Facilitating the elicitation of beliefs for use in Bayesian Belief modelling

Kirsty L. Hassall [a,*], Gordon Dailey [b], Joanna Zawadzka [c], Alice E. Milne [b], Jim A. Harris [c],
Ron Corstanje [c], Andrew P. Whitmore [b]

[a] *Computational and Analytical Sciences, Rothamsted Research, Harpenden, AL5 2JQ, UK*
[b] *Sustainable Agricultural Systems, Rothamsted Research, Harpenden, AL5 2JQ, UK*
[c] *Centre for Environmental and Agricultural Informatics, Cranfield University, Cranfield, MK43 0AL, UK*

ARTICLE INFO

ABSTRACT

Expert opinion is increasingly being used to inform Bayesian Belief Networks, in particular to define the conditional dependencies modelled by the graphical structure. The elicitation of such expert opinion remains a major challenge due to both the quantity of information required and the ability of experts to quantify subjective beliefs effectively. In this work, we introduce a method designed to initialise conditional probability tables based on a small number of simple questions that capture the overall shape of a conditional probability distribution before enabling the expert to refine their results in an efficient way. These methods have been incorporated into a software Application for Conditional probability Elicitation (ACE), freely available at https://github.com/KirstyLHassall/ACE (Hassall, 2019).

1. Introduction

Bayesian Belief Nets, also referred to as Bayes Nets, Belief Networks or often simply BBNs, have, in recent years, seen a dramatic increase in their use for describing and modelling natural systems. Examples include quantifying the risk of erosion in peat bogs (Aalders et al., 2011), modelling ecosystem services (Haines-Young, 2011), applications in natural resource management (see Henriksen et al., 2012, and references therein), mapping risks of soil threats such as soil compaction (Troldborg et al., 2013), predicting soil bulk density at landscape scales (Taalab et al., 2015) and assessing the impact of buffer zones on water protection and biodiversity (Tattari et al., 2003). This explosion in practical BBN modelling may in part be due to the relative simplicity of the intuitive graphical representation of multiple interrelated variables captured through conditional probabilities and more practically, the increasing accessibility to specialist BBN software.

There has been much work in the development of BBN methodology to address the practicalities of BBN modelling (see, for example, Marcot, 2017). BBN modelling largely consists of two interrelated steps; defining the graphical structure and quantifying the form of the conditional dependencies through conditional probability distributions. In each step, one can incorporate both data and expert opinion (Pollino et al., 2007; Aalders et al., 2011). Our interest is to define a fully parameterised BBN that quantifies soil health by capturing the inherent knowledge of experts representative of all aspects of soil science such

as soil microbiology, soil chemistry, soil physics and land management, among others. A full description of the construction and evolution of these soil health network structures will be the topic of a future paper, although an early example is shown in Fig. 2 for demonstrative purposes. In this paper, we focus on the key issues surrounding the use of expert opinion in the characterisation of conditional probability tables.[1]

BBNs first saw an increase in popularity in the 1980s with applications to decision support and expert systems (see, for example, Pearl, 1986; Lauritzen and Spiegelhalter, 1988), whereby a causal network description is used to express expert knowledge that can inform diagnoses and decisions. In this work, our motivation for using BBNs differs from both the practical modelling approaches prolific in the literature and also the traditional usage in expert systems. Specifically, there is an increasing desire to be able to derive and quantify metrics for (often) subjective concepts. An example of such is soil quality and health, which is a term frequently used, albeit qualitatively and subjectively (Wienhold et al., 2004; Doran and Zeiss, 2000). Definitions of soil quality and health vary in the literature but include: the condition or state of soil relative to the requirements of one or more biotic species and/or to any human need or purpose (Johnson et al., 1997); the capacity of a specific kind of soil to function, within natural or managed ecosystem boundaries, to sustain plant and animal productivity, maintain or enhance water and air quality, and support human health

---

* Corresponding author.
E-mail address: kirsty.hassall@rothamsted.ac.uk (K.L. Hassall).
[1] Conditional probability tables or CPTs, described fully in Section 1.1.1 parameterise a BBN by quantifying the relationship between variables.
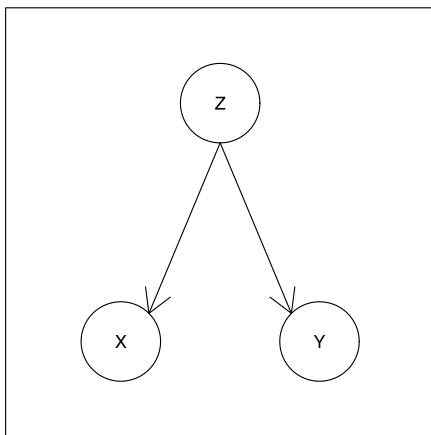
Fig. 1. An illustrative BBN, showing three nodes X, Y, Z connected by two edges where X and Y are conditionally independent given the parent, Z.

and habitation (Friedman et al., 2001; Karlen et al., 1997); the soil's ability to provide ecosystem and social services through its capacities to perform its functions under changing conditions (Tóth et al., 2007). Since these definitions include inherently subjective concepts, quantifying them cannot be done without the use of expert opinion. BBNs are a natural framework for incorporating both expert opinion and data to conceptualise different model systems, (see, for example, Pollino et al., 2007; Aalders et al., 2011).

A BBN is a graphical model made up of nodes representing variables of interest and edges representing direct dependencies between variables. Specifically, each edge represents a conditional distribution, with nodes that are not connected considered to be conditionally independent given all other nodes in the network. BBNs are more precisely directed acyclic graphs (DAGs) meaning each edge has a direction and there are no feedback loops in the network. The direction of each arrow represents the direction of conditioning and the node at the source of the arrow is referred to as a parent and the node at the sink of the arrow, the child. The joint distribution of all the variables within the graphical model can be represented as the product of the conditional distributions,

$$f_{X_1,\dots,X_p}(x_1,\dots,x_p) = \prod_i f_{X_i|\text{Parents}\{X_i\}}(x_i|\text{parents}\{x_i\}),$$

where $\text{Parents}\{X_i\}$ denotes the set of nodes connected by a directed edge to the node i. For example, Fig. 1, shows three nodes X, Y, Z connected by two edges where X and Y are conditionally independent given the parent, Z. Moreover, there exists a dependency between X and Z and between Y and Z. The graph shown in Fig. 1 has joint distribution $f_{X,Y,Z}(x,y,z)$, which can be decomposed as,

$$f_{X,Y,Z}(x,y,z) = f_{X|Z}(x|z)f_{Y|Z}(y|z)f_Z(z).$$

Although the distribution of each node in a BBN can be general, for the remainder of this paper, we consider only discrete BBNs where all random variables $X_1,\dots,X_p$ are categorical.

