Appendix I

Basic statistical background of CPTs and properties of Bns

Basic statistical background

The concept used for the treatment of certainty in Bayesian networks is that of conditional probability. A conditional probability statement is of the following type:

…… if the variable B is in state b_1, then from either evidence or experience, we know that as a result, the probability of the variable A being in state a_1, is x. The notation for this statement is:

$$P(a_1|b_1) = x$$

The expression P(A|B) denotes a CPT containing numbers $P(a_i|b_j)$. Using Table A1.1 as an example, i = 1 to 3 and j = 1 to 3; in other words variables A and B both have 3 possible states. It should be noted that $P(a_1|b_1) = x$ means that whenever B is in state b_1, the probability of A being in state a_1 is x, *provided that everything else that is known is irrelevant for A.* This is important to remember, since other factors may have a significant affect on variable A. For example, in the case of Table A1.1 the variable 'Reservoir Storage' may also be affected by other factors such as 'Water Abstraction'. If this is the case, then this factor should also be built into the CPT. For our example, however, only the 2 variables 'Reservoir Storage' and 'River Flow' are considered.

		River Flow (Variable B)		
		Good (b_1)	Acceptable (b_2)	Bad (b_3)
Reservoir Storage (Variable A)	Good (a_1)	*0.9*	*0.6*	*0*
	Medium (a_2)	*0.1*	*0.3*	*0.1*
	Bad (a_3)	*0*	*0.1*	*0.9*

Table A1.1 Example CPT showing P(A|B); taken from Figure A1.1. Note that the sum of each column is 1

Table A1.1 is a CPT taken from the network shown in Figure A1.1 (Figure 2.2 in the guidelines). It shows the probability of A being in any particular state (a_1, a_2, a_3) given a state of B (b_1, b_2, b_3); it can be written as P(A|B). In the example, if 'River Flow' is in state b_2 (acceptable), the probability of 'Reservoir Storage' being in state a_1 (good) is 0.6 (60%), a_2 (medium) 0.3, and a_3 (bad) 0.1. Note that each *column* must add up to 1. In Table A1.1 the probabilities given, are based on the fact that the variable B has a 100% probability of being in state b_1, b_2 or b_3. But in reality we are unlikely to be certain of the state of river flow; there will always be some uncertainty. For instance, in the case shown in Figure A1.2 (Figure 2.3 in the guidelines) the variable 'River Flow' is not entirely in one state, but has the following probability distribution; 'good' 0.8 (80%), 'medium' 0.15 (15%), and 'bad' 0.05 (5%).

This probability distribution of B, written as P(B), together with the values given in the 'Reservoir Storage' CPT (Table A1.1), can be used to calculate the resulting probability distribution for reservoir storage, P(A). To obtain this distribution Bns use the *fundamental rule*, which can be written as;

