Delft University of Technology

Delft Center for Systems and Control

Technical report 17-002

Bayesian and Dempster-Shafer reasoning for knowledge-based fault diagnosis – A comparative study*

K. Verbert, R. Babuška, and B. De Schutter

If you want to cite this report, please use the following reference instead:
K. Verbert, R. Babuška, and B. De Schutter, "Bayesian and Dempster-Shafer reasoning for knowledge-based fault diagnosis – A comparative study," *Engineering Applications of Artificial Intelligence*, vol. 60, pp. 136–150, Apr. 2017.

Delft Center for Systems and Control
Delft University of Technology
Mekelweg 2, 2628 CD Delft
The Netherlands
phone: +31-15-278.24.73 (secretary)
URL: https://www.dcsc.tudelft.nl

Bayesian and Dempster-Shafer Reasoning for Knowledge-Based Fault Diagnosis – A Comparative Study

K. Verbert[a,*], R. Babuška[a], B. De Schutter[a]

[a]*Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands*

Abstract

Even though various frameworks exist for reasoning under uncertainty, a realistic fault diagnosis task does not fit into any of them in a straightforward way. For each framework, only part of the available data and knowledge is in the desired format. Moreover, additional criteria, like clarity of inference and computational efficiency, require trade-offs to be made. Finally, fault diagnosis is usually just a subpart of a larger process, e.g. condition-based maintenance. Consequently, the final goal of fault diagnosis is not (just) decision making, and the outcome of the diagnosis process should be a suitable input for the subsequent reasoning process. In this chapter, we analyze how a knowledge-based diagnosis task is influenced by uncertainty, investigate which additional objectives are of relevance, and compare how these characteristics and objectives are handled in two well-known frameworks, namely the Bayesian and the Dempster-Shafer reasoning framework. In contrast to previous works, which take the reasoning method as the starting point, we start from the application, knowledge-based fault diagnosis, and examine the effectiveness of different reasoning methods for this specific application. It is concluded that the suitability of each reasoning method highly depends on the problem under consideration and on the requirements of the user. The best framework can only be assigned given that the problem (including uncertainty characteristics) and the user requirements are completely known.

Keywords: fault diagnosis; uncertainty reasoning; Bayesian inference; Dempster-Shafer inference; condition-based maintenance.

1. Introduction

Condition-based maintenance is a promising preventive maintenance strategy to reduce system downtime and costs. An important task within the condition-based maintenance process is the determination of the actual system health based on measurement data, hereafter referred to as "fault diagnosis". In practice, fault diagnosis is a challenging task, among other things, due to the presence of uncertainty. Especially for safety-critical systems, like medical devices, railway systems, and nuclear reactors, is it important to deal with the uncertainty in an adequate way.

Although a lot of research has been devoted to fault diagnosis, relatively little attention has been paid to the consequences of uncertainty. Many existing methods account for part of the uncertainty, e.g. methods based on Kalman filters [1–4] or methods based on set-membership approaches [5, 6]. Such methods however adopt strong assumptions regarding the type of uncertainty present, and require that the system can be described by a specific model, often a linear state space model. Besides, data-based methods, e.g. methods based on neural-networks [7, 8], have been proposed that may implicitly account for various types of uncertainty. However, such methods are, in general, not able to clearly express the uncertainty in the diagnostic result, yielding that the uncertainty cannot be adequately accounted for in the subsequent decision making process.

Because of the aforementioned drawbacks of existing methods with respect to uncertainty handling, in this paper we focus on uncertainty reasoning for *knowledge-based fault diagnosis*. Knowledge-based diagnosis is considered because in many practical applications not enough knowledge is available to define a quantitative model required by model-based approaches. Knowledge-based fault diagnosis is influenced by uncertainty in various ways: First, the available measurement data may be incomplete, incorrect, or imprecise, e.g. due to sensors with a limited accuracy; Second, knowledge is needed to infer system health from these uncertain data. Also this knowledge is generally uncertain, i.e. (partly) incorrect, subjective, or incomplete.

Despite of the development of various methods for reasoning under uncertainty and the many discussions about the correctness and usefulness of these methods [9–16], no agreement has been reached regarding a consistent and uniform framework to handle problems under uncertainty. In particular the disagreement about the correctness and usefulness of the Bayesian and the Dempster-Shafer frame-

*Corresponding author

Email addresses: `k.a.j.verbert@tudelft.nl` (K. Verbert), `r.babuska@tudelft.nl` (R. Babuška), `b.deschutter@tudelft.nl` (B. De Schutter)

work has led to debates. Bayesian proponents claim that the Bayesian theory is the optimal framework to handle all kinds of uncertainty (see e.g. [9, 10]). To quote Dennis Lindley, an eminent probabilist [17], "probability is the only sensible description of uncertainty and is adequate for all problems involving uncertainty. All other methods are inadequate" and "anything that can be done with fuzzy logic, belief functions, upper and lower probabilities, or any other alternative to probability can better be done with probability." While Bayesian proponents are convinced about their framework, shortcomings are claimed by many researchers (see e.g. [11–16, 18, 19]). For example, the authors of [11, 12, 18, 19] argue for the need of belief functions and for their added value over probabilities. Especially, they promote belief functions for being superior in representing incomplete and partially reliable knowledge. In [13] it is concluded that the Bayesian approach is tailored for decision making, but not necessarily for other kinds of reasoning. The authors of [14, 15] consider different sources of uncertainty, all having their own characteristics, and they argue that each of these uncertainty sources requires another reasoning strategy. In contrast, [16] advocates that the Bayesian and Dempster-Shafer frameworks have roughly the same expressive power.

In this paper, we compare Bayesian and Dempster-Shafer reasoning from an application-oriented point of view. In contrast to previous works, which take the reasoning method as the starting point and use examples to illustrate the effectiveness of the method, we start from the application, i.e. knowledge-based fault diagnosis, and examine the effectiveness of different reasoning methods for this specific application. More specifically, the contributions of this paper are:

1. We analyze how the available data and knowledge are influenced by uncertainty;
2. We compare how the knowledge-based fault diagnosis task fits within the Bayesian and Dempster-Shafer reasoning framework;
3. We present additional objectives (e.g. clarity of inference) and analyze how they are accounted for in both reasoning frameworks.

Note that our aim is not to deeply discuss uncertainty methods nor to advocate one of the methods in general. We focus on a specific problem with the related objectives, for which we assess under which circumstances which method is most suitable to reach these objectives.

Note that this paper is an improved and extended version of our conference paper [20]. In particular, the current paper adds the following elements: a thorough analysis of the knowledge-based fault diagnosis problem in both the Bayesian and the Dempster-Shafer framework, as well as a more extensive comparison and example.

The remainder of this paper consists of three parts: The first part (Section 2 till Section 4) discusses general concepts regarding reasoning under uncertainty. In the second part (Sections 5 till 7), we analyze the uncertain reasoning problem of knowledge-based fault diagnosis. The third part (Section 8) covers a specific fault diagnosis example for railway track circuits.

2. Classification of uncertainty

According to e.g. [13–15, 17] various sources of uncertainty need to be treated differently. A distinction is made between the following sources of uncertainty:

1. Randomness;
2. Incompleteness;
3. Imprecision;
4. Conflict.

Randomness, also called intrinsic variability, refers to the situation that a future outcome is uncertain, but a probability distribution of the outcome is available, e.g. throwing a known fair die. *Incompleteness* means that an outcome (or probability distribution) is defined, but the information available is not sufficient to identify this outcome (or probability distribution). For example, the evidence that the winner of a competition is a male is only sufficient to identify the winner in the case that there is only one male candidate winner. Otherwise, this evidence only allows to exclude candidate female winners. *Imprecision* refers to the situation that the outcome is known, but with finite precision. For example, we know that the current outside temperature is between 25.5 and 26.5 degrees Celsius. Finally, uncertainty can arise due to (partially) *conflicting* information. For example, two experts give a different answer to a particular question.

For reasoning purposes, uncertainty is often classified into the following two classes [21, 22]:

1. Aleatory uncertainty;
2. Epistemic uncertainty.

Aleatory uncertainty, also called statistical uncertainty, represents intrinsic variability – i.e. the differences that are observed each time the same experiment is repeated. *Epistemic uncertainty*, also called systematic uncertainty, arises due to a lack of knowledge. This is the uncertainty about things that we could in principle know, but in practice we do not know. The two are often distinguished using the fact that epistemic uncertainty can be reduced by gathering more knowledge or more data, whereas aleatory uncertainty cannot be reduced [14, 21]. To illustrate this, consider the example of throwing a die. When we throw a die of which we know the underlying model, each time we get a different outcome, but throwing it more often will not provide information to reduce uncertainty about the outcome of a future throw. So, the uncertainty referred to is of the aleatory type. In contrast, when we throw an unknown die and we want to construct a probabilistic model of the outcome of a throw, then the more data we gather, the less uncertainty we have in our model. Here, the uncertainty referred to is of the epistemic type. Ideally, we would like to eliminate all epistemic uncertainty, so that only aleatory uncertainty remains. In practice, which part of the uncertainty actually can be reduced depends on the particular problem, practical constraints, and the assumptions adopted [21].

Considering the different uncertainty sources: both imprecision, incompleteness, and conflict refer to a lack of knowledge and they can be regarded as epistemic uncertainty, whereas randomness can be regarded as aleatory uncertainty.

3. Methods for reasoning under uncertainty – An overview

For completeness and to make a link between the different uncertainty sources and the different reasoning frameworks, in this section, we briefly introduce four common frameworks for reasoning under uncertainty, namely the Bayesian framework, the Dempster-Shafer framework, possibility theory, and fuzzy logic. Later on in Section 4, we motivate our choice to focus on Bayesian and Dempster-Shafer reasoning in this paper. Extensive discussions of the frameworks compared in this work, i.e. Bayesian and Dempster-Shafer reasoning, can be found in Appendix A and Appendix B respectively.

3.1. Notation

We denote a variable by an upper-case letter (e.g. X, Y). A variable X can take values in its domain Θ_X. A particular element of Θ_X is denoted by x_i and a subset of Θ_X is denoted by x. A set of variables is denoted by a bold-face upper-case letter (e.g. $\mathbf{U}$, $\mathbf{V}$) and the assignment of a value to each variable in the set by the corresponding bold-face lower-case letter ($\mathbf{u}$, $\mathbf{v}$).

3.2. Bayesian probability theory

Probability theory [23, 24] is an established and well-known framework for reasoning under uncertainty. Roughly there are two interpretations of probability [25]: the Bayesian and frequentist interpretation. Here, the focus is on the (subjective) Bayesian approach. Whereas frequentists only use data, Bayesians use data to improve their initial belief, i.e. "initial belief" + "data" = "improved belief". The combination of these two is beneficial in situations where relatively little data and a reasonable amount of prior knowledge are available [26]. Technical details regarding reasoning in Bayesian networks can be found in Appendix A.

3.3. Dempster-Shafer framework

The Dempster-Shafer (D-S) framework [27–29] was developed to handle incomplete information. This is realized by allowing the assignment of belief to *sets of elements in the domain* instead of assigning belief only to individual elements, like in the Bayesian framework. Different interpretations of the D-S theory exist, among which are the upper and lower probabilities model and the evidentiary value model [12]. In this work, we adopt Smets' well-known Transferable Belief Model (TBM) interpretation [30]. Technical details regarding reasoning in the TBM can be found in Appendix B.