However, we note here a discrepancy in the literature regarding BBNs and the term "Bayesian" (Korb and Nicholson, 2004). The notion of a "prior" in a BBN often refers to the distribution of an ancestral node. This does not preclude the use of expert derived opinion, but rather, the opinions or beliefs are used to directly inform the (conditional) distribution of each node. This differs from what we might term the "truly Bayesian" approach which would consider prior information to be included through a hyperdistribution over the parameters of the node distribution, i.e. the likelihood of node X_i, $f_{X_i|\text{Parents}\{X_i\}}(x_i|\text{parents}\{x_i\},\theta)$, is defined through the parameters θ, prior information is defined through the distribution over θ, and

interest is in deriving a posterior distribution for $\theta|X = x$. More explicitly, for a discrete node X_i with a categorical multinomial distribution defined by a set of probabilities, a prior distribution is defined over this set of probabilities, by e.g. a Dirichlet distribution. In this way, expert opinion or belief would be used to parameterise the Dirichlet prior which when combined with the likelihood, gives the posterior. As with most practical BBNs, our focus is on the direct frequency representation and not the fully Bayesian approach.

1.1. Example BBNs

Throughout the remainder of this paper, we will demonstrate our methods on the following BBN application, shown in Fig. 3. This is an illustrative example aiming to define the concept of Road Safety. This example was developed predominantly to aid in the exposition of Bayesian Belief Networks to the subject specific experts we approached to define soil health and quality. As an aside, introducing the topic of BBNs with an example unrelated to the topic we wished to focus on, i.e. soil health, was useful to put across the main concepts (e.g. of conditional dependence and conditional independence) without prejudicing the question at hand.

Fig. 3 shows how the subjective notion of road safety can be defined by four variables; the presence of cycle lanes, whether the road is on a main school route, the number of car crashes and the number of fatalities. Furthermore, the net shows how, if data are not available on the number of car crashes, we can infer this from other causally related variables such as the speed limit and weather conditions. Moreover, our example stresses the point that graphical representations do not need to be causal, specifically, we include the (somewhat artificial) example that if data were available on the number of umbrella sales, this could be used to infer the weather conditions which, in turn, can be used to infer the number of car crashes.

1.1.1. Conditional probability tables

The conditional probability table (CPT) describes the distribution of the child node for every combination of states of the parent nodes. For example, shown in Table 1 is an example CPT for the node representing the number of car crashes in the road safety network. The top row describes the distribution of the number of car crashes when visibility is poor, there is no surface water and the speed limit is 30.

When a child node has multiple parents, the number of entries in the conditional probability table can quickly become very large. Moreover, the interdependence between the parent states can be difficult to identify. Thus motivating our research into the development of methods to aid the quantification of conditional probability tables from expert opinion.

2. Methods

2.1. BBN conditional probability elicitation

Despite the vast amount of literature and research into elicitation techniques (see supplementary information Appendices A and B), it remains challenging to elicit conditional probability tables in practical BBN modelling. The difficulty in eliciting CPTs is predominantly due to the sheer number of CPT entries required in practical BBNs, often in the order of 100s of individual probability estimates. This hampers the use of digital tools that require a separate input for every row of a CPT (e.g. Spaccasassi and Deleris, 2011). Thus, many practical BBN papers resort to simply asking "How many times out of 100…" or "What is the probability variable X takes value x for information Y…" repeatedly without further aids for the numerical elicitation process (see, for example, Taalab et al., 2015; Pollino et al., 2007).

However, elicitation in this way can easily produce unrepresentative tables, not least of all due to the sheer number of estimates required.
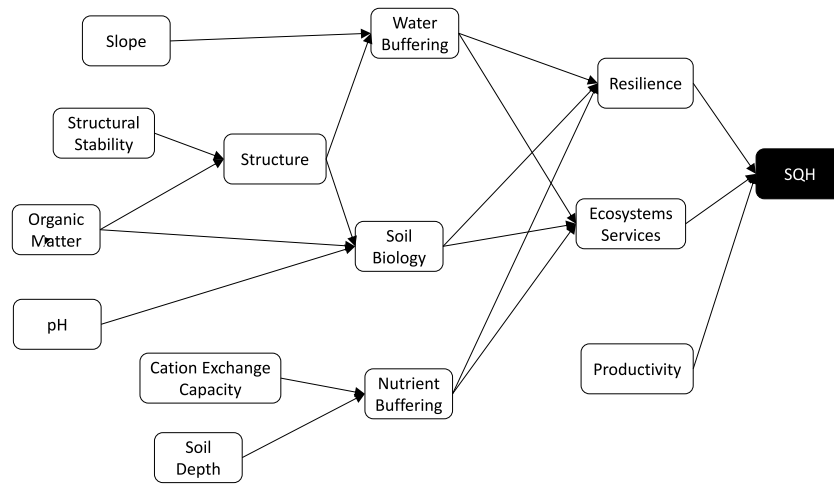
Fig. 2. An illustrative Bayesian Belief Network constructed to define the subjective concept of soil quality and health (SQH).
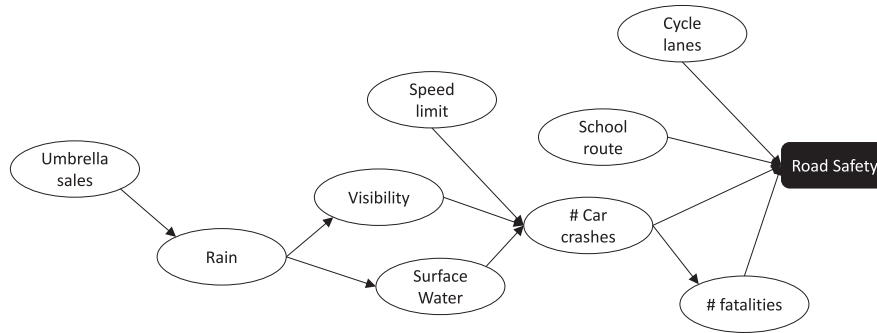

Fig. 3. An illustrative Bayesian Belief Network constructed to define the subjective concept of road safety.

Table 1
Example conditional probability table for the Car Crash node of the road safety network in Fig. 3. $p_i|\{j,k,l\}$ is the probability that the number of car crashes will be in state i for Visibility = j, Surface Water = k and Speed Limit = l, where j = Poor or Good, k = None, Some, Lots and $l = 30, 40, 50, 60$.

