> <li>• The features of the selected analysis method satisfy the requirements of the problem and successful applications in similar problems can justify the choice of the method.</li> </ul>
Quality of input data	<ul style="list-style-type: none"> <li>• There is no sufficient statistical data and the input data is purely based on expert judgements;</li> <li>• epistemic uncertainty in the expert-generated input data is not considered.</li> </ul>
Accuracy of risk calculation	<ul style="list-style-type: none"> <li>• Only errors from the calculation process itself (e.g., the accuracy of Monte Carlo simulations) might exist and</li> <li>• the uncertainties caused by the errors are properly modeled.</li> </ul>

TABLE IV  
ESTIMATED  $Pr(T = k)$

$k$	$Pr(T = k)$
0	0.4807
1	0.4438
2	0.0755

$$Pr(T = 0 | x_1 = 0) = \frac{Pr(x_1 = 0 | T = 0) \cdot Pr(T = 0)}{Pr(x_1 = 0)}$$

$$= \frac{Pr(x_1 = 0 | T = 0) \cdot Pr(T = 0)}{\sum_{i=1}^3 Pr(x_1 = 0 | T = i) \cdot Pr(T = i)} \quad (12)$$

$$= 0.9896 \quad (13)$$

1 That is, if  $x_1$  equals to zero, the expert tends to judge the QRA as unreliable. This is a natural result, since  $x_1$   
2 denotes the completeness of documentations. If the QRA process is not well-documented, it is unlikely to be  
3 repeatable: therefore, the associated QRA is unreliable according to the criteria in Table I.

4 5) *Comparison to existing methods:* In traditional proactive trustworthiness assessment methods, e.g., [32],  
5 expert knowledge is elicited to develop a simple conformance/non-conformance-based framework that relates  
6 the quality criteria to the trustworthiness of the QRA. That is, the conclusion of whether the QRA is trustworthy