3.4. Possibility theory

Another way to handle incomplete information is using possibility theory [15, 31, 32]. Instead of assigning one probability to each individual element in the domain, like in the Bayesian framework, possibility theory uses two values: a possibility value and a necessity value, making it possible to represent incomplete information [33]. The possibility of an event is equal to zero if and only if its negation is known to be true, and is equal to one otherwise. The necessity of an event is equal to one if and only if the event is known to be true. In practice, this binary representation is often not entirely satisfactory and a graded notion of possibility theory is used (see e.g. [33]).

3.5. Fuzzy logic

The fuzzy logic framework [17, 34, 35] was developed to handle perception-based information. Perception-based information is imprecise and cannot be represented by a single number. In fuzzy logic, everything is, or is allowed to be, graduated [36]. So in this sense, a proposition can be partially true. Consider for example the proposition "The room temperature is very high". In standard logic, this proposition is true or false. In fuzzy logic, this proposition can be true with a degree between 0 and 1.

4. Relation between uncertainty sources and reasoning frameworks

In Section 2, we have discussed sources of uncertainty and in Section 3 various reasoning frameworks have been mentioned. The question that remains is "How do these relate to each other?". In this section we give a brief overview of these relations. Moreover, we motivate our choice to focus on Bayesian and Dempster-Shafer reasoning in this paper.

4.1. Overview

According to Bayesian proponents, probabilities are suited to handle all kinds of uncertainty, which is precisely the advantage of the Bayesian approach [37]. According to non-Bayesians [13, 14], probabilities are suited to handle aleatory uncertainty, but are not suited to handle epistemic uncertainty. Fuzzy set theory [38] has been proposed to handle imprecise information, and possibility theory [31, 32] and the theory of belief functions [27, 28] have been proposed to handle incomplete information. An overview of these relations is given in Table 1. Note that in Table 1, for each uncertainty class, it is indicated which framework is particularly tailored to handle uncertainty from this class. This does however not mean that the other frameworks cannot be used to handle uncertainty from that particular class. These methods may however be less efficient or less accurate.

Table 1: Uncertainty classifications and reasoning frameworks.

Uncertainty class	Uncertainty source	Reasoning framework
Aleatory	Randomness	Bayesian probability theory
Epistemic	Incompleteness	D-S theory, possibility theory
	Imprecision	Fuzzy logic
	Conflict	D-S theory

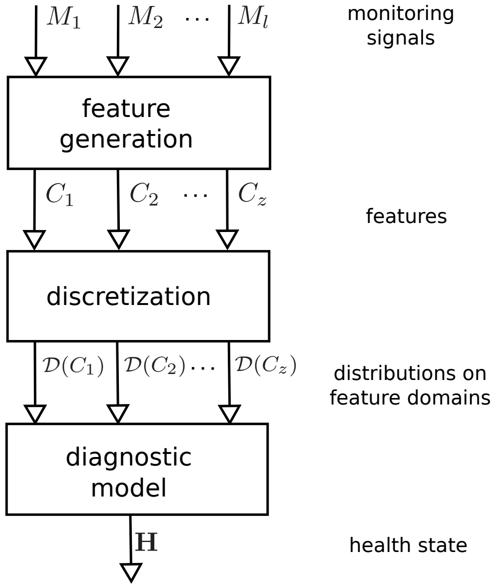


Figure 1: Overview knowledge-based diagnosis.

4.2. Motivation for Bayesian and Dempster-Shafer reasoning

For clarity, in the remainder of this paper we focus on two frameworks only, namely the Bayesian and the Dempster-Shafer framework. We have chosen for these two frameworks because a knowledge-based fault diagnosis problem is often subject to randomness and incompleteness. Since the Dempster-Shafer framework is particularly suited to handle incompleteness, and the Bayesian framework is particularly suited to handle randomness, and, according to Bayesian proponents, also to handle incompleteness, these two frameworks are a natural choice for knowledge-based fault diagnosis. Note that we could also have opted to consider possibility theory because of its ability to handle incompleteness. However, for fault diagnosis, we prefer the representation as used in the D-S framework (i.e. belief functions) over the representation used by the possibility theory (i.e. possibility values).

5. Knowledge-based fault diagnosis

5.1. Overview

Fault diagnosis comprises the determination of the cause(s) of any faulty system behavior. This paper considers knowledge-based fault diagnosis, which is a model-based diagnosis strategy that uses knowledge to define the diagnostic model[1] in the form of a qualitative model or a rule-based system [40]. Figure 1 gives an overview of the knowledge-based fault diagnosis process. The monitoring signals M_1 till M_l serve as input for the diagnosis and the output is the system health represented by a set $\mathbf{H}$ of variables, indicating whether or not the system is healthy, and if not, what actually causes the faulty behavior. To determine the system health, first, characteristic features C_1 till C_z are extracted from the monitoring signals. Next, the values of features C_1 till C_z are determined and, in the presence of uncertainty, represented by distribution functions over the associated domains $\Theta_{C_1} = \{c_{1,1}, c_{1,2}, ..., c_{1,k_1}\}$ till $\Theta_{C_z} = \{c_{z,1}, c_{z,2}, ..., c_{z,k_z}\}$. The type of distribution function depends on the reasoning framework used for the fault diagnosis, e.g. in the Bayesian framework, a probability distribution is used, while in the D-S framework, a D-S belief function is used (see Appendix A and Appendix B for more details regarding the different distribution functions). Finally, based on the distributions over the feature domains, the presence and type of faults is inferred by using the diagnostic model.

So, the reasoning task of knowledge-based fault diagnosis is the determination of the system health based on the values of the features C_1 till C_z. Therefore, we distinguish between two groups of variables:

1. The set $\mathbf{C}$ of observable variables $\mathbf{C} = \{C_1, .., C_z\}$;
2. The set $\mathbf{H}$ of target variables representing the system health.

Assuming that there are ℓ different fault causes f_1 till f_ℓ, the system health is represented by one $(\ell+1)$-valued variable H with $\Theta_H = \{h, f_1, .., f_\ell\}$ or by ℓ two-valued variables F_1 till F_ℓ all taking on values in the set $\{0, 1\}$, indicating the absence (0) or presence (1) of the respective fault cause f_j. Generally, the first option is preferred when only single-fault scenarios are considered, while the second option is used when also multiple-fault scenarios are taken into account. A combination of the two can be used when only part of the faults can occur simultaneously. Unless otherwise stated, we allow multiple-fault scenarios and use one binary variable for each possible fault cause.

[1] A diagnostic model is a set of static or dynamic relations that link specific input variables – the feature values – to specific output variables – the faults [39].

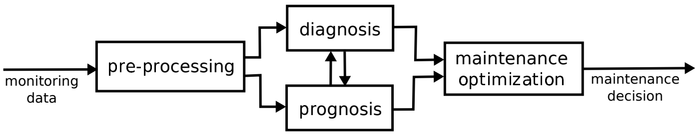


Figure 2: Fault diagnosis within the condition-based maintenance process.

5.2. Fault diagnosis within a condition-based maintenance framework

In general, fault diagnosis is not an isolated task, but it is part of a larger process. As reasoning under uncertainty and decision making in the presence of uncertainty impose different requirements on uncertainty characterization, it is important to consider the purpose(s) of fault diagnosis.

Often, fault diagnosis is performed as part of a condition-based maintenance process (see Figure 2). The final goal of condition-based maintenance is maintenance planning, i.e. deciding on the required maintenance activities. This is done based on the diagnosis and prognosis[2] result. The diagnosis outcome serves as an input for both the prognosis and the maintenance optimization step (see Figure 2). So, the final goal of condition-based maintenance is decision making. However, the main goal of fault diagnosis is reasoning about the system health based on (multiple) observed variables. Therefore, in this work, the main focus is on information fusion and reasoning under uncertainty, and less on decision making.

5.3. Uncertainty sources

As already indicated, uncertainty can originate from different sources. For knowledge-based fault diagnosis, we identify the following main sources of uncertainty:

1. Uncertainty arising from imperfect sensors;
2. Uncertainty regarding the relations between features and faults;
3. Uncertainty arising from the conversion from measurement data to the feature space.

More specifically, we characterize the above-mentioned uncertainty sources as follows:

5.3.1. Sensors

In general, sensors are imprecise (i.e. they have limited accuracy) and may suffer from structural errors (e.g. offsets, drift). Due to imperfect sensors, our assumed world differs from reality. Therefore, this type of uncertainty refers to a lack of knowledge. In that sense, the uncertainty can be reduced e.g. by calibrating sensors, implementing better sensors, or using additional sensors. In practice, the available sensors are generally fixed (and cannot be changed) and their precision is approximately known. In this case, the corresponding uncertainty is regarded as intrinsic variability, so it is of the aleatory type.

[2]The prognosis result is a prediction of how the system health will evolve over time.

5.3.2. Relations between features and faults

Here, two sources of uncertainty play a role. First uncertainty arises because the relations between features and faults are not completely deterministic due to unmodeled influences. Second, the available knowledge relating faults and features may be incomplete or imprecise. The latter reflects a lack of knowledge (epistemic uncertainty); the former is, for diagnosis purposes, generally regarded as aleatory uncertainty.

5.3.3. Conversion from measurement data to the feature space

Based on the monitoring signals, the features have to be determined. In general, a derived feature C_k does not behave exactly according to one element in its domain Θ_{C_k} (Bayesian framework) or to one element in the power set $2^{\Theta_{C_k}}$ (D-S framework). So, it has to be determined to what extent the observed behavior corresponds to each element of Θ_{C_k} or $2^{\Theta_{C_k}}$. The exact uncertainty characteristics depend on the system behavior and the way the behavior is evaluated, e.g. by subjective human judgment or mathematical (computer) calculations.

6. Reasoning under uncertainty for knowledge-based fault diagnosis

In this section, we discuss how the knowledge-based fault diagnosis problem is handled in the Bayesian and the D-S framework.

6.1. Bayesian networks

The considered knowledge-based fault diagnosis problem (see Section 5.1) is graphically represented by a Bayesian network such as the one shown in Figure 3. The edges indicate that fault f_1 has a direct influence on both feature C_1 and feature C_2, that both fault f_2 and fault f_3 influence feature C_3, and that feature C_z is influenced by fault f_ℓ.

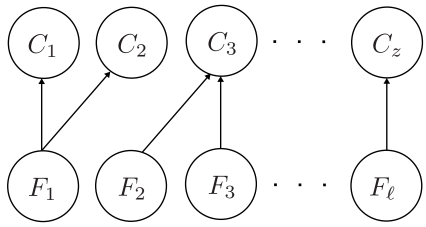


Figure 3: Bayesian network representation of the knowledge-based fault diagnosis problem. The variables F_1 till F_ℓ represent the different system faults and the variables C_1 till C_z are the diagnostic features.

Before the Bayesian network can be used for reasoning, the prior probability distributions of F_1 till F_ℓ (root nodes), and the conditional probability tables of C_1 till C_z need to be determined. The prior probabilities indicate the likelihood of a particular fault f_j, i.e. $P(F_j = 1)$, before any evidence is collected. The conditional probability table of C_i contains the probabilities of each feature value $c_{i,n}$ given the value of each parent of C_i. For example in Figure 3, feature C_3 has parents F_2 and F_3; so, for C_3, the conditional probability table as given in Table 2 needs to be defined.