Visibility	Surface water	Speed limit	Number of car crashes		
			0–4	5–10	>10
Poor	None	30	$p_1|\{Poor, None, 30\}$	$p_2|\{Poor, None, 30\}$	$p_3|\{Poor, None, 30\}$
Poor	None	40	$p_1|\{Poor, None, 40\}$	$p_2|\{Poor, None, 40\}$	$p_3|\{Poor, None, 40\}$
Poor	None	50	$p_1|\{Poor, None, 50\}$	$p_2|\{Poor, None, 50\}$	$p_3|\{Poor, None, 50\}$
Poor	None	60	$p_1|\{Poor, None, 60\}$	$p_2|\{Poor, None, 60\}$	$p_3|\{Poor, None, 60\}$
Poor	Some	30	$p_1|\{Poor, Some, 30\}$	$p_2|\{Poor, Some, 30\}$	$p_3|\{Poor, Some, 30\}$
Poor	Some	40	$p_1|\{Poor, Some, 40\}$	$p_2|\{Poor, Some, 40\}$	$p_3|\{Poor, Some, 40\}$
Poor	Some	50	$p_1|\{Poor, Some, 50\}$	$p_2|\{Poor, Some, 50\}$	$p_3|\{Poor, Some, 50\}$
Poor	Some	60	$p_1|\{Poor, Some, 60\}$	$p_2|\{Poor, Some, 60\}$	$p_3|\{Poor, Some, 60\}$
Poor	Lots	30	$p_1|\{Poor, Lots, 30\}$	$p_2|\{Poor, Lots, 30\}$	$p_3|\{Poor, Lots, 30\}$
Poor	Lots	40	$p_1|\{Poor, Lots, 40\}$	$p_2|\{Poor, Lots, 40\}$	$p_3|\{Poor, Lots, 40\}$
Poor	Lots	50	$p_1|\{Poor, Lots, 50\}$	$p_2|\{Poor, Lots, 50\}$	$p_3|\{Poor, Lots, 50\}$
Poor	Lots	60	$p_1|\{Poor, Lots, 60\}$	$p_2|\{Poor, Lots, 60\}$	$p_3|\{Poor, Lots, 60\}$
Good	None	30	$p_1|\{Good, None, 30\}$	$p_2|\{Good, None, 30\}$	$p_3|\{Good, None, 30\}$
Good	None	40	$p_1|\{Good, None, 40\}$	$p_2|\{Good, None, 40\}$	$p_3|\{Good, None, 40\}$
Good	None	50	$p_1|\{Good, None, 50\}$	$p_2|\{Good, None, 50\}$	$p_3|\{Good, None, 50\}$
Good	None	60	$p_1|\{Good, None, 60\}$	$p_2|\{Good, None, 60\}$	$p_3|\{Good, None, 60\}$
Good	Some	30	$p_1|\{Good, Some, 30\}$	$p_2|\{Good, Some, 30\}$	$p_3|\{Good, Some, 30\}$
Good	Some	40	$p_1|\{Good, Some, 40\}$	$p_2|\{Good, Some, 40\}$	$p_3|\{Good, Some, 40\}$
Good	Some	50	$p_1|\{Good, Some, 50\}$	$p_2|\{Good, Some, 50\}$	$p_3|\{Good, Some, 50\}$
Good	Some	60	$p_1|\{Good, Some, 60\}$	$p_2|\{Good, Some, 60\}$	$p_3|\{Good, Some, 60\}$
Good	Lots	30	$p_1|\{Good, Lots, 30\}$	$p_2|\{Good, Lots, 30\}$	$p_3|\{Good, Lots, 30\}$
Good	Lots	40	$p_1|\{Good, Lots, 40\}$	$p_2|\{Good, Lots, 40\}$	$p_3|\{Good, Lots, 40\}$
Good	Lots	50	$p_1|\{Good, Lots, 50\}$	$p_2|\{Good, Lots, 50\}$	$p_3|\{Good, Lots, 50\}$
Good	Lots	60	$p_1|\{Good, Lots, 60\}$	$p_2|\{Good, Lots, 60\}$	$p_3|\{Good, Lots, 60\}$

For example, the relatively simple road safety network in Fig. 3 consists of 10 nodes. The corresponding 6 conditional probability tables produce almost 70 individual combinations of parent states across all CPTs, each of which is associated with a distribution of the corresponding child node to be elicited. In total, this results in more than 150 individual probability estimates required from an elicitee. Even for this
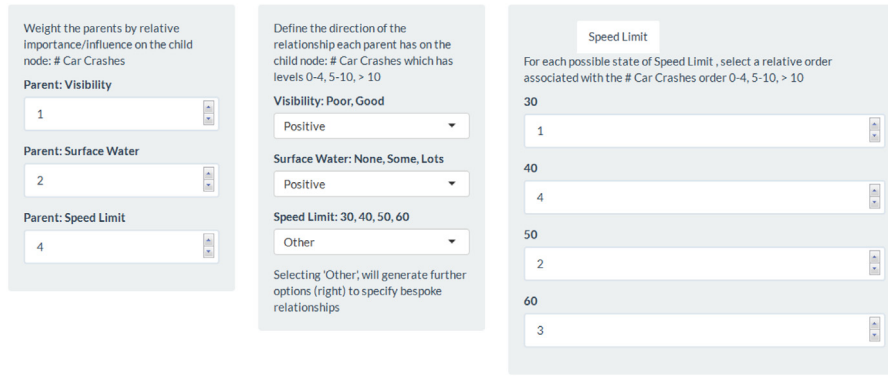
Fig. 4. Screenshots of the ACE software demonstrating the specification of relative parental importance and the associated relationship required to construct the score and initialise the conditional probability table.

very simple network, it can be difficult to complete all necessary CPTs both because it is time-consuming and also because it is difficult to maintain concentration consistently over so many distributions.

Perhaps more importantly, viewing each row of a CPT independently can make it very difficult to characterise an expert's belief about the relative changes in the different parent states (see, for example, Marcot, 2017).

Our approach addresses these issues through the definition of a simple scoring system based on two questions per parent node that is then used to initialise a CPT. This initial CPT provides a logical starting point for experts to fine-tune whilst ensuring the higher-level relationships between parent nodes has been efficiently encapsulated. This has been implemented in ACE, a freely available R-Shiny software Application for Conditional probability Elicitation (Hassall, 2019).

2.2. CPT Scoring algorithm

To specify a score that captures the relative effects of different parent nodes, an expert first assigns a weight of relative importance to each parent node. This weighting is used to define the relative effects of each parent on the probability distribution of the child node. Parents with a larger weight are assigned a greater level of influence in determining the conditional probability table such that changes in the states of the parent with the largest weight will result in the biggest differences in the distribution of the child node.

The second step is to define the direction of the relationship between each parent and child. Each parent can have either a positive, negative or other relationship with the child node. A relationship is considered positive if, as the states of the parent changes according to the order they have been defined, the probability the child node is in its higher states also increases. Conversely, a negative relationship is appropriate if as the states of the parent changes according to the order they have been defined, the probability the child node is in its higher states decreases. Not every parent–child relationship can be categorised as having either a positive or negative relationship. Although it is impractical to incorporate a full set of relationships into the software implementation, we have instead incorporated the option to define an "Other" relationship. This enables experts to define a "relative order" to the states of the parent node. This relative order describes the order of the parent states that would result in an increasing probability that the child node is in its higher states.