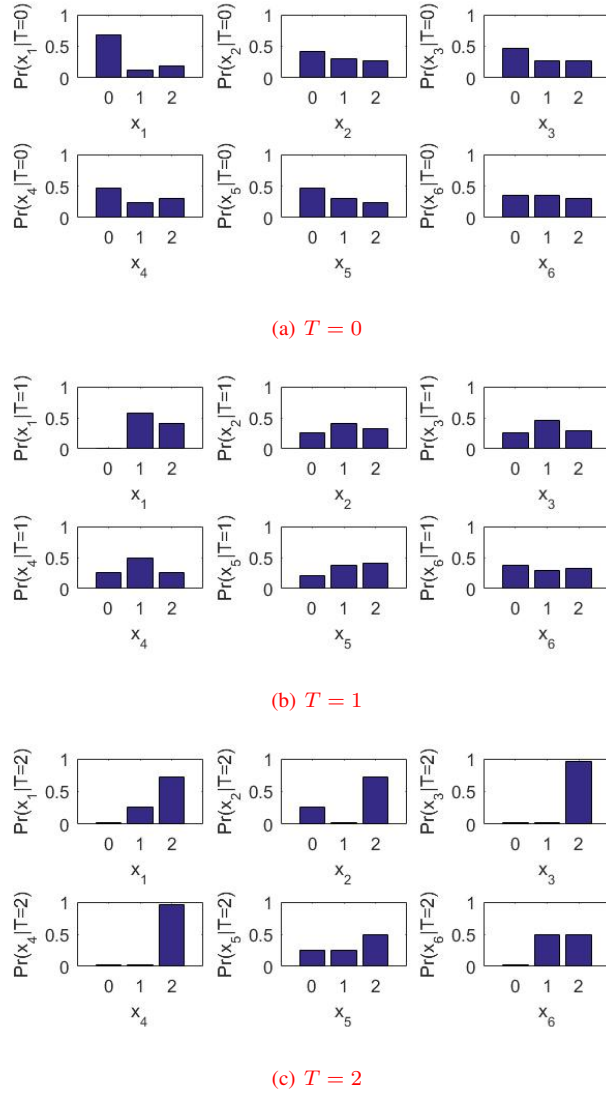


Fig. 11. Estimated  $Pr(x_i = j | T = k)$

1 or not is made by comparing the number of the conformed quality criteria to a predefined threshold value  $n_{th}$ .  
 2 In this paper, we assume that a quality criterion  $i$  is conformed when  $x_i = 2$ . Table V shows a comparison  
 3 between the classification-based framework and the conformance/non-conformance-based framework, using the  
 4 training data in Table II. It can be seen that in general, the existing conformance/non-conformance-based  
 5 framework cannot accurately model the complex expert knowledge expressed in the empirical data in Table II.  
 6 The developed method, on the other hand, is capable of capturing the complex behavior of expert judgement  
 7 in assessing the trustworthiness of the QRA.

8 6) *Trustworthiness assessment*: To assess the trustworthiness of the methanol QRA using the developed  
 9 NBC, its six quality criteria are first evaluated based on the QRA report [43] and following the scaling rules  
 10 in Table A.1-A.6:

- 11 • The scaling rule for completeness of documentation ( $x_1$ ) is listed in Table A.1. In general, the methanol  
 12 QRA report contains sufficient information on the scope and objective of the analysis, the system under

TABLE V  
A COMPARISON TO EXISTING METHODS

Methods	Correct classification rate
Classification-based method	$CR = 0.944$
Conformance-based method	$CR = 0.130$ , for $n_{th} = 0$
	$CR = 0.315$ , for $n_{th} = 1$
	$CR = 0.463$ , for $n_{th} = 2$
	$CR = 0.556$ , for $n_{th} = 3$
	$CR = 0.500$ , for $n_{th} = 4$
	$CR = 0.500$ , for $n_{th} = 5$
	$CR = 0.482$ , for $n_{th} = 6$

1 investigation and the adopted analysis methods. However, according to Table A.1, the presentation of the  
 2 analysis results is incomplete, since no accounts of uncertainty are given in the report. Therefore, we have  
 3  $x_1 = 1$ .

4 • The scaling rule for understanding of problem setting ( $x_2$ ) is listed in Table A.2. As reflected in the report,  
 5 the analyst understood well the purposes of the QRA, the system under investigation and the resources  
 6 constraints of the analysis. Therefore, we have  $x_2 = 2$ .

7 • The scaling rule for coverage of accident scenario ( $x_3$ ) is listed in Table A.3. The methanol QRA iden-  
 8 tifies accident scenarios by an initial HAZard IDentification (HAZID) workshop. The identified accident  
 9 scenarios are verified by expert reviews, conducted by experts from the QRA analysis team, the operator  
 10 of the methanol plant and other organizations in related fields [43]. However, no field data are used to  
 11 verify the accident scenarios. Therefore, we have  $x_3 = 1$ .

12 • The scaling rule for appropriateness of analysis method ( $x_4$ ) is listed in Table A.4. The QRA is mainly  
 13 based on Event Tree Analysis (ETA), which assumes that the occurrence probabilities of the intermediate  
 14 events do not change with time. In practice, however, the safety barriers of the methanol plant might be  
 15 time-dependent due to the degradation of critical components or systems. ETA is not able to capture such  
 16 time-dependence. Therefore, we have  $x_4 = 0$ .

17 • The scaling rule for quality of input data ( $x_5$ ) is listed in Table A.5. The input data used in the QRA  
 18 come from expert judgements based on handbooks (parts count reliability prediction standards and TNO  
 19 “purple book” [43]). Epistemic uncertainty in the expert judgements on the input data is not considered.  
 20 Therefore, we have  $x_5 = 0$ .

21 • The scaling rule for accuracy of risk calculation ( $x_6$ ) is listed in Table A.6. The risk calculation is based  
 22 on ETA, which is an analytical method. Therefore, its accuracy can be ensured and we have  $x_6 = 2$ .