Table 2: Example of a conditional probability table of C_3

		C_3			
F_2	F_3	$c_{3,1}$	$c_{3,2}$	$\dots$	c_{3,k_3}
0	0	$\dots$	$\dots$	$\dots$	$\dots$
0	1	$\dots$	$\dots$	$\dots$	$\dots$
1	0	$\dots$	$\dots$	$\dots$	$\dots$
1	1	$\frac{1}{k_3-1}$	$\frac{1}{k_3-1}$	$\dots$	0

Often, the available knowledge is not in probabilistic form, e.g. we are uncertain about the prior probability of F_j, or we are not sure about the conditional probability distribution $P(C_i|\mathbf{U}_{C_i})$ of feature C_i given the values of its parents $\mathbf{U}_{C_i}$. For example, we only know that given that $F_2 = F_3 = 1$, it holds that $P(C_3 = c_{3,k_3}) = 0$. In such case, the remaining probabilities are assigned according to the additivity axiom and the principle of maximum entropy [41, 42] (see e.g. the last row of Table 2). The *additivity axiom* states that $P(a) + P(\bar{a}) = 1$, and the *principle of maximum entropy* is a strategy in which missing probabilities are assigned such that the distribution is consistent with known constraints, but is otherwise as unbiased as possible.

After the Bayesian network is initialized, i.e. the structure G and the set of local probability functions D are defined, it can be used for reasoning. So, we can update the model based on evidences regarding the features C_1 till C_z (the observable variables) and compute the marginal probability distributions of the fault variables F_1 till F_ℓ (the target variables). When the available evidences are hard evidences[3], they can be easily propagated based on standard Bayesian inference algorithms (see Appendix A.3). When the available evidences are uncertain, it first needs to be assured that they are specified by likelihood ratios as required by Pearl's method of virtual evidence (see Appendix A.3). When the evidences are specified as probabilistic evidence[4], standard rules can be used for the conversion (see Appendix A.3). In practice, evidences are often specified by human experts, which do not necessarily follow the Bayesian laws. For example, a (partially) incomplete answer, like "the value of C_i is $c_{i,2}$ or $c_{i,4}$" is also plausible. Again, probabilistic information is derived from this incomplete information based on the principle of maximum entropy.

To summarize, knowledge-based fault diagnosis in Bayesian networks may require the following pre-processing steps to match the available information with the Bayesian format:

1. Transformation of the uncertain knowledge base (i.e. the relations between features and faults) into a set of conditional probability tables. Usually, the available knowledge is already conditional. Only missing probabilities in the case of incomplete information have to be estimated.
2. Determination of the prior fault probabilities.
3. Transforming the evidence into the format specified by the virtual evidence method, i.e. likelihood ratios (see Appendix A.3).

6.2. Dempster-Shafer belief networks

In the D-S framework, the considered knowledge-based fault diagnosis problem (see Section 5.1) is represented by a D-S valuation network (see Figure 4 for an example). The notation $\Theta_{A\times B\times C}$ is used as a shorthand for the multidimensional space $\Theta_A \times \Theta_B \times \Theta_C$.

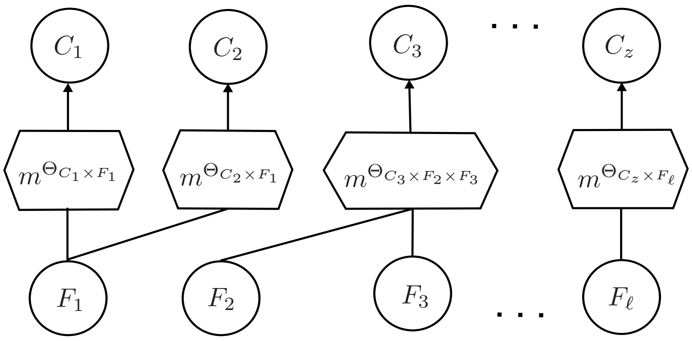


Figure 4: D-S valuation network representation of the knowledge-based fault diagnosis problem. The variables F_1 till F_ℓ represent the different system faults, the variables C_1 till C_z represent the diagnostic features, and the valuations (multivariate mass functions) $m^{\Theta_{C_1\times F_1}}$, $m^{\Theta_{C_2\times F_1}}$, $m^{\Theta_{C_3\times F_2\times F_3}}$, and $m^{\Theta_{C_z\times F_\ell}}$ express the (uncertain and incomplete) relationships between the respective variables.

Before the valuation network can be used for reasoning, the prior mass distributions of F_1 till F_ℓ and the multivariate mass functions describing the valuations (hexagons in Figure 4) need to be defined. The prior mass distributions indicate the likelihood of a particular fault before any evidence is collected. The important difference with the Bayesian analysis is that, in the D-S framework, the prior mass functions of F_1 till F_ℓ can be defined as vacuous mass functions, expressing total ignorance, i.e. $m^{\Theta_{F_j}}(\Theta_{F_j}) = 1$. The relationships between variables (valuations) need to be defined by multivariate mass functions on the product spaces of the domains of the connected variables. For example, the relation between C_3, F_2, and F_3 is characterized by a mass function on the space $\Theta_{C_3} \times \Theta_{F_2} \times \Theta_{F_3}$. A

[3]Hard (or certain) evidence for a variable X is evidence that states that X takes a particular value $x_i \in \Theta_X$.

[4]Probabilistic evidence for a variable X is specified by a probability distribution over Θ_X.

mass needs to be attached to each combination of possible values. For example, to capture the relation between F_1 and C_1, assuming that C_1 can take values in $\Theta_{C_1} = \{c_{1,1}, c_{1,2}\}$, the masses given in Table 3 need to defined. When all mass is assigned to the masses in the first column, the information available is complete, but possibly uncertain. The more mass is assigned to the masses in the right columns, the more ignorant we are. Note that even for a two-dimensional mass function with small domains, a large number of masses is needed to capture the available (incomplete) knowledge. In the worst case, 15 nonzero masses need to be assigned for the given example. In comparison, the Bayesian model requires only 4 conditional probabilities to be specified. These are the costs that have to be paid for including the possibility of expressing ignorance. For fault diagnosis, the information available is often specified in conditional form, in which case the joint masses are estimated by using the ballooning extension (see Appendix B.1).

A D-S valuation network is used for reasoning as follows: When new evidence becomes available, the network is updated according to Dempster's rule of combination (B.8). These evidences should be represented in the form of a mass function.

To summarize, knowledge-based fault diagnosis in D-S valuation networks may require the following pre-processing steps to match the available information with the D-S demands:

1. Transformation of the uncertain knowledge base into the desired format, i.e. multivariate mass functions on the joint domains. Usually, the knowledge is conditional and the joint distributions need to be estimated using the ballooning extension.
2. Transformation of the available evidences into mass functions.

7. Comparison and additional criteria

7.1. Diagnostic reasoning performance

From the analysis in Section 6, we conclude that the Bayesian model is particularly suited for reasoning about conditional relationships, like the relations between faults and features. In practice, the relationships between faults and features, as well as the available evidences are however not purely probabilistic, and approximations need to be made when using the Bayesian model. In contrast, the D-S model is perfectly suited to handle knowledge that is not purely probabilistic, e.g. incomplete or imprecise. The D-S model is however particularly suited for non-causal reasoning tasks [16], e.g. information fusion, and, compared to the Bayesian model, less tailored to diagnostic reasoning. So, when we have to chose for one of the two methods, a trade-off needs to be made. In general, when the problem mainly concerns causal/diagnostic reasoning and the information available is (almost) complete, i.e. probabilistic, the use of the Bayesian model is recommended. When the problem concerns mainly non-causal reasoning and the available information is incomplete, the D-S model is recommended. As the exact reasoning task and the associated uncertainty characteristics are application-specific, this trade-off needs to be made for each diagnosis problem individually. Unfortunately, a good insight into the characteristics of all uncertain influences is often missing for practitioners, which complicates the choice of the method. Table 4 gives an overview of the advantages and disadvantages of the Bayesian and the D-S model. The first three properties follow from the previous analysis, the remaining properties are discussed in the remainder of this section. Note that in this table, the two methods are compared qualitatively relative to each other, i.e. a minus sign merely indicates that the method is less suited compared to the other method.

7.2. Additional criteria

For practical problems, additional criteria like computational efficiency, suitability for decision making, clarity of inference, and adaptability are of importance (see Table 4).

7.2.1. Computational efficiency

Computationally, D-S networks are more expensive to evaluate than Bayesian networks [16, 43]. The worst-case complexity of a Bayesian network is $\mathcal{O}(n)$, whereas the worst-case complexity of a D-S network is $\mathcal{O}(2^n)$, with n the dimension of the state space of the largest clique in the join tree[5] [16]. The size n of the state space of the largest clique depends on the dimensions of the state spaces of variables, the dimensions of state spaces of valuations, and the structure of the graph [16]. To what extent the higher computational complexity of D-S networks is practically disadvantageous depends on the size of the network and on the available calculation time and power. For online diagnosis this implies that the Bayesian approach has the advantage that the diagnosis can be carried out with a smaller delay due to calculations.

7.2.2. Suitability for decision making

Often, it is argued that only the Bayesian model is appropriate for rational decision making, as probabilities fit within the expected-utility theory [9]. However, mass functions can be easily transformed to probability distributions at the moment decisions have to be made by using the pignistic transformation. Note that in the case of incomplete information, non-probabilistic information is transformed to probabilities without any fundamental reason to do so, except to facilitate decision making. Consider e.g. the extreme case that we have a non-informative mass function m^{Θ_H} regarding variable H:

$$m^{\Theta_H}(\Theta_H) = 1.$$

[5]A join tree is the moralization of a directed graph into a tree structure that supports efficient inference.

Table 3: Masses capturing the relation between F_1 and C_1

$m(0, c_1)$	$m(\{(0, c_1), (0, c_2)\})$	$m(\{(0, c_1), (0, c_2), (1, c_1)\})$	$m(\Theta_{C_1} \times \Theta_{F_1})$
$m(0, c_2)$	$m(\{(0, c_1), (1, c_1)\})$	$m(\{(0, c_2), (1, c_1), (1, c_2)\})$	
$m(1, c_1)$	$m(\{(0, c_1), (1, c_2)\})$	$m(\{(0, c_1), (1, c_1), (1, c_2)\})$	
$m(1, c_2)$	$m(\{(0, c_2), (1, c_1)\})$	$m(\{(0, c_1), (0, c_2), (1, c_2)\})$	
	$m(\{(0, c_2), (1, c_2)\})$		
	$m(\{(1, c_1), (1, c_2)\})$		

Table 4: Comparison of Bayesian and Dempster-Shafer reasoning

	Bayesian framework	D-S framework
Suitability for causal/diagnostic reasoning	+	−
Suitability for non-causal reasoning (e.g. information fusion)	−	+
Handling incomplete information	−	+
Computational efficiency	+	−
Suitability for decision making	+	+
Clarity of inference	+	−
Adaptability	+	−

We can transform this mass function into a probability distribution. However, as we have no knowledge, every probability distribution is equally good (or bad). Is it justified to make decisions based on guessed odds? In addition, if a decision needs to be made, is it justified to ignore that the outcome was just (or partly) based on a guess? Incomplete information indicates that the information collected so far is not sufficient to make a sound decision [18], so more information should be gathered or the diagnosis setup should be improved. In some situations, decisions need to be made, but even in these cases it seems beneficial to have insight into the underlying mass distributions, e.g. to give feedback about the quality of the monitoring setup. In addition, measures of uncertainty may provide information about the severeness of the fault [44]. Generally, it holds that the more severe the fault, the lower the ignorance and conflict. This is because for severe faults relatively large amounts of data are available. Moreover severe faults manifests itself more clearly in the data compared to incipient faults. Analyzing and exploiting the uncertainty present require that all computations are done in the D-S framework, which is computationally less attractive. However, applying a technique based on probabilities using information that is not probabilistic, may yield erroneous results [14].

Based on the considerations presented, we conclude that the Bayesian model naturally fits decision making. Decision making in the D-S framework is slightly more involved compared to decision making in the Bayesian framework. However, mass functions contain more information, so allowing more informed decisions. Therefore, we consider Bayesian and D-S reasoning equally suitable for decision making.