Fig. 4, shows an example of specifying this information into the ACE software for the car crash node of the road safety network. Here, a change in the state of Surface Water will have an effect twice as large as a change in the state of Visibility. The effect of both Visibility and Surface Water on the increasing number of Car Crashes is positive, whilst the effect of Speed Limit is defined as non-monotonic. Specifically, the relative order of Speed Limit states describes the order that

would result in an increasing probability that the child node is in its higher states. In the example depicted in Fig. 4, a speed limit of 30 [order 1] is associated with the fewest car crashes, speed limits of 50 and 60 are associated with some car crashes and a speed limit of 40 [order 4] is associated with the most car crashes.

Mathematically, this relative weighting and order relationship defines a score, from which an initial draft CPT is created. Let P_{ij} denote the score of the jth state of the ith parent, which is given by,

$$
P_{ij} = \begin{cases}
\frac{j-1}{n_i-1} & \text{if Parent } i \text{ has a positive relationship with the child node} \\[2mm]
\frac{n_i-j}{n_i-1} & \text{if Parent } i \text{ has a negative relationship with the child node} \\[2mm]
\frac{ord[j]-1}{n_i-1} & \text{if Parent } i \text{ has an "other" relationship with the child node}
\end{cases}
$$

where n_i is the number of states of parent i and $ord[j]$ denotes the ordered index of state j.

An overall score is then calculated for each combination of parent states, given by a weighted average of the constituent scores,

$$
Score_{\{k\}} = \frac{\sum_i w_i P_{i\{k\}}}{\sum_i w_i},
$$

where w_i is the weight associated with Parent i, and $\{k\}$ is the kth combination of parent states, with $P_{i\{k\}}$ denoting the associated score of parent i for combination k.

For a child node with two states, this score will correspond to the probability the child node is in its highest state. For a child node with $M > 2$ states, a conversion is made for each parent combination $\{k\}$. Specifically, the probability that the child node is in state m is given by twice the area of the mth trapezium formed when the linear line between the two probabilities of a corresponding two state child is cut into M equal intervals. This is depicted in Fig. 5. For a $Score_{\{k\}} = 0.8$, the distribution of a child node with two states is given by Prob(Child = State 1) = 0.2 and Prob(Child = State 2) = 0.8. For a $Score_{\{k\}} = 0.8$, the distribution of a child node with four states, a $Score_{\{k\}} = 0.8$, corresponds to a distribution of the child node of Prob(Child = State 1) =0.1375, Prob(Child = State 2) =0.2125, Prob(Child = State 3)=0.2875 and Prob(Child = State 4) = 0.3625.

This scoring system assumes a) that all states can be considered on an equally spaced linear scale and b) that the range of CPT rows for a two-state child node will contain values in the full range of $0 - 100\%$. These assumptions act as a constraint on the construction of the scores which can be relaxed if needed within the underlying open source code. Due to the construction of this score, one major limitation is in the mapping depicted in Fig. 5. For child nodes with an odd number of states (M), the middle category will always have a probability of $1/M$. We reiterate here, that this score is not designed to fully define a CPT, but rather to provide an initialisation that captures the relative effects
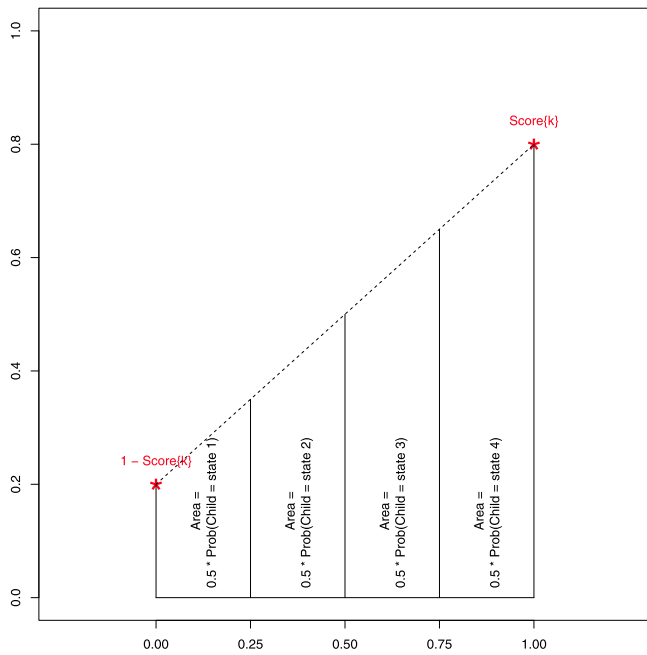
Fig. 5. An illustration of the conversion between a single score (in this case, $Score_{\{k\}} = 0.8$) to a probability distribution of the child node.

of the parent nodes whilst still enabling experts to refine their beliefs through individual edits.

The ACE software aids the process of elicitation by firstly using the above scoring system to provide a logical initial CPT based on an expert's belief of the overall effect of the different parent nodes. Secondly, the ACE software provides a fully editable interface with the initialised CPT for an expert to fine-tune and edit individual values. Furthermore, the software has been encoded with a number of warning messages that check the CPT for incongruities for users to refine. For example, if all "middle categories" are left unchanged, an appropriate warning message is shown.

To further aid the elicitation, a visual aid on a relative frequency scale is provided as shown in Fig. 6. This allows a user to visualise the full conditional distribution as well as to see the relative changes in the child node for different combinations of parent states. This graphical representation of the CPT can be reordered according to the relative weights defined for the parent set, thus providing a more intuitive display of the overall effects of the parent nodes.