23 By running the NBC with the input feature vector  $\mathbf{x} = [1, 2, 1, 0, 0, 2]$ , we can calculate the posterior  
 24 probabilities from (4), as shown in Figure 12. We can conclude that  $T = 1$  for the QRA of the methanol  
 25 plant, which means, according to Table I, that the QRA of the methanol plant is reliable but invalid. Such a  
 26 QRA can be used to support decision making, but not for safety-critical decisions.

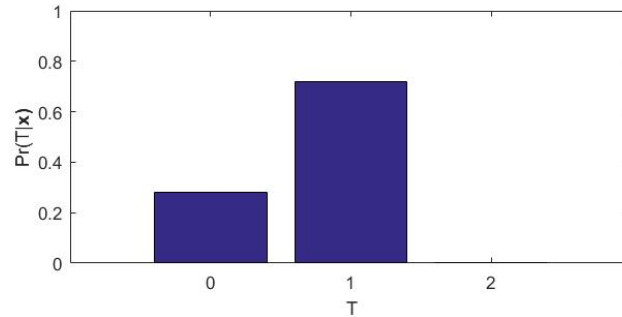


Fig. 12. Posterior probabilities for each value of  $T$

1

## VI. CONCLUSION

2 In this paper, a classification-based framework is developed for the trustworthiness assessment of QRA. In the  
 3 developed framework, trustworthiness is assessed in terms of six criteria, i.e., completeness of documentations,  
 4 understanding of problem settings, coverage of accident scenarios, appropriateness of analysis methods, quality  
 5 of input data, accuracy of risk calculation. For each criteria, three levels are distinguished and corresponding  
 6 scaling rules are presented for the assessment. The trustworthiness of QRA is also divided into three discrete  
 7 levels, based on its reliability and validity. A Naive Bayes Classifier (NBC) is constructed for trustworthiness  
 8 assessment given the values of the six criteria. **The developed NBC is able to learn from the training data the**  
 9 **expert's behavior when assessing the trustworthiness. Therefore, once accurately constructed, the NBC is able**  
 10 **to reasonably approximate the expert's assessment of the trustworthiness.** A stochastic hypothesis testing-based  
 11 approach is also developed to check the consistency of the training data. The performance of the developed  
 12 methods are tested by ten simulated case studies. The results demonstrate that the developed framework can  
 13 accurately approximate various expert assessment behaviors. Finally, a real application has been considered.

14 An inherent assumption of NBC is that the features (criteria) are independent among one another. In practice,  
 15 however, various dependencies might exist among the features. How to consider these dependencies needs  
 16 to be explored in future researches in order to derive a more accurate classifier. Furthermore, uncertainties  
 17 might affect the evaluation of the criteria  $x_1, x_2, \dots, x_6$ , since their assessments involve a lot of subjective  
 18 judgements. Future researches can be carried out to integrate these uncertainties in the assessment framework,  
 19 with mathematical theories dealing with subjective uncertainties, e.g., evidence theory [44], possibility theory  
 20 [45], uncertainty theory [46], etc.

21

## REFERENCES

- 22 [1] G. E. Apostolakis, "How useful is quantitative risk assessment?," *Risk analysis*, vol. 24, no. 3, pp. 515–520,  
 23 2004.
- 24 [2] E. Zio and T. Aven, "Industrial disasters: Extreme events, extremely rare. some reflections on the treatment  
 25 of uncertainties in the assessment of the associated risks," *Process Safety and Environmental Protection*,  
 26 vol. 91, no. 1, pp. 31–45, 2013.