7.2.3. Clarity of inference

Clarity of inference is of importance for most practical applications, as the implementation of a decision support system is much easier when the reasoning is intuitive and understandable. In this sense, Bayesian networks outperform D-S networks, since the causal representation in Bayesian networks is more natural and easier for the user to provide and understand [45].

Although the Bayesian reasoning is considered clearer, the D-S output is clearer, as the D-S framework makes a distinction between probabilistic information and incomplete information. In the D-S framework, two distinct outcomes are obtained in the situation that no information regarding a variable H is available, i.e. $m^{\Theta_H}(\Theta_H) = 1$, and the situation in which we have the information that all elements in Θ_H are equally likely, i.e. $m^{\Theta_H}(h_1) = m^{\Theta_H}(h_2) = \ldots = m^{\Theta_H}(h_n) = 1/n$. In contrast, in the Bayesian framework the two situations are represented by the same probability distribution, $P(H = h_1) = P(H = h_2) = \ldots = P(H = h_n) = 1/n$. The additional information provided by the D-S outcome can be used to reconsider the diagnostic setup (e.g. an incomplete outcome may be a reason to extend the knowledge base, whereas a probabilistic answer may be a reason to implement better sensors) or to assist decision making (e.g. by choosing a conservative decision when the diagnostic result is incomplete).

7.2.4. Adaptability

Adaptability indicates how easily new knowledge can be incorporated in the network, e.g. when we want to include new faults or features in the model or update the relations between faults and features. This property is mainly important for large networks when it is expected that the model needs to be updated multiple times over

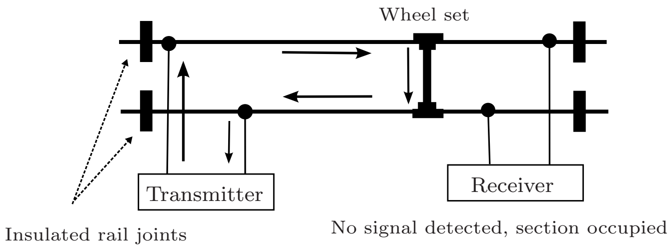


Figure 5: Current flow in an occupied section of track.

time. Both frameworks allow for the incorporation of new knowledge without the need to redefine the whole model. As (new) knowledge relating faults and features is generally in causal form, the incorporation in the Bayesian model is more straightforward.

8. Railway case study

A representative example of knowledge-based fault diagnosis under uncertainty is the diagnosis of railway track circuits using temporal and spatial information as considered in [46, 47]. In this section, we illustrate the reasoning concepts discussed in the current paper based on the track circuit diagnosis problem.

8.1. Problem formulation

To guarantee the safe operation of a railway network, track circuits are used to detect the absence of a train in a section of railway track. Trains are only allowed to enter sections that are reported free. The track circuit uses the rails as conductors that connect a transmitter at one end of the section to a receiver at the other end, as shown in Figure 5. When no train is present in the section, the current will activate a relay in the receiver, which indicates that the section is free. When a train enters the section, the wheels and axles of the train short the circuit. Consequently, the current through the receiver drops, the relay de-energizes, and the section is reported as occupied.

Track circuits only work properly if the conductance properties of the rails are high. When the conductance is below a certain level, the section will be reported as occupied regardless of the presence of a train, leading to unnecessary train delays. Two main causes have been identified that negatively influence rail conductance [47], namely:

1. Mechanical rail defects f_{rd}
2. Electrical disturbances f_{ed}

The goal is to determine, which fault (f_{rd} or f_{ed}) is present. Assume that from previous analysis, we can already conclude that the section suffers from a conductance problem and that we are concerned with the determination of its cause. To distinguish between these two faults, we proposed to monitor the temporal and spatial dependencies of the measured currents:

1. Temporal dependencies T, with $\Theta_T = \{\text{L}, \text{E}, \text{A}, \text{I}\}$
2. Spatial dependencies S, with $\Theta_S = \{\text{NC}, \text{CSS}, \text{CAS}\}$

with L = linear, E = exponential, A = abrupt, I = intermittent, NC = no correlation with other sections, CSS = correlation with sections on the same track, CAS = correlation with all nearby sections.

The Bayesian and D-S graphical representations of the diagnosis problem are given at the top of Table 5. Since only single-fault scenarios are allowed, we use one fault variable H with $\Theta_H = \{f_{\text{rd}}, f_{\text{ed}}\}$. Quantitatively, fault variable H is linked to the features S and T as follows:

k_1 : **If** $H = f_{\text{rd}}$ **then** $P(T = \text{E}) = 0.85$

k_2 : **If** $H = f_{\text{ed}}$ **then** $P(T = \text{A} \vee T = \text{I}) = 1$

k_3 : **If** $H = f_{\text{rd}}$ **then** $P(S = \text{NC}) = 1$

k_4 : **If** $H = f_{\text{ed}}$ **then** $P(T = \text{CSS}) = 0.7$

It encodes that a rail defect f_{rd} likely evolves exponentially over time, whereas an electrical disturbance is characterized by an intermittent or abrupt time behavior. A rail defect only influences the behavior of one particular section, while electrical disturbances likely influence the behavior of sections on the same track (i.e. connected sections). This system knowledge is conditional, uncertain, and incomplete.

We assume that no prior knowledge about the relative occurrence of the two faults is available and that the following uncertain pieces of evidence are available for diagnosis:

e_1: $P(T = \text{I}) = 0.3, P(T \neq \text{I}) = 0.7$

e_2: $P(T = \text{A} \vee T = \text{I}) = 1$

e_3: $P(S = \text{CCS}) = 0.8$

Evidence e_1 provides information about the temporal dependencies, but can only distinguish between intermittent and non-intermittent behavior. The second evidence indicates that the temporal behavior is not gradual, i.e. not linear or exponential, but cannot discriminate between intermittent and abrupt behavior. Evidence e_3 corresponds to an unreliability information source providing that $S =$ CCS.

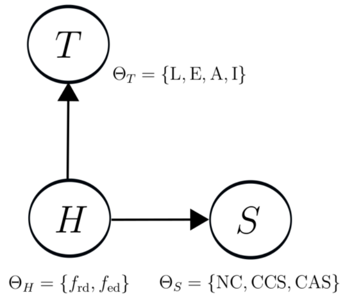
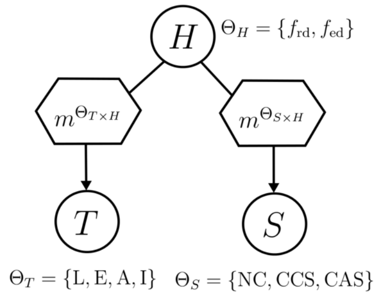
8.2. Bayesian solution

8.2.1. Information preprocessing

As indicated in Section 6.1, fault diagnosis using Bayesian networks requires three preprocessing steps.

Transformation of the knowledge base. The knowledge specified by the rules k_1 till k_4 needs to be represented by two conditional probability tables, one for T and one for S. The knowledge is already in conditional form, so we only have to represent the incomplete knowledge by probabilities. This is done based on the additivity axiom and the principle of maximum entropy. The obtained probability tables are included in Table 5.

Table 5: Summary of the diagnosis example

<table>
<tr><th></th><th>Bayesian</th><th>Dempster-Shafer</th></tr>
<tr><td>Graph</td><td>
</td><td>
</td></tr>
<tr><td>Knowledge $T \times H$</td><td>
<table>
<tr><th></th><th colspan="4">T</th></tr>
<tr><th>H</th><th>L</th><th>E</th><th>A</th><th>I</th></tr>
<tr><td>f_{rd}</td><td>0.05</td><td>0.85</td><td>0.05</td><td>0.05</td></tr>
<tr><td>f_{ed}</td><td>0</td><td>0</td><td>0.5</td><td>0.5</td></tr>
</table></td><td>$m^{\Theta_{T\times H}}(\{(E, f_{\mathrm{rd}}), (A, f_{\mathrm{ed}}), (I, f_{\mathrm{ed}})\}) = 0.85$
$m^{\Theta_{T\times H}}(\{(\cdot, f_{\mathrm{rd}}), (A, f_{\mathrm{ed}}), (I, f_{\mathrm{ed}})\}) = 0.15$</td></tr>
<tr><td>Knowledge $S \times H$</td><td>
<table>
<tr><th></th><th colspan="3">S</th></tr>
<tr><th>H</th><th>NC</th><th>CCS</th><th>CAS</th></tr>
<tr><td>f_{rd}</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>f_{ed}</td><td>0.15</td><td>0.7</td><td>0.15</td></tr>
</table></td><td>$m^{\Theta_{S\times H}}(\{(\mathrm{NC}, f_{\mathrm{rd}}), (\mathrm{CCS}, f_{\mathrm{ed}})\}) = 0.7$
$m^{\Theta_{S\times H}}(\{(\mathrm{NC}, f_{\mathrm{rd}}), (\cdot, f_{\mathrm{ed}})\}) = 0.3$</td></tr>
<tr><td>Prior knowledge H</td><td>
<table>
<tr><th colspan="2">H</th></tr>
<tr><th>f_{rd}</th><th>f_{ed}</th></tr>
<tr><td>0.5</td><td>0.5</td></tr>
</table></td><td>$m^{\Theta_H}(\Theta_H) = 1$</td></tr>
<tr><td>Temporal dependencies</td><td>
<table>
<tr><th></th><th>T_{obs}</th></tr>
<tr><th>T</th><th>$e_1 \wedge e_2$</th></tr>
<tr><td>L</td><td>0</td></tr>
<tr><td>E</td><td>0</td></tr>
<tr><td>A</td><td>7</td></tr>
<tr><td>I</td><td>9</td></tr>
</table></td><td>$m^{\Theta_T}(I) = 0.3$
$m^{\Theta_T}(A) = 0.7$</td></tr>
<tr><td>Spatial dependencies</td><td>
<table>
<tr><th></th><th>S_{obs}</th></tr>
<tr><th>S</th><th>e_3</th></tr>
<tr><td>NC</td><td>1</td></tr>
<tr><td>CCS</td><td>8</td></tr>
<tr><td>CAS</td><td>1</td></tr>
</table></td><td>$m^{\Theta_S}(\mathrm{CCS}) = 0.8$
$m^{\Theta_S}(\Theta_S) = 0.2$</td></tr>
<tr><td>Diagnosis result</td><td>$P(f_{\mathrm{rd}}) = 0.0167$
$P(f_{\mathrm{ed}}) = 0.9833$</td><td>$m^{\Theta_H}(f_{\mathrm{ed}}) = 0.97$
$m^{\Theta_H}(\Theta_H) = 0.03$</td></tr>
</table>

Prior probability distribution. For the root node H a prior probability distribution is needed. As we have no prior knowledge regarding the relative occurrence of the two faults, we adopt a uniform prior distribution (principle of maximum entropy).

Temporal evidences. Both evidence e_1 and evidence e_2 relate to the temporal dependencies T. In the Bayesian model, e_1 is represented by the following likelihood ratios:

$$\begin{aligned} P(e_1|\mathrm{L}) : P(e_1|\mathrm{E}) : P(e_1|\mathrm{A}) : P(e_1|\mathrm{I}) &= 0.23 : 0.23 : 0.23 : 0.3 \\ &= 7 : 7 : 7 : 9 \end{aligned} \tag{1}$$

Conditioning this information based on e_2, yields:

$$P(e_1, e_2|\mathrm{L}) : P(e_1, e_2|\mathrm{E}) : P(e_1, e_2|\mathrm{A}) : P(e_1, e_2|\mathrm{I}) = 0 : 0 : 7 : 9 \tag{2}$$

These ratios are reflected in the conditional probability of the virtual node T_{obs} (see Table 5).