2.3. Quantifying uncertainty

Throughout we have focused on the elicitation of CPTs through the quantification of an expert's frequency distribution for a particular scenario, the aleatoric uncertainty. This does not capture a user's uncertainty in the resulting estimate. As discussed in the appendix, there are approaches available in the literature that look at formally quantifying the additional epistemic uncertainty in an expert's belief. In a practical BBN application, this would involve the elicitation of a multi-dimensional hyperdistribution for each combination of parent states and rapidly becomes infeasible. However, we do believe it is important to capture this epistemic uncertainty, as expressed in Marcot (2017). As a pragmatic approach, we included the notion of confidence in a user's estimate. A confidence value can take one of three qualitative values:

Low: Low confidence in the final beliefs and the expert would consider it likely the values could vary.
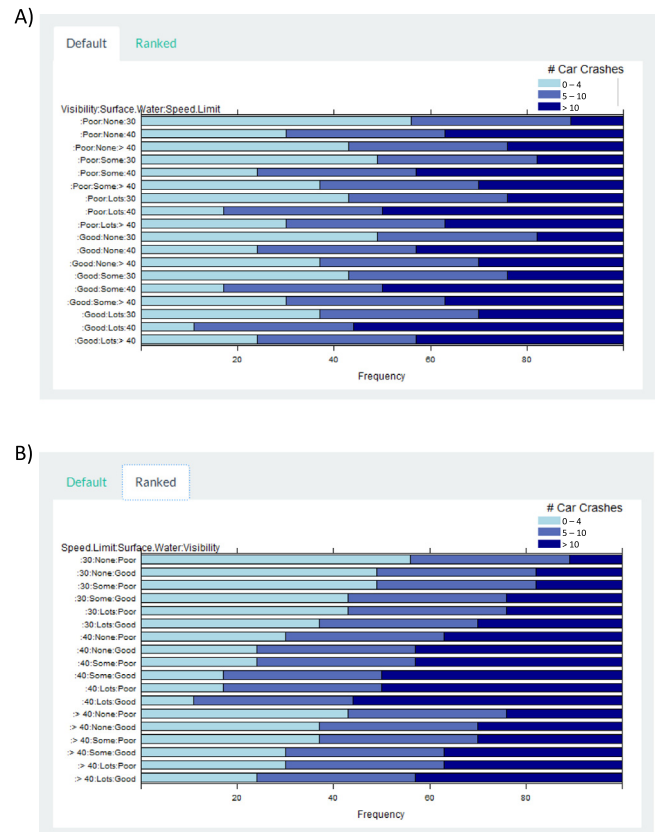

Fig. 6. Visual aid of a conditional probability table available in ACE. Each bar represents the probability distribution of the child node for each combination of parent states. A) orders the bars according to the default order of parent nodes. B) orders the bars by the weight of each parent node.

Medium: Reasonably confident/certain in the final beliefs, although the final values could vary.

High: Highly confident in the final beliefs, and the expert would not consider it likely for these values to vary much.

As stated above, this definition of uncertainty in the probability estimate is pragmatic and qualitative. More sophisticated measures include those aiming to numerically quantify the uncertainty in an estimate through the identification of quantiles of the distribution as developed in the SHELF methodology (Oakley and O'Hagan, 2016), IDEA (Hanea et al., 2017), and approaches using Cooke's Classical Model (Cooke, 1991; Aspinall and Cooke, 2013) and those that identify confidence intervals of the estimates (see e.g. Christophersen et al. (2018)). Thus, users and developers of ACE can access the open source code to incorporate more precise definitions of uncertainty.

Notions of confidence can be incorporated in the downstream analysis of an elicited BBN in multiple ways. For example, Pollino et al. (2007) combined the estimates for each probability from multiple experts through a weighting associated with the confidence. In this way, confidence was used to form an equivalent sample size for each experts' contribution. In comparison, Van Allen et al. (2008) showed how with an assessment of variance, error bars can be incorporated into the BBN. This does however, rely on a numerical estimate of variance rather than a qualitative assessment. An alternative, would be to use the confidence as a form of sensitivity analysis, e.g. through a Monte Carlo simulation study, perturbing the derived conditional distributions relative to their associated confidence.
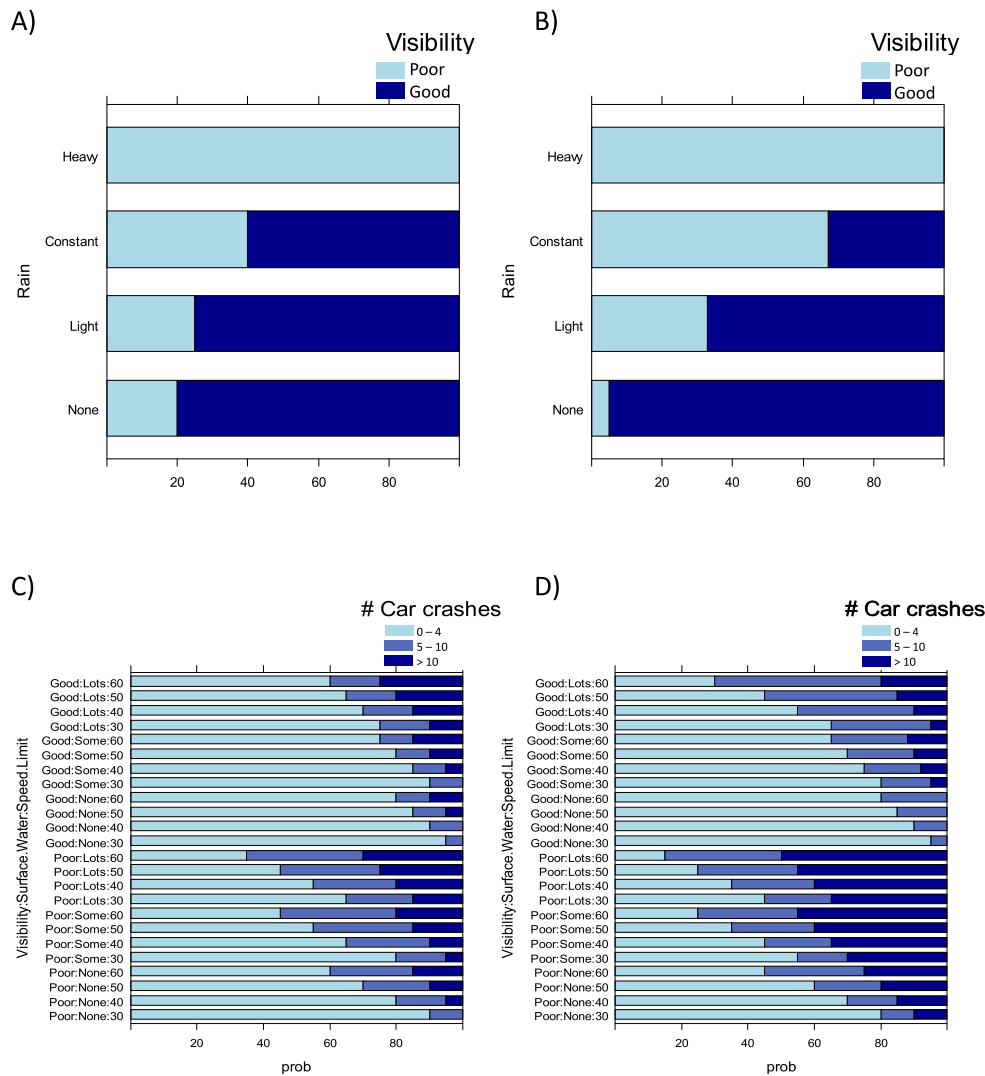
Fig. 7. The elicited distribution for Visibility (A–B) conditional on the parent node Rain. (A) was elicited using the manual process and (B) using the automatic scoring approach. The elicited distribution for the Number of car crashes (C–D) conditional on the parent nodes Visibility, Surface Water and Speed limit. (C) was elicited using the manual process and (D) using the automatic scoring approach.