- 1 [3] N. Khakzad, G. Reniers, R. Abbassi, and F. Khan, "Vulnerability analysis of process plants subject to  
2 domino effects," *Reliability Engineering & System Safety*, vol. 154, pp. 127–136, 2016.
- 3 [4] M. Abimbola, F. Khan, and N. Khakzad, "Risk-based safety analysis of well integrity operations," *Safety  
4 Science*, vol. 84, pp. 149–160, 2016.
- 5 [5] F. Goerlandt and J. Montewka, "Maritime transportation risk analysis: Review and analysis in light of  
6 some foundational issues," *Reliability Engineering & System Safety*, vol. 138, pp. 115–134, 2015.
- 7 [6] E. Zio and N. Pedroni, "Estimation of the functional failure probability of a thermalhydraulic passive  
8 system by subset simulation," *Nuclear Engineering and Design*, vol. 239, no. 3, pp. 580–599, 2009.
- 9 [7] F. Khan, S. J. Hashemi, N. Paltrinieri, P. Amyotte, V. Cozzani, and G. Reniers, "Dynamic risk management:  
10 a contemporary approach to process safety management," *Current Opinion in Chemical Engineering*,  
11 no. 14, pp. 9–17, 2016.
- 12 [8] F. Khan, S. Rathnayaka, and S. Ahmed, "Methods and models in process safety and risk management:  
13 Past, present and future," *Process Safety and Environmental Protection*, vol. 98, pp. 116–147, 2015.
- 14 [9] F. I. Khan and S. A. Abbasi, "Ophazopan effective and optimum approach for hazop study," *Journal of  
15 Loss Prevention in the Process Industries*, vol. 10, no. 3, pp. 191–204, 1997.
- 16 [10] H. Yu, F. Khan, and B. Veitch, "A flexible hierarchical bayesian modeling technique for risk analysis of  
17 major accidents," *Risk analysis*, 2017.
- 18 [11] N. Khakzad, F. Khan, and P. Amyotte, "Dynamic safety analysis of process systems by mapping bow-tie  
19 into bayesian network," *Process Safety and Environmental Protection*, vol. 91, no. 1-2, pp. 46–53, 2013.
- 20 [12] T. Rosqvist, "On the validation of risk analysis: A commentary," *Reliability Engineering & System Safety*,  
21 vol. 95, no. 11, pp. 1261–1265, 2010.
- 22 [13] A. Rae, R. Alexander, and J. McDermid, "Fixing the cracks in the crystal ball: A maturity model for  
23 quantitative risk assessment," *Reliability Engineering & System Safety*, vol. 125, pp. 67–81, 2014.
- 24 [14] H. J. Paskan, S. Jung, K. Prem, W. J. Rogers, and X. Yang, "Is risk analysis a useful tool for improving  
25 process safety?," *Journal of Loss Prevention in the Process Industries*, vol. 22, no. 6, pp. 769–777, 2009.
- 26 [15] J. Suokas and V. Rouhiainen, "Quality control in safety and risk analysis," *Journal of Loss Prevention in  
27 Process Industries*, vol. 2, pp. 67–77, 1989.
- 28 [16] T. Aven and E. Zio, "Foundational issues in risk assessment and risk management," *Risk Analysis*, vol. 34,  
29 no. 7, pp. 1164–1172, 2014.
- 30 [17] F. Goerlandt and P. Kujala, "On the reliability and validity of shipship collision risk analysis in light of  
31 different perspectives on risk," *Safety Science*, vol. 62, pp. 348–365, 2014.
- 32 [18] F. Goerlandt, N. Khakzad, and G. Reniers, "Validity and validation of safety-related quantitative risk  
33 analysis: A review," *Safety Science*, 2016.
- 34 [19] J. Graham, "Verifiability isn't everything," *Risk Analysis*, vol. 15, p. 109, 1995.
- 35 [20] J. S. Busby, R. E. Alcock, and E. J. Hughes, "Credibility in risk assessment," in *Probabilistic Safety  
36 Assessment and Management*, Probabilistic Safety Assessment and Management, pp. 2809–2814, Springer,  
37 2004.
- 38 [21] K. Lauridsen, M. Christou, A. Amendola, F. Markert, I. Kozine, and M. Fiori, "Assessing the uncertainties  
39 in the process of risk analysis of chemical establishments: Part 1 and 2," in *Proceedings of European*