Spatial evidence. Evidence e_3 is related to the spatial dependencies. Following the additivity axiom and the principle of maximum entropy, evidence e_3 is represented by the following likelihood ratios:

$$P(e_3|\mathrm{NC}) : P(e_3|\mathrm{CCS}) : P(e_3|\mathrm{CAS}) = 1 : 8 : 1 \tag{3}$$

which are reflected in the conditional probability table of the virtual node S_{obs} (see Table 5).

8.2.2. Fault diagnosis

To obtain the posterior probability distribution of H, we propagate the hard evidences on the virtual events T_{obs} and S_{obs} through the augmented Bayesian network. Updating (A.3) with $T_{\mathrm{obs}} = e_1 \wedge e_2$ yields:

$$P(H = f_{\mathrm{rd}}|e_1, e_2) = 0.0909 \tag{4}$$
$$P(H = f_{\mathrm{ed}}|e_1, e_2) = 0.9091 \tag{5}$$

Subsequently updating (A.3) with $S_{\mathrm{obs}} = e_3$ yields:

$$P(H = f_{\mathrm{rd}}|e_1, e_2, e_3) = 0.0167 \tag{6}$$
$$P(H = f_{\mathrm{ed}}|e_1, e_2, e_3) = 0.9833 \tag{7}$$

So, we conclude with a probability of slightly more than 98% that electrical disturbances are responsible for the conductance problem.

8.3. Dempster-Shafer solution

8.3.1. Information preprocessing

Transformation of the knowledge base. To convert the conditional knowledge regarding T and S to a mass function $m^{\Theta_{T\times H}}$ on the space $\Theta_T \times \Theta_H$ and a mass function $m^{\Theta_{S\times H}}$ on the space $\Theta_S \times \Theta_H$, we first use the ballooning extension (B.7) to derive two mass functions on both spaces. Next, we use Dempster's rule of combination (B.8) to combine the two mass functions on each space.

On the space $\Theta_T \times \Theta_H$, the ballooning extension (B.7) of rules k_1 and k_2 yields the following two mass functions:

$$\begin{aligned} m^{\Theta_T}[f_{\mathrm{rd}}]^{\Uparrow \Theta_{T\times H}}(\{(\mathrm{E}, f_{\mathrm{rd}}), (\cdot, f_{\mathrm{ed}})\}) &= 0.85 \\ m^{\Theta_T}[f_{\mathrm{rd}}]^{\Uparrow \Theta_{T\times H}}(\Theta_T \times \Theta_H) &= 0.15 \end{aligned} \tag{8}$$

$$m^{\Theta_T}[f_{\mathrm{ed}}]^{\Uparrow \Theta_{T\times H}}(\{(\mathrm{A}, f_{\mathrm{ed}}), (\mathrm{I}, f_{\mathrm{ed}}), (\cdot, f_{\mathrm{rd}})\}) = 1 \tag{9}$$

Combining them using (B.8) gives:

$$\begin{aligned} m^{\Theta_{T\times H}}(\{(\mathrm{E}, f_{\mathrm{rd}}), (\mathrm{A}, f_{\mathrm{ed}}), (\mathrm{I}, f_{\mathrm{ed}})\}) &= 0.85 \\ m^{\Theta_{T\times H}}(\{(\mathrm{A}, f_{\mathrm{ed}}), (\mathrm{I}, f_{\mathrm{ed}}), (\cdot, f_{\mathrm{rd}})\}) &= 0.15 \end{aligned} \tag{10}$$

On the space $\Theta_{S\times H}$, the ballooning extension (B.7) of rules k_3 and k_4 yields the following two mass functions:

$$m^{\Theta_S}[f_{\mathrm{rd}}]^{\Uparrow \Theta_{S\times H}}(\{(\mathrm{NC}, f_{\mathrm{rd}}), (\cdot, f_{\mathrm{ed}})\}) = 1 \tag{11}$$

$$\begin{aligned} m^{\Theta_S}[f_{\mathrm{ed}}]^{\Uparrow \Theta_{S\times H}}(\{(\mathrm{CCS}, f_{\mathrm{ed}}), (\cdot, f_{\mathrm{rd}})\}) &= 0.7 \\ m^{\Theta_S}[f_{\mathrm{ed}}]^{\Uparrow \Theta_{S\times H}}(\Theta_S \times \Theta_H) &= 0.3 \end{aligned} \tag{12}$$

Combining them using (B.8) gives:

$$\begin{aligned} m^{\Theta_{S\times H}}(\{(\mathrm{CCS}, f_{\mathrm{ed}}), (\mathrm{NC}, f_{\mathrm{rd}})\}) &= 0.7 \\ m^{\Theta_{S\times H}}(\{(\mathrm{NC}, f_{\mathrm{rd}}), (\cdot, f_{\mathrm{ed}})\}) &= 0.3 \end{aligned} \tag{13}$$

Temporal evidences. In the D-S framework, evidence e_1 is represented by the following mass function:

$$\begin{aligned} m^{\Theta_T}(\mathrm{I}) &= 0.3 \\ m^{\Theta_T}(\mathrm{L} \vee \mathrm{E} \vee \mathrm{A}) &= 0.7 \end{aligned} \tag{14}$$

Conditioning this knowledge based on evidence e_2 yields:

$$\begin{aligned} m^{\Theta_T}(\mathrm{I}) &= 0.3 \\ m^{\Theta_T}(\mathrm{A}) &= 0.7 \end{aligned} \tag{15}$$

Spatial evidence. In the D-S framework, evidence e_3, related to the spatial dependencies, is represented as:

$$\begin{aligned} m^{\Theta_S}(\mathrm{CCS}) &= 0.8 \\ m^{\Theta_S}(\Theta_S) &= 0.2 \end{aligned} \tag{16}$$

8.3.2. Fault diagnosis

To infer the fault cause, we first combine the mass functions m^{Θ_T} and m^{Θ_S} with the corresponding valuation functions $m^{\Theta_{T\times H}}$ and $m^{\Theta_{S\times H}}$. So, m^{Θ_T} is combined with $m^{\Theta_{T\times H}}$ and m^{Θ_S} with $m^{\Theta_{S\times H}}$. Next, we project the mass function on Θ_H. To combine two mass functions on different spaces we use the cylindrical extension (B.5). So, we vacuously extend m^{Θ_T} to the space $\Theta_T \times \Theta_H$ and m^{Θ_S} to the space $\Theta_S \times \Theta_H$.

On the space $\Theta_T \times \Theta_H$ the following results are obtained: The cylindrical extension (B.5) of m^{Θ_T} on $\Theta_T \times \Theta_H$ yields:

$$m^{\Theta_T \uparrow \Theta_{T\times H}}\big(\{(I, f_{\mathrm{rd}}), (I, f_{\mathrm{ed}})\}\big) = 0.3 \tag{17}$$
$$m^{\Theta_T \uparrow \Theta_{T\times H}}\big(\{(A, f_{\mathrm{rd}}), (A, f_{\mathrm{ed}})\}\big) = 0.7 \tag{18}$$

Combining this mass function with the valuation function $m^{\Theta_{T\times H}}$ according to Dempster's rule of combination (B.8) gives:

$$\begin{aligned} m^{\Theta_{T\times H}}(\mathrm{I}, f_{\mathrm{ed}}) &= 0.3 \cdot 0.85 \\ m^{\Theta_{T\times H}}(\mathrm{A}, f_{\mathrm{ed}}) &= 0.7 \cdot 0.85 \\ m^{\Theta_{T\times H}}(\{(\mathrm{I}, f_{\mathrm{rd}}), (\mathrm{I}, f_{\mathrm{ed}})\}) &= 0.3 \cdot 0.15 \\ m^{\Theta_{T\times H}}(\{(\mathrm{A}, f_{\mathrm{rd}}), (\mathrm{A}, f_{\mathrm{ed}})\}) &= 0.7 \cdot 0.15 \end{aligned} \tag{19}$$

Marginalization of $m^{\Theta_{T\times H}}$ on Θ_H accordingly (B.6) gives:

$$\begin{aligned} m^{\Theta_{T\times H} \downarrow \Theta_H}(f_{\mathrm{ed}}) &= 0.3 \cdot 0.85 + 0.7 \cdot 0.85 \\ m^{\Theta_{T\times H} \downarrow \Theta_H}(\Theta_H) &= 0.3 \cdot 0.15 + 0.7 \cdot 0.15 \end{aligned} \tag{20}$$

On the space $\Theta_S \times \Theta_H$, the following results are obtained: The cylindrical extension (B.5) of Θ_S yields:

$$\begin{aligned} m^{\Theta_S \uparrow \Theta_{S\times H}}\big(\{(\mathrm{CCS}, f_{\mathrm{rd}}), (\mathrm{CCS}, f_{\mathrm{ed}})\}\big) &= 0.8 \\ m^{\Theta_S \uparrow \Theta_{S\times H}}\big(\Theta_S \times \Theta_H\big) &= 0.2 \end{aligned} \tag{21}$$

Combining (21) with the valuation function $m^{\Theta_{S\times H}}$ according to (B.8) gives:

$$\begin{aligned} m^{\Theta_{S\times H}}(\mathrm{CCS}, f_{\mathrm{ed}}) &= 0.7 \cdot 0.8 + 0.3 \cdot 0.8 \\ m^{\Theta_{S\times H}}(\{(\mathrm{CCS}, f_{\mathrm{ed}}), (\mathrm{NC}, f_{\mathrm{rd}})\}) &= 0.7 \cdot 0.2 \\ m^{\Theta_{S\times H}}(\{(\mathrm{NC}, f_{\mathrm{rd}}), (\cdot, f_{\mathrm{ed}})) &= 0.3 \cdot 0.2 \end{aligned} \tag{22}$$

Marginalization of $m^{\Theta_{S\times H}}$ on Θ_H accordingly (B.6) gives:

$$\begin{aligned} m^{\Theta_{S\times H} \downarrow \Theta_H}(f_{\mathrm{ed}}) &= 0.7 \cdot 0.8 + 0.3 \cdot 0.8 \\ m^{\Theta_{S\times H} \downarrow \Theta_H}(\Theta_H) &= 0.7 \cdot 0.2 + 0.3 \cdot 0.2 \end{aligned} \tag{23}$$

Combining (20) and (23) according to the conjunctive rule of combination (B.9) results in the final mass distribution:

$$\begin{aligned} m^{\Theta_H}(f_{\mathrm{ed}}) &= 0.97 \\ m^{\Theta_H}(\Theta_H) &= 0.03 \end{aligned} \tag{24}$$

In the case that the diagnosis result serves as input for a decision making process, the following pignistic probability distribution is obtained:

$$\begin{aligned} P_{\mathrm{pig}}(f_{\mathrm{rd}}) &= 0.015 \\ P_{\mathrm{pig}}(f_{\mathrm{ed}}) &= 0.985 \end{aligned} \tag{25}$$

Like in the Bayesian model, it is concluded with a probability of slightly more than 98% that the conductance problem is caused by electrical disturbances.