3. Results

To investigate the potential efficiency gains in using the scoring system described above to initialise CPTs, we recruited 8 volunteers to test our methods using the ACE software implementation. Each volunteer received training on graphical modelling and the association with conditional independence along with a description of the road safety network shown in Fig. 3. The volunteers were then allocated to 1 of 2 groups. Group A were given 25 minutes to complete as many CPTs of the road safety network (in a prespecified order) as they could using the scoring system described above. After a short interlude, they were then given 25 minutes to fill in as many CPTs (in the same order as before) without using the scoring system, i.e. to fill in the tables manually, although a graphical aid remained accessible in the software. Group B had the same tasks but in the reverse order, i.e. to first fill the tables in manually and secondly to fill the table in using the scoring system.

In generality, the majority of tables were found to be consistent across the two methods. It was notable that the automated scoring algorithm enabled more tables to be completed in the given time, however, some discrepancies between the methods were seen. Three types of discrepancy were found:

1. The automatic method could result in a more "linear" distribution compared to the manual process. This is illustrated in Fig. 7A) - B) which shows the elicited distribution for Visibility conditional on the parent node of Rain under the two approaches. This suggests a tendency to stick with the default values when using the automated approach.

2. Although the shape of the distributions closely match, there is a shift in the location. This is demonstrated in Fig. 7C)–D) which shows the elicited distribution for the Number of car crashes conditional on the parent nodes Visibility, Surface Water and Speed limit under the two approaches. It can be seen that under the manual approach, there is a consistently larger proportion of the distribution located in the 0–4 crash category compared to the automated approach. For this particular example, it demonstrates the limitations of the linear mapping of the score function. In particular, to obtain the same distribution as obtained under the manual process, a non-linear mapping of the score to three states would be required.

3. The consistency of the relative importance of parent nodes. By construction, the default of the automated approach is to ensure the relative importance of the parent nodes is consistent over all combinations of parent node states. This was not always

observed in the corresponding manual tables. Following up on these discrepancies, different views were expressed as to which table best represents the true belief;

(a) Mistakes were made in the manual process, due to (i) too many scenarios to follow the relative importance through logically, (ii) a difficulty in expressing the relative importance in the parent nodes, which could result in an equal weight given to each parent.

(b) The automated table was not edited to reflect the scenarios which do not follow the general trend. A specific example was given that for a road with a high number of crashes (>10) and more than 1 fatality, the likelihood the road safety is good decreases when a cycle lane is present compared to being absent, whereas in all other scenarios, the presence of cycle lane increases the likelihood of good road safety. This particular example was captured in the manual process but not in the automated process as the overall trend of cycle presence increasing the likelihood of good road safety was used throughout all scenarios.

It should be noted, that one of the drawbacks to BBNs in general is the sheer number of distributions to be elicited. We have found that elicitees fatigue with this process (regardless of method) which may also be reflected in the discrepancies observed above. It is difficult to motivate an individual to repeat the process twice with equal attentiveness.

After the volunteers completed the comparative exercise, they were asked the following three questions:

1. In general, did you find it easy or difficult to quantify your beliefs numerically?
2. Which method did you prefer?
3. Why?

Unanimously, the automated method was preferred with the general process of numerical elicitation found to be difficult. In particular, most participants greatly favoured the automated process for nodes with multiple parents as they found it very difficult to translate the interrelationships between parents into a set of well-defined consistent probability distributions. This did sometimes come with the caveat that the manual process may have produced a more detailed representation.

Although, the study presented comes from a small set of volunteers, both the manual elicitation and the automatic initialisation were used in expert elicitation workshops we have run to formulate a working definition of soil quality and health. Over four workshops, 27 (13) experts from across soil science and associated disciplines used the automated (manual) approach to initialise the CPTs, respectively. The findings of the above study are largely representative of what we observed through these workshops, particularly the ability for experts to complete the CPTs in the time allowed for in the workshop; the appreciation in being able to initialise a full CPT with the relative importance of parents defined consistently; the observed linearity in the distributions from the automated approach and the fact that few edits were made to the initialised CPTs. A full description of these workshops and associated findings will be the topic of a future paper.

3.1. Additional guidance

In implementing the ACE software for capturing expert opinion, we provide some additional guidance based on our own experiences.

- Avoid double negatives in the definition of nodes and their associated states
- A consistent ordering in node states can be particularly helpful when defining the direction of a relationship, e.g. A Cycle Lane node with states $\{Absent, Present\}$ resulted in fewer mistakes compared to node states defined by $\{Present, Absent\}$.

- It is encouraged to keep the number of child states to a minimum. This ensures a more dichotomous definition of the node. Many participants found a two-state definition particularly intuitive. For example, a natural definition of an ordinal node is to have three states $\{low, medium, high\}$. In our experience, we found that when experts were uncertain of the distribution of the child node, they would intuitively put the majority of the distribution into the medium category. However, when a node is defined to have two states $\{low, high\}$, an uncertain distribution would be intuitively reflected from an equal weight to each category. This highlights the importance of communicating what the distribution actually represents, a notion that is difficult for non-domain specific experts. As discussed in Christophersen et al. (2018) there are often circumstances where the discretisation of continuous variables is limiting and extensions to include continuous variables in a BBN are desirable.
- If more than two states are necessary for a child node, the scoring approach works best when there is an equidistant definition to the ordinal states of the child node.
- On hand facilitation is needed along with clear guidance on what the numerical quantities mean in terms of the practical application.
- Practice via training exercises is a fundamental necessity. In our experience, experts became much more comfortable with the concept of graphical representations after they had practised the process of capturing their belief in conditional probability tables. For example, prior to training, it was difficult to convince participants that a hierarchical graphical structure with intermediary parents was a desirable (and meaningful) structure (see e.g. Section 9.3.4.1 of Korb and Nicholson (2004)) until they had experienced the process of translating their beliefs into the large interdependent conditional probability tables.

4. Discussion

Ultimately, the optimal method by which the conditional probability tables are captured will differ depending on elicitee. In practice, we have found the automated method to be far less daunting to many experts who are not versed in probabilistic descriptions. In addition, the automated scoring enables an instinctive and qualitative knowledge to be captured numerically through a simpler definition of relative parent weights and directional relationships. The automated approach was found to greatly ease the process of elicitation but at the cost of specificity. Although designed purely as an approach to initialise the CPTs, we found in practice, relatively little editing of these initialised tables was done. Users tended to accept the prepopulated distributions and move on to the next CPT. Although we cannot say whether this phenomenon will occur in general, we found that it did occur both in our soil health workshops and the volunteer study presented in Section 3. This was primarily due to either time constraints with users daunted by the number of tables they needed to complete or low confidence in the elicited relationships with users opting for a generalised representation of their belief in the absence of any strong feelings to counter that state. If time is the limiting factor, our recommendation would be to encourage experts to identify the scenarios which deviate from the overall trend they have specified and to edit these specific individuals. If the issue is low confidence, this can be recorded directly in ACE for use in downstream analysis.