- 1 *Conference on Safety and Reliability (ESREL)*, European Conference on Safety and Reliability (ESREL),  
2 (Torino, Italy), pp. 592–606, 2001.
- 3 [22] D. Sornette, T. Maillart, and W. Krger, “Exploring the limits of safety analysis in complex technological  
4 systems,” *International Journal of Disaster Risk Reduction*, vol. 6, pp. 59–66, 2013.
- 5 [23] B. J. Garrick, “On pra quality and use,” tech. rep., School of Engineering and Applied Science, University  
6 of Colifornia at Los Angeles, 1982.
- 7 [24] T. Aven and E. Zio, “Model output uncertainty in risk assessment,” *International Journal of Performability  
8 Engineering*, vol. 9, no. 5, pp. 475–486, 2013.
- 9 [25] T.-R. Wang, N. Pedroni, and E. Zio, “Identification of protective actions to reduce the vulnerability  
10 of safety-critical systems to malevolent acts: a sensitivity-based decision-making approach,” *Reliability  
11 Engineering & System Safety*, vol. 147, pp. 9–18, 2016.
- 12 [26] T. Aven and B. Heide, “Reliability and validity of risk analysis,” *Reliability Engineering & System Safety*,  
13 vol. 94, no. 11, pp. 1862–1868, 2009.
- 14 [27] P. A. Fenner-Crisp and V. L. Dellarco, “Key elements for judging the quality of a risk assessment,”  
15 *Environmental health perspectives*, 2016.
- 16 [28] V. Rouhiainen, “Importance of the quality management of safety analysis,” *Reliability Engineering &  
17 System Safety*, vol. 40, no. 1, pp. 5–16, 1993.
- 18 [29] A. Pinto, R. A. Ribeiro, and I. L. Nunes, “Ensuring the quality of occupational safety risk assessment,”  
19 *Risk analysis*, vol. 33, no. 3, pp. 409–419, 2013.
- 20 [30] E. Vergison, “A quality-assurance guide for the evaluation of mathematical models used to calculate the  
21 consequences of major hazards,” *Journal of Hazardous Materials*, 1996.
- 22 [31] V. Rouhiainen, “Quasa: A method for assessing the quality of safety analysis,” *Safety Science*, vol. 15,  
23 no. 3, pp. 155–172, 1992.
- 24 [32] “Regulatory guide 1.200: An approach for determining the technical adequacy of probabilistic risk  
25 assessment results for risk-informed activities,” tech. rep., US Nuclear Regulatory Commission, 2009.
- 26 [33] T. R. Wang, V. Mousseau, N. Pedroni, and E. Zio, “Assessing the performance of a classificationbased  
27 vulnerability analysis model,” *Risk Analysis*, vol. 35, no. 9, pp. 1674–1689, 2015.
- 28 [34] C. Bishop, *Pattern Recognition and Machine Learning*. London: Springer, 2006.
- 29 [35] A. McCallum and K. Nigam, “A comparison of event models for naive bayes text classification,” in  
30 *AAAI-98 workshop on learning for text categorization*, vol. 752 of *AAAI-98 workshop on learning for text  
31 categorization*, pp. 41–48, Citeseer, 1998.
- 32 [36] I. Kononenko, “Inductive and bayesian learning in medical diagnosis,” *Applied Artificial Intelligence an  
33 International Journal*, vol. 7, no. 4, pp. 317–337, 1993.
- 34 [37] T.-R. Wang, V. Mousseau, N. Pedroni, and E. Zio, “An empirical classification-based framework for the  
35 safety criticality assessment of energy production systems, in presence of inconsistent data,” *Reliability  
36 Engineering & System Safety*, vol. 157, pp. 139–151, 2017.
- 37 [38] Z. Zeng, F. Di Maio, E. Zio, and R. Kang, “A hierarchical decision making framework for the assessment of  
38 the prediction capability of prognostic methods,” *Proceedings of the Institution of Mechanical Engineers,  
39 Part O: Journal of Risk and Reliability*, 2016. Accepted.