8.4. Modified case

Consider the case as introduced in Section 8.1, but with rule k_2 redefined as:

k_2': **If** $H = f_{\mathrm{ed}}$ **then** $P(T = I) = 1$

The associated conditional probability table of T is given in Table 6. The corresponding valuation function $m^{\Theta_{T\times H}}$ is:

$$\begin{aligned} m^{\Theta_{T\times H}}(\{(\mathrm{E}, f_{\mathrm{rd}}), (\mathrm{I}, f_{\mathrm{ed}})\}) &= 0.85 \\ m^{\Theta_{T\times H}}(\{(\mathrm{I}, f_{\mathrm{ed}}), (\cdot, f_{\mathrm{rd}})\}) &= 0.15 \end{aligned} \tag{26}$$

Following the same analysis as before, the following diagnosis results or obtained. According to the Bayesian model:

$$\begin{aligned} P(f_{\mathrm{rd}}) &= 0.015 \\ P(f_{\mathrm{ed}}) &= 0.985 \end{aligned} \tag{27}$$

According to the D-S model:

$$\begin{aligned} m(f_{\mathrm{ed}}) &= 0.718 \\ m(f_{\mathrm{rd}}) &= 0.052 \\ m(\Theta_H) &= 0.022 \\ m(\emptyset) &= 0.207 \end{aligned} \tag{28}$$

Both the Bayesian and the D-S solution point towards a conductance problem. The D-S solution encodes more uncertainty about this conclusion compared to the Bayesian solution.

8.5. Evaluation

We have illustrated how the track circuit diagnosis problem is handled in both the Bayesian and the D-S framework. In the original case, the available information is almost complete and non-conflicting, and both frameworks conclude with a high confidence that electrical disturbances are responsible for the conductance problem. In the modified case, the different evidences are partially conflicting and the results obtained in the two frameworks differ. The Bayesian model, again, concludes with a high confidence that electrical disturbances are responsible for the conductance problems. The D-S model also concludes that the conductance problem is most likely caused by electrical disturbances, but the model is less confident and also indicates that there is some conflicting information. The conflict may e.g. indicate that a fault not included in

Table 6: Conditional probability table T

	T			
H	L	E	A	I
f_{rd}	0.05	0.85	0.05	0.05
f_{ed}	0	0	0	1

Θ_H is responsible for the conductance problem, or that one or more of the evidences is unreliable. The different conclusions can be partly explained by the way evidence e_1 and e_2 are interpreted in the two frameworks: According to the Bayesian model, the temporal behavior is most likely intermittent (I). According to the D-S interpretation, the temporal behavior is most likely abrupt (A). The most likely feature values in the D-S model, T = A and S = CSS, are partially conflicting with respect to the fault cause, explaining the conflict in the D-S solution.

In summary, the preferred reasoning framework depends on how and to which extent the available knowledge and evidences are disturbed by uncertainty. When the available information is almost complete and non-conflicting, the Bayesian and D-S diagnosis outcome will be close. Considering Table 4, in such cases, the Bayesian model seems to be the preferred one since it is computationally less demanding, clearer, and easier to adapt. When the available knowledge is partially incomplete or conflicting, the D-S outcome is more informative and consequently may be preferred over the Bayesian outcome. Whether this advantage outweighs the Bayesian advantages as listed in Table 4, depends on the degree to which the information is incomplete and conflicting and on application-specific preferences, e.g. what are the consequences of an incorrect decision, and how important are intuitiveness and adaptability.

9. Conclusions

In this paper, Bayesian and Dempster-Shafer reasoning have been compared for knowledge-based fault diagnosis. The Bayesian model is based on probabilities and is tailored to causal reasoning based on probabilistic knowledge. The Dempster-Shafer model is based on belief functions and is tailored to non-causal reasoning, e.g. information fusion, based on both probabilistic and incomplete information. Fault diagnosis comprises causal reasoning, often based on incomplete information. So, none of the two reasoning models fits the diagnostic reasoning task in a straightforward way. In addition, real-life diagnosis problems often include additional criteria, e.g. we want to know how reliable the reasoning results are or we want to retrieve why a certain conclusion has been reached. For such problems, without an exactly defined performance criterion, it is not possible to unambiguously conclude what the best method is. We have concluded that the final choice for a reasoning framework depends on the problem under consideration (including uncertainty characteristics), requirements of the user, and personal preferences. In general, the better the match between the probabilistic description and the real information, the more suitable the Bayesian approach is. The more conflicting and incomplete the available information, the more informative the D-S solution is compared to the Bayesian solution.

As a topic for further research, we propose to apply the two reasoning methods in a model-based diagnosis frameworks based on residuals. Moreover, we will apply Bayesian and Dempster-Shafer reasoning on a representative fault diagnosis problem and examine their diagnostic performance for this problem. Finally, we will develop methods for failure prognosis and condition-based maintenance planning based on the uncertain fault diagnosis results.

Acknowledgment

This research is supported by ProRail and the Dutch Technology Foundation STW, which is part of the Netherlands Organization for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs. The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement nr 324432.

References

[1] J. Chen, R. J. Patton, Optimal filtering and robust fault diagnosis of stochastic systems with unknown disturbances, IEE Proceedings-Control Theory and Applications 143 (1) (1996) 31–36.

[2] K. Li, Y. L. Zhang, Z. X. Li, Application research of kalman filter and svm applied to condition monitoring and fault diagnosis, in: Applied Mechanics and Materials, Vol. 121, 2012, pp. 268–272.

[3] M. Mrugalski, An unscented kalman filter in designing dynamic GMDH neural networks for robust fault detection, International Journal of Applied Mathematics and Computer Science 23 (1) (2013) 157–169.

[4] C. Combastel, Merging kalman filtering and zonotopic state bounding for robust fault detection under noisy environment, IFAC-PapersOnLine 48 (21) (2015) 289–295.

[5] V. Puig, Fault diagnosis and fault tolerant control using set-membership approaches: Application to real case studies, International Journal of Applied Mathematics and Computer Science 20 (4) (2010) 619–635.

[6] J. Blesa, V. Puig, J. Romera, J. Saludes, Fault diagnosis of wind turbines using a set-membership approach, IFAC Proceedings Volumes 44 (1) (2011) 8316–8321.

[7] S. S. Tayarani-Bathaie, Z. S. Vanini, K. Khorasani, Dynamic neural network-based fault diagnosis of gas turbine engines, Neurocomputing 125 (2014) 153–165.

[8] Z. Du, B. Fan, X. Jin, J. Chi, Fault detection and diagnosis for buildings and hvac systems using combined neural networks and subtractive clustering analysis, Building and Environment 73 (2014) 1–11.

[9] D. Lindley, The probability approach to the treatment of uncertainty in artificial intelligence and expert systems, Statistical Sciences 2 (1) (1987) 17–24.

[10] P. Cheeseman, In defense of probability., in: Proceedings of the 9th International Joint Conference on Artificial Intelligence, Los Angeles, California, USA, 1985, pp. 1002–1009.

[11] P. Smets, Resolving misunderstandings about belief functions, International Journal of Approximate Reasoning 6 (3) (1992) 321–344.

[12] P. Smets, What is Dempster-Shafer's model?, in: R. Yager, J. Kacprzyk, M. Fedrizzi (Eds.), Advances in the Dempster-Shafer Theory of Evidence, John Wiley & Sons, Inc., 1994, pp. 5–34.

[13] D. Dubois, H. Prade, P. Smets, Representing partial ignorance, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 26 (3) (1996) 361–377.
[14] S. Ferson, L. Ginzburg, Different methods are needed to propagate ignorance and variability, Reliability Engineering & System Safety 54 (2) (1996) 133–144.
[15] D. Dubois, H. Prade, Possibility theory, probability theory and multiple-valued logics: A clarification, Annuals of Mathematics and Artificial Intelligence 32 (1-4) (2001) 35–66.
[16] B. R. Cobb, P. P. Shenoy, A comparison of Bayesian and belief function reasoning, Information Systems Frontiers 5 (4) (2003) 345–358.
[17] L. Zadeh, Is there a need for fuzzy logic?, Information Sciences 178 (13) (2008) 2751–2779.
[18] R. Haenni, Ignoring ignorance is ignorant, Tech. rep., Center for Junior Research Fellows, University of Konstanz, Konstanz, Germany (2003).
[19] G. Shafer, Belief functions, in: G. Shafer, J. Pearl (Eds.), Readings in Uncertain Reasoning, Morgan Kaufman, 1990.
[20] K. Verbert, B. De Schutter, R. Babuška, Reasoning under uncertainty for knowledge-based fault diagnosis: A comparative study, in: Proceedings of the 9th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes, Paris, France, 2015, pp. 422–427.
[21] A. D. Kiureghian, O. Ditlevsen, Aleatory or epistemic? does it matter?, Structural Safety 31 (2) (2009) 105–112.
[22] R. Billinton, D. Huang, Aleatory and epistemic uncertainty considerations in power system reliability evaluation, in: Proceedings of the 10th International Conference on Probabilistic Methods Applied to Power Systems, Rincon, Bonaire, 2008, pp. 1–8.
[23] P. Laplace, A Philosophical Essay on Probabilities, New York: Dover Publications Inc., 1814.
[24] L. Savage, The Foundations of Statistics, New York: John Wiley & Sons, 1954.
[25] D. Howie, Interpreting Probability: Controversies and Developments in the Early Twentieth Century, Cambridge University Press, 2002.
[26] M. Goldstein, Subjective Bayesian analysis: principles and practice, Bayesian Analysis 1 (3) (2006) 403–420.
[27] A. Dempster, Upper and lower probabilities induced by a multivalued mapping, The Annals of Mathematical Statistics (1967) 325–339.
[28] G. Shafer, A Mathematical Theory of Evidence, Vol. 1, Princeton University Press Princeton, 1976.
[29] J. Kohlas, P. Monney, A Mathematical Theory of Hints: An Approach to the Dempster-Shafer Theory of Evidence, Springer-Verlag, 1995.
[30] P. Smets, R. Kennes, The transferable belief model, Artificial Intelligence 66 (2) (1994) 191–234.
[31] D. Dubois, H. Prade, Possibility Theory, Plenum Press, New-York, 1988.
[32] L. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems 100, Supplement 1 (1999) 9–34.
[33] D. Dubois, Possibility theory and statistical reasoning, Computational Statistics & Data Analysis 51 (1) (2006) 47–69.
[34] L. A. Zadeh, Fuzzy logic and approximate reasoning, Synthese 30 (3-4) (1975) 407–428.
[35] G. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic, Vol. 4, Prentice Hall New Jersey, 1995.
[36] L. A. Zadeh, Generalized theory of uncertainty (GTU)-principal concepts and ideas, Computational Statistics & Data Analysis 51 (1) (2006) 15–46.
[37] A. O'Hagan, Dicing with the unknown, Significance 1 (3) (2004) 132–133.
[38] L. Zadeh, Fuzzy sets, Information and Control 8 (3) (1965) 338–353.
[39] R. Isermann, Fault-Diagnosis Applications: Model-Based Condition Monitoring: Actuators, Drives, Machinery, Plants, Sensors, and Fault-Tolerant Systems, Springer Science & Business Media, 2011.
[40] P. Frank, E. A. Garcıa, B. Köppen-Seliger, Modelling for fault detection and isolation versus modelling for control, Mathematics and Computers in Simulation 53 (4) (2000) 259–271.
[41] E. T. Jaynes, Information theory and statistical mechanics, Physical review 106 (4) (1957) 620.
[42] E. T. Jaynes, Information theory and statistical mechanics. ii, Physical review 108 (2) (1957) 171.
[43] R. Haenni, N. Lehmann, Implementing belief function computations, International Journal of Intelligent Systems 18 (1) (2003) 31–49.
[44] S. Engel, B. Gilmartin, K. Bongort, A. Hess, Prognostics, the real issues involved with predicting life remaining, in: Proceedings of the IEEE Aerospace Conference, Big Sky, MT, USA, 2000, pp. 457–469.
[45] B. B. Yaghlane, K. Mellouli, Inference in directed evidential networks based on the transferable belief model, International Journal of Approximate Reasoning 48 (2) (2008) 399–418.
[46] K. Verbert, B. De Schutter, R. Babuška, Exploiting spatial and temporal dependencies to enhance fault diagnosis: Application to railway track circuits, in: Proceedings of the 2015 European Control Conference, Linz, Austria, 2015, pp. 3052–3057.
[47] K. Verbert, B. De Schutter, R. Babuška, Fault diagnosis using spatial and temporal information with application to railway track circuits, Engineering Applications of Artificial Intelligence 56 (2016) 200–211.
[48] A. Darwiche, Modeling and Reasoning with Bayesian Networks, Cambridge University Press, 2009.
[49] D. Heckerman, A tutorial on learning with Bayesian networks, in: M. Jordan (Ed.), Learning in Graphical Models, Vol. 89, Springer Netherlands, 1998, pp. 301–354.
[50] A. B. Mrad, V. Delcroix, S. Piechowiak, P. Leicester, M. Abid, An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence, Applied Intelligence 43 (4) (2015) 802–824.
[51] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers Inc., 1988.
[52] J. Pearl, Causality; Models, Reasoning, and Inference, Cambridge University Press, 2000.
[53] D. Dubois, S. Moral, H. Prade, Belief change rules in ordinal and numerical uncertainty theories, in: D. Dubois, H. Prade (Eds.), Belief Change, Vol. 3 of Handbook of Defeasible Reasoning and Uncertainty Management Systems, Springer Netherlands, 1998, pp. 311–392.
[54] A. B. Mrad, V. Delcroix, S. Piechowiak, P. Leicester, From information to evidence in a Bayesian network, in: Probabilistic Graphical Models, Springer, 2014, pp. 33–48.
[55] H. Chan, A. Darwiche, On the revision of probabilistic beliefs using uncertain evidence, Artificial Intelligence 163 (1) (2005) 67–90.
[56] D. V. Lindley, Making Decisions, Wiley New York, 1985.
[57] B. B. Yaghlane, P. Smets, K. Mellouli, Directed evidential networks with conditional belief functions, in: Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer, 2003, pp. 291–305.
[58] P. Smets, Un modéle mathématico-statistique stimulant le processus du diagnostic médical, Ph.D. thesis, Université de Bruxelles (1978).
[59] P. P. Shenoy, A valuation-based language for expert systems, International Journal of Approximate Reasoning 3 (5) (1989) 383–411.
[60] P. P. Shenoy, Valuation-based systems: A framework for managing uncertainty in expert systems, in: L. A. Zadeh, J. Kacprzyk (Eds.), Fuzzy Logic for the Management of Uncertainty, John Wiley & Sons, Inc., New York, NY, USA, 1992, pp. 83–104.
[61] P. Shenoy, Valuation-based systems for Bayesian decision analysis, Operations Research 40 (3) (1992) 463–484.
[62] P. Smets, The transferable belief model and other interpretations of Dempster-Shafer's model, in: Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence, Amsterdam, the Netherlands, 1990, pp. 375–383.
[63] R. Yager, On the Dempster-Shafer framework and new combi-