The main drawback to the automated approach is the prescriptive conversion of the derived score into a distribution over multiple states (Fig. 5). Many extensions to this score could be considered, predominantly to a non-linear conversion between the score and the frequency distribution over multiple states. However, these conversions may be difficult to convey to individual experts. An alternative approach would be to ensure the states of a node are defined to be equidistant, e.g. in the Road Safety example above, the states of the Number of car crashes

were defined as $\{0-4, 5-10, >10\}$, a more equidistant definition might be $\{0, 1-3, 4-10, >10\}$.

In all our studies, the visualisation aids have greatly facilitated the process of numerically quantifying beliefs and may be improved through further developments. It is well known that trellis graphics (Cleveland, 1993) provide an intuitive display of multiple interrelated variables and could be incorporated into the software. In addition, an interactive graphical display could further aid the insight obtained from visualising such CPTs.

It remains a major challenge to elicit CPTs under uncertainty. Work by Zapata-Vázquez et al. (2014) address this by extending the ideas within the SHELF package to elicit a Dirichlet distribution over the set of probabilities. However, applying this methodology to the number of cases in a typical BBN remains impractical, although certainly a desirable objective for the future.

5. Conclusion

We have investigated two approaches to filling in the conditional probability tables of a Bayesian Belief Net, a manual approach requiring the expert to consider how many times out of 100 would a particular outcome be expected from a set of specified scenarios and an automatic approach which initialises the CPT based on two simple questions before enabling further edits to be made.

The manual approach was found to work well for experts with a good quantitative background who were practised at translating relative relationships into numerical form. However, for many domain specific experts, quantifying interdependent relationships is incredibly difficult.

Through the development of the ACE software, we have provided the community with a digitised data capture method for recording conditional probability tables in conjunction with both visual and numerical aids. Moreover, through the development of a well-defined score, we have shown how a potentially large, complex set of interactions can be encapsulated in a CPT without having to specify the outcome of every single scenario. Automating this approach in freely available software allows its incorporation in many elicitation techniques, whether by group consensus; independently or through a Delphi recursive approach.

Although Bayesian Belief Networks are increasingly being used to describe and model both natural systems and public health concerns, robust expert elicitation of the CPTs remains a major bottleneck in the process. This implementation has substantially reduced the burden associated with filling in CPTs through expert derived opinion and provides efficient data collection for use in Bayesian Belief Networks. Thus, this methodology has wide applicability to the research community for modelling systems and developing policy support tools.

Software availability

All methods described in this paper have been incorporated into an open source software Application for Conditional probability Elicitation (ACE), freely available at https://github.com/KirstyLHassall/ACE (Hassall, 2019).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Funding

Appendix A. General elicitation methods

There is much debate around how expert opinion should be elicited, with some controversy surrounding the allowance for experts to discuss and revise opinions (see, for example, Cooke, 1991). Despite this, all approaches to structured elicitation aim to reduce known biases within the elicitation process. These biases include (Kuhnert et al., 2010; Tversky and Kahneman, 1973; Cooke, 1991);

- Availability bias, whereby an expert's response is based on most recent available information and not considering past events
- Hindsight bias, where too much emphasis is placed on past events
- Anchoring, the tendency to anchor around initial first guesses irrespective of the accuracy of the initial estimate
- Law of small numbers, where opinions are based on small pieces of evidence which are then extrapolated
- Representativeness, where opinions are based on situations that are rightly or wrongly perceived to be similar to the scenario in question.

Whether one chooses to use a process with discussion or not, depends on which biases are most likely to occur and whether the group dynamics will add to or lessen such biases. For example, the presence of a single "strong" opinion may cause anchoring around this opinion.

The other consideration when choosing the elicitation approach is around how opinions from multiple experts will be combined. Opinions from multiple experts can be combined either through allowing a group of experts to reach a group consensus through repeated revisions and discussion or through mathematical aggregation. There is evidence to suggest that allowing experts to interact and discuss may impact the validity of mathematical aggregation as it induces a dependence between responses (see Hanea et al., 2017, and references therein).

Appendix B. Numerical elicitation methods

There is an additional layer of complexity when it comes to eliciting quantitative responses that characterise the relationships between variables. It is well known that humans are inherently poor at estimating numerical quantities. To this end, there has been much work to overcome these shortcomings (Kuhnert et al., 2010). Research has focused on the task of effectively and accurately eliciting estimates of proportions from different experts. For example, it is considered that expressing this information in a frequency format enables more accurate estimates compared to specifying a proportion between 0 and 1 (Gigerenzer, 1996; Price, 1998). Other methods for eliciting proportions can be characterised through probability scale approaches and gambling methods (Jenkinson, 2005). Gambling methods assist experts to think about their probabilities in terms of an event upon which they might place a bet (see Cooke, 1991, and references therein). This can enable a greater engagement into the thinking behind each probability estimate. However, the basic gambling methods are focused on binary outcomes. In all cases, visual assessments are generally believed to greatly aid the elicitation process by allowing experts to "see" their beliefs quantified.

However, a conditional probability table consists of multiple probability entries. As discussed in Section 1.1.1, CPTs can become very large very quickly, resulting in 100s of individual probability estimates to be filled in. Werner et al. (2017) provides a review of the methods aiming at reducing this burden in completing CPTs. The methods vary

from piecewise interpolation based on the influence of parents (Wisse et al., 2008), a method similar to that developed in ACE, to making use of the causal structure in a BBN e.g. the noisy-OR and noisy-MAX methods (Pearl, 2014; Diez, 1993).

In addition, it is a highly active area of research that focuses on eliciting beliefs under uncertainty, through (for example) the methods of SHELF (Oakley and O'Hagan, 2016) and IDEA (Hanea et al., 2017). These methods look to ascertain a prior distribution for a parameter of interest based on an expert's belief, i.e. to isolate the epistemic uncertainty associated with a belief and separate this from the aleatoric uncertainty, the uncertainty due to randomness in the process.

References

Aalders, I., Hough, R., Towers, W., 2011. Risk of erosion in peat soils–an investigation using Bayesian belief networks. Soil Use Manag. 27 (4), 538–549.

Aspinall, W., Cooke, R., 2013. Quantifying scientific uncertainty from expert judgement elicitation. In: Risk and Uncertainty Assessment for Natural Hazards. Cambridge University Press Cambridge, UK, p. 64.