- 1 [39] “Naive bayes, guassian distributions and practical applications.” [http://http://www.cs.cmu.edu/~tom/](http://http://www.cs.cmu.edu/~tom/10601_sp09/lectures/NBayes2_2-2-2009-ann.pdf)  
2 [10601\\_sp09/lectures/NBayes2\\_2-2-2009-ann.pdf](http://http://www.cs.cmu.edu/~tom/10601_sp09/lectures/NBayes2_2-2-2009-ann.pdf). Accessed: 2016-10-24.
- 3 [40] T. L. Saaty, “A scaling method for priorities in hierarchical structures,” *Journal of mathematical psychology*,  
4 vol. 15, no. 3, pp. 234–281, 1977.
- 5 [41] Y. Chen, Z. Zeng, and R. Kang, “Validation methodology for distribution-based degradation model,”  
6 *Systems Engineering and Electronics, Journal of*, vol. 23, no. 4, pp. 553–559, 2012.
- 7 [42] J. L. Devore, *Probability and statistics for engineering and the sciences*. Boston, MA :: Cengage Learning,,  
8 8th ed. ed., 2010.
- 9 [43] “Quantitative risk assessment: Final report (prepared for nw innovation works, by  
10 acutech consulting group).” [http://http://kalamamfgfacilitysepa.com/wp-content/uploads/2016/09/](http://http://kalamamfgfacilitysepa.com/wp-content/uploads/2016/09/FEIS-Appendix-G1-Quantitative-Risk-Assessment.pdf)  
11 [FEIS-Appendix-G1-Quantitative-Risk-Assessment.pdf](http://http://kalamamfgfacilitysepa.com/wp-content/uploads/2016/09/FEIS-Appendix-G1-Quantitative-Risk-Assessment.pdf). Accessed: 2016-11-08.
- 12 [44] G. Shafer, *A mathematical theory of evidence*, vol. 1. Princeton university press Princeton, 1976.
- 13 [45] L. A. Zadeh, “Fuzzy sets as a basis for a theory of possibility,” *Fuzzy sets and systems*, vol. 1, pp. 3–28,  
14 1978.
- 15 [46] B. Liu, *Uncertainty Theory, Fifth Edition*. Springer, 2015.

APPENDIX  
SCALING RULES FOR  $x_1$ - $x_6$

TABLE A.1  
SCALING RULES FOR  $x_1$

Levels	Descriptions
$x_1 = 0$	Some the following elements are missing in the documentations: <ul style="list-style-type: none"> <li>• scopes and objectives of the QRA;</li> <li>• descriptions of the system under investigation and related references;</li> <li>• accounts of the adopted analysis methods;</li> <li>• presentation of source data needed for the analysis;</li> <li>• report of the analysis results.</li> </ul>
$x_1 = 1$	At least one of the following flaws present in the documentations: <ul style="list-style-type: none"> <li>• descriptions of scopes and objectives are incomplete or ambiguous;</li> <li>• descriptions of the system under investigation are unclear;</li> <li>• no sufficient references on the system under investigation are given;</li> <li>• descriptions of the adopted methods are unclear;</li> <li>• presentations of the results are incomplete (e.g., no uncertainty is considered) or ambiguous.</li> </ul>
$x_1 = 2$	The documentation of the QRA process contains sufficient information for its repetition: <ul style="list-style-type: none"> <li>• the documentation contains all the necessary parts;</li> <li>• no flaws in level <math>x_1 = 1</math> present.</li> </ul>