nation rules, Information Science 41 (2) (1987) 93–137.
[64] S. Destercke, D. Dubois, Idempotent conjunctive combination of belief functions: Extending the minimum rule of possibility theory, Information Sciences 181 (18) (2011) 3925–3945.
[65] L. Zadeh, Review of a mathematical theory of evidence, AI Magazine 5 (3) (1984) 81–83.
[66] P. Smets, Decision making in the TBM: the necessity of the pignistic transformation, International Journal of Approximate Reasoning 38 (2) (2005) 133–147.
[67] P. Smets, Application of the transferable belief model to diagnostic problems, International Journal of Intelligent Systems 13 (2-3) (1998) 127–157.

Appendix A. Reasoning in Bayesian networks

Appendix A.1. Uncertainty representation

In the Bayesian framework uncertainty is represented by (conditional) probabilities. At each time and for each variable X, a conditional probability $P(x_i|\mathcal{E})$ between zero and one is assigned to *each individual element* x_i in the domain Θ_X of X such that [48]:

$$\sum_{x_i \in \Theta_X} P(x_i|\mathcal{E}) = 1 \tag{A.1}$$

with $\mathcal{E}$ the collection of the currently available information.

Appendix A.2. Bayesian networks

A Bayesian network is a graphical model for probabilistic relationships among a set of variables that provides a powerful way to embed knowledge and to update one's beliefs about target variables given new information about other variables [49, 50]. Formally, a Bayesian network for a set of variables $\mathbf{V}$ is a pair (G, D) [49, 50], with:

1. $G = (\mathbf{V}, \mathbf{E})$ a directed acyclic graph with nodes $\mathbf{V}$ and directed edges $\mathbf{E}$ that encodes a set of conditional independence assertions about the variables in $\mathbf{V}$;
2. D a set of local probability distributions associated with each variable in $\mathbf{V}$.

In a Bayesian network, a directed edge from a variable X to a variable Y indicates that X has a direct influence on variable Y. Variable X is called a *parent* of variable Y and variable Y is called a *child* of variable X. The lack of possible edges in G encodes conditional independence [49]. Bayesian networks satisfy the Markov condition, meaning that any node is conditionally independent of its non-descendants given its parents. Thanks to the Markov assumption, the joint distribution of the complete system can be obtained in an efficient way by combining the conditional distributions of each variable given its parents [48, 51]: Given the network structure (G, D), the joint probability distribution for $\mathbf{V}$ is given by:

$$P(\mathbf{V} = \mathbf{v}) = \prod_{x_i \in \mathbf{v}} P(x_i|\mathbf{u}_X) \tag{A.2}$$

with $\mathbf{U}_X \subset \mathbf{V}$ the parents (immediate predecessors) of $X \in \mathbf{V}$ and $P(x_i|\mathbf{u}_X)$ the local probabilities associated with variable X, which are collected in D. Consequently, the pair (G, D) uniquely defines the joint probability distribution of $\mathbf{V}$.

Appendix A.3. Reasoning under uncertainty

Once the Bayesian network has been constructed (from prior knowledge, data, or a combination of both), we can use it to determine the probabilities of interest. This process is known as *probabilistic inference* [49].

Appendix A.3.1. Inference with hard evidences

Probabilistic inference with hard evidences can be regarded as a mechanism for automatically applying Bayes' rule:

$$P(y_i|x_i) = \frac{P(x_i|y_i)P(y_i)}{\sum_{y_j \in \Theta_Y} P(x_i|y_j)P(y_j)} \tag{A.3}$$

with:

- $P(y_i)$: prior probability that $Y = y_i$
- $P(y_i|x_i)$: posterior probability, i.e. the probability that $Y = y_i$ after observing $X = x_i$
- $P(x_i|y_i)$: likelihood function, i.e. the probability of observing $X = x_i$ given $Y = y_i$

The importance of Bayes' rule is that it expresses a quantity $P(y_i|x_i)$, which is often difficult to assess, in terms of quantities that often can be drawn directly from expert knowledge [51]. For a more thorough discussion on inference algorithms in Bayesian networks, we refer the interested reader to e.g. [52].

Appendix A.3.2. Inference with uncertain evidences

In practice, the available evidences are often uncertain, in which case Bayes' rule is not directly applicable. With respect to uncertain evidences, a distinction can be made between [50]:

1. Likelihood (or virtual) evidence
2. Probabilistic evidence
   (a) fixed
   (b) non-fixed

A likelihood evidence on a variable $X \in \mathbf{V}$ is specified by likelihood ratios $L(X)$:

$$L(X) = P(\eta|x_1) : \ldots : P(\eta|x_n) \tag{A.4}$$

with $P(\eta|x_i)$ the probability of the observation η given $X = x_i$. Likelihood evidence concerns evidence with uncertainty, i.e. the uncertainty bears on the meaning of the input [53]; the existence of the input itself is uncertain due to e.g. the unreliability of the source that supplies the input [54]. Note that likelihood evidence is specified "without a prior", as a consequence, its correct propagation requires both the evidence and the current belief in X to be taken into account.

A probabilistic evidence on a variable $X \in \mathbf{V}$ is specified by a local probability distribution $R(X)$ that defines a constraint on the beliefs on the variable X after the evidence has been propagated, i.e. $R(X)$ is an absolute constraint on the posterior probability distribution of X. Probabilistic evidence concerns evidence of uncertainty, i.e. the uncertainty is part of the input [53]. Fixed probabilistic evidence cannot be altered by any further information, while non-fixed probabilistic information can be modified based on later evidences [50].

Two main methods exist for revising probabilistic belief in the case of uncertain evidence [55]:

1. Jeffrey's rule of probability kinematics;
2. Pearl's method of virtual evidence.

Likelihood evidence is propagated by Pearl's method of virtual evidence, while probabilistic evidence is propagated following Jeffrey's rule. Note that for the propagation of multiple fixed probabilistic evidences, specific iterative algorithms, such as big clique or BN-IPFP, are needed to ensure that all constraints imposed by the different evidences are satisfied [50]. Although the various belief revision principles seem to be different, they are all based on the principle of probability kinematics [55], which can be viewed as a principle for minimizing belief change while satisfying the (absolute or relative) constraints imposed by the evidence. In addition, it has been shown that one can translate an evidential constraint used by Jeffrey's rule into one used by Pearl's method and vice versa [55]. Furthermore, as Pearl's method is directly applicable to Bayesian networks, while Jeffrey's rule is not, we will only elaborate on Pearl's method here. For a more thorough discussion on belief propagation based on uncertain evidences, we refer the interested reader to [50, 55] and the references therein.

Pearl's method of virtual evidence [55]: Given an original distribution $P(\mathbf{V})$ and some uncertain evidence η regarding variable $X \in \mathbf{V}$, a likelihood evidence is specified by $\lambda_1, ..., \lambda_n$ as:

$$P(\eta|x_1) : ... : P(\eta|x_n) = \lambda_1 : ... : \lambda_n \tag{A.5}$$

The method assumes that the observation η depends only on variable X and is independent of any other variable $Y \in \mathbf{V}$ given X:

$$P(\eta|x_i, y_i) = P(\eta|x_i), \quad \text{for } i = 1, ..., n \tag{A.6}$$

This results in the following expression for the revised distribution:

$$P(y_i|\eta) = \frac{\sum_{j=1}^{n} \lambda_i P(y_i, x_j)}{\sum_{j=1}^{n} P(x_j)} \tag{A.7}$$

In a Bayesian network, the virtual evidence is represented by adding an auxiliary variable Z and a directed edge $X \rightarrow Z$, where one value of Z, say z_i, corresponds to the virtual event η. This ensures assumption (A.6): the virtual event η is independent of every variable Y given X. The uncertainty of the evidence is quantified by the likelihood ratios $\lambda_1, ..., \lambda_n$ and the conditional probability table of variable Z is assigned such that $P(z_i|x_1), ..., (z_i|x_n) = \lambda_1 : ... : \lambda_n$. The Bayesian network (augmented with variable Z) is updated in the standard way with observation $Z = z_i$, which is a hard evidence.

Appendix A.4. Decision making

Often, decisions have to be made given uncertain information regarding the situation you are in. When the uncertain information is represented by a probability distribution, the expected-utility theory is generally used for decision making. The *expected-utility theory* [56] provides a framework for determining the optimal action given probabilistic information regarding the situation you are in. Its two main ingredients are:

1. Utilities, which indicate the desirability of a particular action in a particular situation, i.e. utilities express preferences among the available choices.
2. Probabilities, which indicate how likely a particular situation is.