Christophersen, A., Deligne, N.I., Hanea, A.M., Chardot, L., Fournier, N., Aspinall, W.P., 2018. Bayesian network modeling and expert elicitation for probabilistic eruption forecasting: Pilot study for whakaari/white island, New Zealand. Front. Earth Sci. 6, 211.

Cleveland, W.S., 1993. Visualizing Data. Hobart Press.

Cooke, R.M., 1991. Experts in Uncertainty. Oxford University Press Oxford.

Diez, F.J., 1993. Parameter adjustment in Bayes networks. The generalized noisy OR–gate. In: Uncertainty in Artificial Intelligence. Elsevier, pp. 99–105.

Doran, J.W., Zeiss, M.R., 2000. Soil health and sustainability: Managing the biotic component of soil quality. Appl. Soil Ecol. 15 (1), 3–11.

Friedman, D., Hubbs, M., Tugel, A., Seybold, C., Sucik, M., 2001. Guidelines for soil quality assessment in conservation planning. In: United States Department of Agriculture: Natural Resources Conservation Service.

Gigerenzer, G., 1996. The psychology of good judgment: Frequency formats and simple algorithms. Med. Decis. Mak. 16 (3), 273–280.

Haines-Young, R., 2011. Exploring ecosystem service issues across diverse knowledge domains using Bayesian belief networks. Prog. Phys. Geogr. 35 (5), 681–699.

Hanea, A., McBride, M., Burgman, M., Wintle, B., Fidler, F., Flander, L., Twardy, C., Manning, B., Mascaro, S., 2017. Investigate discuss estimate aggregate for structured expert judgement. Int. J. Forecast. 33 (1), 267–279.

Hassall, K.L., 2019. ACE: Application for conditional probability elicitation. http://dx.doi.org/10.5281/zenodo.2563240, https://github.com/KirstyLHassall/ACE.

Henriksen, H.J., Zorrilla-Miras, P., de la Hera, A., Brugnach, M., 2012. Use of Bayesian belief networks for dealing with ambiguity in integrated groundwater management. Integr. Environ. Assess. Manag. 8 (3), 430–444.

Jenkinson, D., 2005. The Elicitation of Probabilities: A Review of the Statistical Literature. Technical Report, Citeseer.

Johnson, D.L., Ambrose, S.H., Bassett, T.J., Bowen, M.L., Crummey, D.E., Isaacson, J.S., Johnson, D.N., Lamb, P., Saul, M., Winter-Nelson, A.E., 1997. Meanings of environmental terms. J. Environ. Qual. 26 (3), 581–589.

Karlen, D., Mausbach, M.J., Doran, J., Cline, R., Harris, R., Schuman, G., 1997. Soil quality: A concept, definition, and framework for evaluation (a guest editorial). Soil Sci. Am. J. 61 (1), 4–10.

Korb, K.B., Nicholson, A.E., 2004. Bayesian Artificial Intelligence. CRC Press.

Kuhnert, P.M., Martin, T.G., Griffiths, S.P., 2010. A guide to eliciting and using expert knowledge in Bayesian ecological models. Ecol. Lett. 13 (7), 900–914.

Lauritzen, S.L., Spiegelhalter, D.J., 1988. Local computations with probabilities on graphical structures and their application to expert systems. J. R. Stat. Soc. Ser. B Stat. Methodol. 157–224.

Marcot, B.G., 2017. Common quandaries and their practical solutions in Bayesian network modeling. Ecol. Model. 358, 1–9.

Oakley, J., O'Hagan, A., 2016. SHELF: Tools to support the sheffield elicitation framework. http://tonyohagan.co.uk/shelf.

Pearl, J., 1986. Fusion, propagation, and structuring in belief networks. Artif. Intell. 29 (3), 241–288.

Pearl, J., 2014. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Elsevier.

Pollino, C.A., Woodberry, O., Nicholson, A., Korb, K., Hart, B.T., 2007. Parameterisation and evaluation of a Bayesian network for use in an ecological risk assessment. Environ. Model. Softw. 22 (8), 1140–1152.

Price, P.C., 1998. Effects of a relative-frequency elicitation question on likelihood judgment accuracy: The case of external correspondence. Org. Behav. Hum. Decis. Process. 76 (3), 277–297.

Spaccasassi, C., Deleris, L., 2011. A web-based tool for expert elicitation in distributed teams. In: Proceedings of the 8th Bayesian Modelling Applications Workshop of UAI, vol. 2011. Citeseer.

Taalab, K., Corstanje, R., Mayr, T., Whelan, M., Creamer, R., 2015. The application of expert knowledge in Bayesian networks to predict soil bulk density at the landscape scale. Eur. J. Soil Sci. 66 (5), 930–941.

Tattari, S., Schultz, T., Kuussaari, M., 2003. Use of belief network modelling to assess the impact of buffer zones on water protection and biodiversity. Agric. Ecosys. Environ. 96 (1–3), 119–132.

Tóth, G., Stolbovoy, V., Montanarella, L., 2007. Soil quality and sustainability evaluation-an integrated approach to support soil-related policies of the European Union. In: Institute for Environment and Sustainability. Luxembourg, European Communities.

Troldborg, M., Aalders, I., Towers, W., Hallett, P.D., McKenzie, B.M., Bengough, A.G., Lilly, A., Ball, B.C., Hough, R.L., 2013. Application of Bayesian belief networks to quantify and map areas at risk to soil threats: Using soil compaction as an example. Soil Tillage Res. 132, 56–68.

Tversky, A., Kahneman, D., 1973. Availability: A heuristic for judging frequency and probability. Cogn. Psychol. 5 (2), 207–232.

Van Allen, T., Singh, A., Greiner, R., Hooper, P., 2008. Quantifying the uncertainty of a belief net response: Bayesian error-bars for belief net inference. Artificial Intelligence 172 (4–5), 483–513.

Werner, C., Bedford, T., Cooke, R.M., Hanea, A.M., Morales-Napoles, O., 2017. Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions. European J. Oper. Res. 258 (3), 801–819.

Wienhold, B.J., Andrews, S., Karlen, D., 2004. Soil quality: A review of the science and experiences in the USA. Environ. Geochem. Health 26 (2), 89–95.

Wisse, B., van Gosliga, S.P., van Elst, N.P., Barros, A.I., 2008. Relieving the elicitation burden of Bayesian belief networks. In: BMA.

Zapata-Vázquez, R.E., O'Hagan, A., Soares Bastos, L., 2014. Eliciting expert judgements about a set of proportions. J. Appl. Stat. 41 (9), 1919–1933.

Facilitating the elicitation of beliefs for use in Bayesian Belief modelling

Hassall, Kirsty L.