TABLE A.2  
SCALING RULES FOR  $x_2$

Levels	Descriptions
$x_2 = 0$	The analysts are unaware of the problem settings of the QRA due to the presence of all the following flaws: <ul style="list-style-type: none"> <li>• the purposes of the QRA are not clearly understood;</li> <li>• the systems of interests are not well defined;</li> <li>• the resources constraints (e.g., time, computational resources, etc) are not clearly defined.</li> </ul>
$x_2 = 1$	The analysts misunderstand part of the problem settings due to some of the following flaws: <ul style="list-style-type: none"> <li>• the analysts fail to clearly understand the purposes of the QRA;</li> <li>• the analysts fail to clearly understand the systems of interests;</li> <li>• the analysts fail to clearly understand the resources constraints.</li> </ul>
$x_2 = 2$	The analysts clearly understand the problem settings: no flaws in levels $x_2 = 0$ and $x_2 = 1$ occur.



TABLE A.3  
SCALING RULES FOR  $x_3$

Levels	Descriptions
$x_3 = 0$	Some critical accident scenarios are highly likely to be missed by the identification process: <ul style="list-style-type: none"> <li>the coverage of the identified accident scenarios is not validated;</li> <li>the validation shows that some critical accident scenarios might be missing.</li> </ul>
$x_3 = 1$	<ul style="list-style-type: none"> <li>Validation reveals that most critical accident scenarios are covered by the identification process but</li> <li>the validation is conducted based on peer review rather than using real data.</li> </ul>
$x_3 = 2$	<ul style="list-style-type: none"> <li>Validation reveals that most critical accident scenarios are covered by the identification process and</li> <li>the validation is reliable by using field data.</li> </ul>

TABLE A.4  
SCALING RULES FOR  $x_4$

Levels	Descriptions
$x_4 = 0$	The features of the selected analysis method cannot satisfy the requirements of the problem: <ul style="list-style-type: none"> <li>the methods cannot capture some features of the problem (e.g., time-dynamics, dependencies, etc.) or</li> <li>the methods require more resources (e.g., data, computational power, etc.) than that can be provided.</li> </ul>
$x_4 = 1$	<ul style="list-style-type: none"> <li>The features of the selected analysis method satisfy the requirements of the problem but</li> <li>the conclusion is drawn based on expert experience.</li> </ul>
$x_4 = 2$	<ul style="list-style-type: none"> <li>The features of the selected analysis method satisfy the requirements of the problem and</li> <li>successful applications in similar problems can justify the choice of the method.</li> </ul>

TABLE A.5  
SCALING RULES FOR  $x_5$

Levels	Descriptions
$x_5 = 0$	<ul style="list-style-type: none"> <li>There is no sufficient statistical data and the input data is purely based on expert judgements;</li> <li>epistemic uncertainty in the expert-generated input data is not considered.</li> </ul>
$x_5 = 1$	<ul style="list-style-type: none"> <li>There is no sufficient statistical data;</li> <li>the input data are based on expert judgements with fully consideration of epistemic uncertainty.</li> </ul>
$x_5 = 2$	<ul style="list-style-type: none"> <li>Sufficient statistical data can be used for risk analysis.</li> </ul>

TABLE A.6  
SCALING RULES FOR  $x_6$

Levels	Descriptions
$x_6 = 0$	<ul style="list-style-type: none"> <li>The process of risk calculation contains major flaws;</li> <li>large errors might exist in the calculated risks.</li> </ul>
$x_6 = 1$	<ul style="list-style-type: none"> <li>The process of risk calculation does not contain major flaws;</li> <li>only errors from the calculation process itself (e.g., the accuracy of Monte Carlo simulations) might exist but</li> <li>the uncertainties caused by the errors are not modeled.</li> </ul>
$x_6 = 2$	<ul style="list-style-type: none"> <li>Only errors from the calculation process itself (e.g., the accuracy of Monte Carlo simulations) might exist and</li> <li>the uncertainties caused by the errors are properly modeled.</li> </ul>