The expected utility $\mathbb{E}(u|a)$ of action a is computed as:

$$\mathbb{E}(u|a) = \sum_{\mathbf{v}_i \in \Theta_{\mathbf{V}}} P(\mathbf{v}_i) u(a, \mathbf{v}_i) \tag{A.8}$$

with $u(a, \mathbf{v}_i)$ the utility of action a given situation $\mathbf{V} = \mathbf{v}_i$. Then the optimal decision a^* is:

$$a^* = \arg\max_a \mathbb{E}(u|a) \tag{A.9}$$

Appendix B. Reasoning in D-S networks

We adopt Smets' Transferable Belief Model (TBM) interpretation of the Dempster-Shafer theory [30]. In the TBM, uncertainty is managed at two levels: the *credal level* where beliefs are entertained and the *pignistic level* where beliefs are used to make decisions. The model does not rely on a probabilistic quantification, but on a more general system based on belief functions [12]. In contrast to Bayesian probabilities, belief functions can express states of ignorance.

Appendix B.1. Uncertainty representation

To enable the expression of (partial) ignorance, in the D-S framework, belief is assigned to each subset of the domain Θ_Y of a variable Y. The power set of Θ_Y, denoted as 2^{Θ_Y}, is a set containing all the possible subsets of Θ_Y. The mapping $\text{bel} : 2^{\Theta_Y} \rightarrow [0, 1]$ is a belief function if and only if there exist a basic belief assignment (bba) $m^{\Theta_Y}: 2^{\Theta_Y} \to [0,1]$ such that:

$$\sum_{y \subseteq \Theta_Y} m^{\Theta_Y}(y) = 1 \tag{B.1}$$

$$\text{bel}(y) = \sum_{\emptyset \neq x \subseteq y} m^{\Theta_Y}(x), \text{ and bel}(\emptyset) = 0 \tag{B.2}$$

$$\text{pl}(y) = \sum_{x \cap y \neq \emptyset} m^{\Theta_Y}(x), \text{ and pl}(\emptyset) = 0 \tag{B.3}$$

The mass $m^{\Theta_Y}(y)$ allocated to $y \subseteq \Theta_Y$ is the degree of belief that is exactly committed to y and that cannot be allocated to a more specific subset. The value $\text{bel}(y)$ quantifies the strength of the belief that the event y occurs. The value $\text{pl}(y)$ quantifies the maximum amount of potential specific support that could be given to y. It can be interpreted as the degree to which the evidence is not contradictory with y, i.e. $\text{pl}(y) = 1 - \text{bel}(\bar{y})$, where $\bar{y}$ is the complement of y with respect to the domain Θ_Y. When mass is only assigned to singleton elements, the mass distribution reduces to a probability distribution.

For illustration consider the uncertain variable H, with $\Theta_H = \{h_1, h_2, h_3\}$ and belief assignment:

$$\begin{aligned} m^{\Theta_H}(\{h_1\}) &= 0.1 \\ m^{\Theta_H}(\{h_2, h_3\}) &= 0.9 \\ m^{\Theta_H}(\{h_2\}) = m^{\Theta_H}(\{h_3\}) = m^{\Theta_H}(\{h_1, h_2\}) &= 0 \\ m^{\Theta_H}(\{h_1, h_3\}) = m^{\Theta_H}(\Theta_H) = m^{\Theta_H}(\emptyset) &= 0 \end{aligned} \tag{B.4}$$

Mass distribution (B.4) indicates that no information is available to discriminate between the outcomes h_2 and h_3. Note that the Bayesian model cannot represent such incomplete information. In the Bayesian model, the mass assigned to $\{h_2, h_3\}$ would typically be equally divided between the two elements (principle of maximum entropy). The situation in which one knows that h_2 and h_3 are equally likely and the situation in which one does not know anything about the individual probabilities $P(h_2)$ and $P(h_3)$ result in the same probability distribution. This is precisely what D-S proponents claim as the main shortcoming of the Bayesian framework [16].

When we model aspects of the real world, we often have to deal with multivariate situations [57], where the state space is a product space and information may be available in a conditional form. Multivariate belief function theory is well suited to handle real-world problems. A multivariate mass function $m^{\Theta_X \times \Theta_Y}$ on $\Theta_X \times \Theta_Y$ can be seen as an uncertain relation between variables X and Y. To extend the theory discussed so far to multivariate problems, the following operations are defined [16, 57]:

1. Cylindrical extension to convert a mass function to a mass function on a larger space;
2. Marginalization to convert a mass function to a mass function on a smaller space;
3. Ballooning extension to convert conditional information to a mass function on the joint space.

Cylindrical extension [16]: Let m^{Θ_X} be a mass function on Θ_X. To extend this information to the space $\Theta_X \times \Theta_Y$, we use the cylindrical extension defined as:

$$m^{\Theta_X \uparrow \Theta_X \times \Theta_Y}(z) = \begin{cases} m^{\Theta_X}(x) & \text{if } z = x \times \Theta_Y \\ 0 & \text{otherwise} \end{cases} \quad \forall z \subseteq \Theta_X \times \Theta_Y \tag{B.5}$$

Marginalization [16]: Let $m^{\Theta_X \times \Theta_Y}$ be a mass function on $\Theta_X \times \Theta_Y$. The marginal mass function $m^{\Theta_X \times \Theta_Y \downarrow \Theta_Y}$ on Θ_Y is defined as:

$$m^{\Theta_X \times \Theta_Y \downarrow \Theta_Y}(y) = \sum_{z \subseteq \Theta_X \times \Theta_Y | \text{Proj}(z \downarrow \Theta_Y) = y} m^{\Theta_X \times \Theta_Y}(z) \tag{B.6}$$
$$\forall y \subseteq \Theta_Y$$

with $\text{Proj}(z \downarrow \Theta_Y) = \{y \in \Theta_Y | \exists x \in \Theta_X, (y, x) \in z\}$

Ballooning extension [58]: Let $m^{\Theta_Y}[x]$ denote the conditional mass function on Θ_Y given $x \subseteq \Theta_X$. The ballooning extension of $m^{\Theta_Y}[x]$ on the space $\Theta_X \times \Theta_Y$ is the least committed mass function whose conditioning on x yields $m^{\Theta_Y}[x]$. It is obtained as:

$$m^{\Theta_Y}[x]^{\Uparrow \Theta_X \times \Theta_Y}(z) = \mathbb{1}_Z \cdot m^{\Theta_Y}[x](y), \quad \forall z \subseteq \Theta_X \times \Theta_Y, \tag{B.7}$$

with:

$$\mathbb{1}_Z = \begin{cases} 1 & \text{if } z = (x \times y) \cup (\bar{x} \times \Theta_Y), \\ 0 & \text{otherwise,} \end{cases}$$

Appendix B.2. Valuation networks

Valuation networks are a suitable graphical tool to represent uncertain knowledge in the form of belief functions [57, 59–61]. In contrast to Bayesian networks, which emphasize conditional independent relations, valuation networks emphasize factorizations of the joint distribution function. Formally, a valuation network can be regarded as a 3-tuple $(\mathbf{V}, \{\Theta_X\}_{X \in \mathbf{V}}, \{W_1, ..., W_m\})$ with operators $\{\oplus, \downarrow\}$ [57], where:

1. $\mathbf{V}$ is the set of variables representing the universe of discourse
2. $\{\Theta_X\}$ is the set of frames associated with each variable $X \in \mathbf{V}$
3. $\{W_1, ..., W_m\}$ is a collection of valuations[6] defined on the subsets of variables
4. $\oplus$ is the combination operation. Intuitively, combination corresponds to the aggregation of knowledge
5. $\downarrow$ is the marginalization operation. Intuitively, marginalization corresponds to the coarsening of knowledge.

[6] A valuation is a function representing the relationship among the variables in its domain.

Table B.7: A comparison of Bayesian networks and valuation networks [57]

Feature	Bayesian network	Valuation network
Graphical structure		
1. Type of graph	directed acyclic graph	hypergraph
2. Relations	conditional independence relations	joint form
3. Nodes	random variables	variables & valuations
Inference procedure		
4. Type of uncertainty	probabilistic	several
5. Inference	quantitative based on probability propagation	quantitative based on fusion algorithm

When the uncertainty is represented by belief functions, the valuations are multivariate basic belief assignments, and the combination operator corresponds to the conjunctive rule of combination (see Section Appendix B.3).

Graphically, there are two types of vertices in a valuation network. One set of vertices represents variables, indicated by circles, and the other set represents valuations, indicated by diamonds. In a valuation network, there are edges only between variables and valuations. There is an edge between a variables and a valuation if and only if the variable is in the domain of the valuation.

A comparison of the Bayesian networks and the valuation networks is given in Table B.7.

Appendix B.3. Reasoning under uncertainty

When new evidences become available, this is incorporated by combining the existing mass function with the mass function describing the new evidence. Consider two distinct mass functions $m_1^{\Theta_\mathbf{X}}$ and $m_2^{\Theta_\mathbf{X}}$ on $\Theta_\mathbf{X}$. The belief function $m^{\Theta_\mathbf{X}}$ that quantifies the combined impact of the two mass functions according to Dempster's original rule of combination is: $\forall \mathbf{x} \subseteq \Theta_\mathbf{X}$

$$m^{\Theta_\mathbf{X}}(\mathbf{x}) = \begin{cases} 0 & \text{if } A = \emptyset, \\ k \sum_{\mathbf{x}' \cap \mathbf{x}'' = \mathbf{x}} m_1^{\Theta_\mathbf{X}}(\mathbf{x}') m_2^{\Theta_\mathbf{X}}(\mathbf{x}'') & \text{otherwise,} \end{cases} \tag{B.8}$$

with $m_1^{\Theta_\mathbf{X}}$ and $m_2^{\Theta_\mathbf{X}}$ two mass functions on the same (multivariate) space $\Theta_\mathbf{X}$, $m^{\Theta_\mathbf{X}}$ the combined mass function, and k a normalization constant.

In the TBM, an open world assumption is allowed. In this case, two pieces of evidence are combined using the conjunctive rule of combination, which is an unnormalized form of Dempster's original rule of combination [62, 63]:

$$m^{\Theta_\mathbf{X}}(\mathbf{x}) = \sum_{\mathbf{x}' \cap \mathbf{x}'' = \mathbf{x}} m_1^{\Theta_\mathbf{X}}(\mathbf{x}') m_2^{\Theta_\mathbf{X}}(\mathbf{x}'') \tag{B.9}$$

The mass assigned to the empty set can be regarded as a measure of conflict. Combination rules (B.8) and (B.9) assume that the two sources $m_1^{\Theta_\mathbf{X}}$ and $m_2^{\Theta_\mathbf{X}}$ are both reliable and independent. Alternative combination rules, e.g. the disjunctive rule of combination and the cautious rule of combination, have been proposed to handle dependent and unreliable sources of evidence [64]. A detailed discussion about combination rules is beyond the scope of this paper. For a more thorough discussion, we refer the interested reader to [62, 63, 65].

Appendix B.4. Decision making

In the TBM, decisions are made by transforming the mass distribution to a probability distribution and then applying the expected-utility theory (see Section Appendix A.4). Belief masses are transformed to probabilities using the pignistic transformation [66]:

$$P_{\text{pig}}(x_i) = \sum_{x \subseteq \Theta_X} \frac{|x_i \cap x|}{x} \frac{m^{\Theta_X}(x)}{1 - m^{\Theta_X}(\emptyset)} \tag{B.10}$$

So, the mass allocated to a non-singleton set x is proportionally divided among the singleton elements in x, and the mass allocated to the empty set is proportionally distributed among all focal sets[7].

[7]The focal sets of a bba m are all subsets $A \subseteq \Theta$ for which holds $m(A) > 0$ [67].