$$P(A|B)P(B) = P(A,B)$$

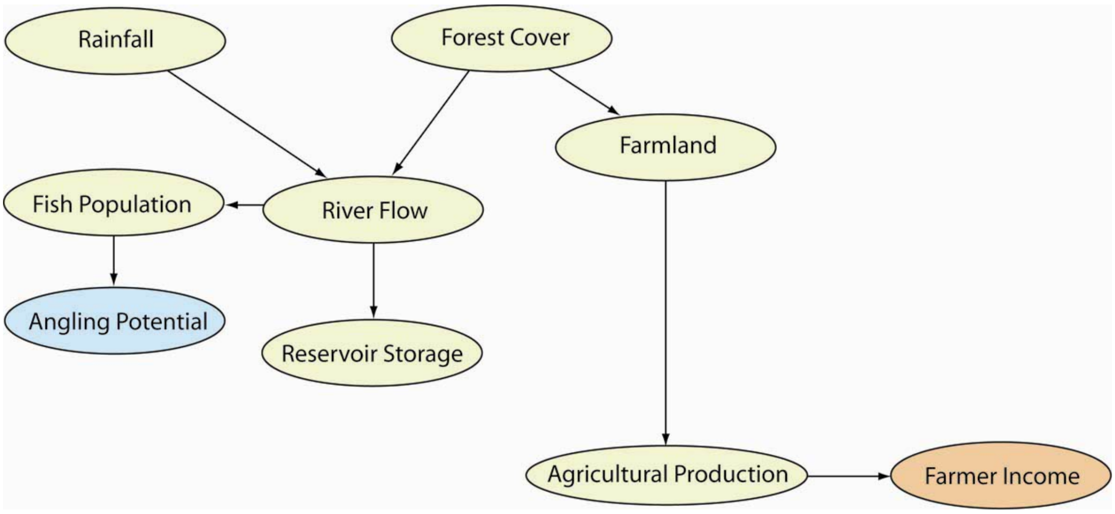


Figure A1.1 A simple Bayesian network

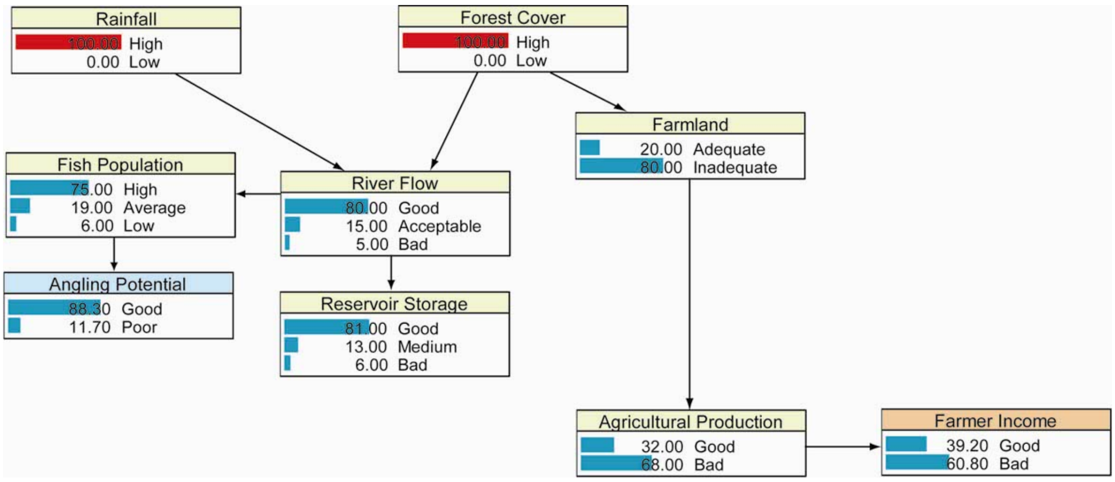


Figure A1.2 Compiled version of the simple Bayesian network shown in Figure A1.1

Here the term P(A,B) is an expression of the *joint* probability for the variables A and B. It consists of a table of all possible configurations (e.g. $a_{1\text{-}3}$, $b_{1\text{-}3}$ in Table A1.1). To construct the table P(A,B) it is necessary to know P(A|B), which we have from the CPT (Table A1.1), and P(B) the probability of B, given by the 'River Flow' node. Using the probability values for B (b_1=0.8; b_2=0.15 and b_3=.05), and the values in the CPT (Table A1.1) we can use the fundamental equation to calculate (P,A). Thus for the configuration a_1,b_1 we can write:

$$P(a_1|b_1)\ P(b_1) = P(a_1,b_1) = 0.9 \times 0.8 = 0.72$$

Values for all the other configurations can be worked out in the same way to produce the full P(A,B) table (Table A1.2). Note that in this table the sum of *all* the entries is 1. This Table gives the joint probability distribution for A and B, but what we really need is the probability distribution for variable A, Reservoir Storage'.

		River Flow (Variable B)		
		Good (b_1)	Acceptable (b_2)	Bad (b_3)
Reservoir Storage (Variable A)	Good (a_1)	0.72	0.09	0
	Medium (a_2)	0.08	0.045	0.005
	Bad (a_3)	0	0.15	0.045

Table A1.2 The *joint* probability distribution P(A,B). Note that the sum of *all* entries is 1

This distribution is derived using a calculation called *marginalisation* whereby the variable is B is marginalised out of P(A,B), resulting in P(A). From Table A1.2 it is clear that for any particular state of A (a_1, a_2 or a_3) there are 3 possible states of B (b_1, b_2 and b_3). Thus for each state of A there are 3 mutually exclusive conditions (a_1, b_1) (a_1, b_2) and (a_1, b_3). From this it follows that:

$$P(a_1) = \sum_{j=1}^{3} P(a_1, b_{1-3})$$

By marginalising B out of Table A1.2 we get:

$P(a_1) = 0.72 + 0.09 + 0 = \underline{0.81}$ (81%)
$P(a_2) = 0.08 + 0.045 + 0.005 = \underline{0.13}$ (13%)
$P(a_3) = 0 + 0.015 + 0.045 = \underline{0.06}$ (6%)

Examination of the compiled network in Figure A1.2 shows that these are the probabilities that appear in the 'Reservoir Storage' window for P(A).

Finally, it should be noted that the fundamental rule gives rise to the well known *Bayes' rule* in the following way:

$P(A|B)P(B) = P(A,B)$ ……………………….. *fundamental rule*

Thus it follows: $P(A|B)P(B) = P(B|A)P(A)$

Then: $P(B \mid A) = \dfrac{P(A \mid B)P(B)}{P(A)}$ ……………………….. *Bayes' rule*

Bayes' rule can be used to obtain the Table P(B|A), which is the CPT showing the likely state of the river flow given the reservoir storage; in other words the river flow is conditional upon reservoir storage, the reverse of situation shown in Table A1.1. Applying Bayes' rule to Table A1.1 gives Table A1.3. Thus in the case of $P(b_1|a_1)$:

$P(a_1|b_1)\ P(b_1)\ /\ P(a_1) = P(b_1|a_1)$

$0.9 \quad x \quad 0.8 \quad / \quad 0.81 = 0.89$

The value $P(a_1|b_1) = 0.9$ is from the CPT in Table A1.1; $P(b_1) = 0.8$, the probability of river flow is given by the probability distribution of the 'River Flow' node shown in Figure A1.2, and $P(a_1) = 0.81$ is the probability given by the 'Reservoir Storage' node in Figure A1.2.

		Reservoir storage (Variable A)		
		Good (a_1)	Medium (a_2)	Bad (a_3)
River Flow (Variable B)	Good (b_1)	0.89	0.615	0
	Acceptable (b_2)	0.11	0.346	0.25
	Bad (b_3)	0	0.039	0.75

Table A1.3 CPT of P(B|A); obtained by applying Bayes' rule to the CPT of P(A|B) in Table A1.1. Note each column adds up to 1

Calculation of these equations is performed automatically by the Bayesian network software, so you will be relieved to know that it is not necessary to solve any of them manually. Nonetheless, although detailed knowledge of the statistical techniques is not essential to run networks, it is clearly preferable to have at least some understanding of the principles upon which the technique is based.

Some basic properties of networks

Direction and feedback loops

Although Bns are flexible and can be used to represent even the most complex systems, there are two basic properties that need to be observed during construction. Firstly they must be *directed*; that is they must act in one direction only. A connection between nodes A and B must be in the direction A → B, or B → A, but not in both directions at once. This is a basic property of networks that cannot be violated.

Secondly the network must also be *acyclic*. This simply means that links must not be allowed to form a closed loop or feedback cycle, because no efficient calculus has yet been developed to model this type of configuration. All networks must, therefore, be both directed and acyclic.

Conditional Independence and d-separation

The links between variables show the direction of cause and effect, but the way in which evidence is transmitted depends on the configuration of the connections, and whether or not evidence has been entered into key nodes. The type of configuration and the evidence available determines whether variables are dependent or independent, and it is important to be able to recognize the different types when they occur. When two variables are independent (i.e. the evidence about one does not affect our belief about the state of the other) they are said to be dependent – separated, or *d-separated* for short. When they are dependent (i.e. our knowledge of the state of one variable will affect our belief about the state of the other), they are described as dependent – connected, or *d-connected*.

In the following paragraphs we look at 3 types of connection - diverging, serial and converging connections - and examine the way in which evidence is transmitted through them.

Diverging connection

The situation shown in Figure A1.3 is a *diverging connection*. Influence can pass between the two child nodes, unless the state of the parent variable is known. This means that once we know the state of the parent (C in Figure A1.3) then any further evidence about B cannot change our belief about A and vice versa.

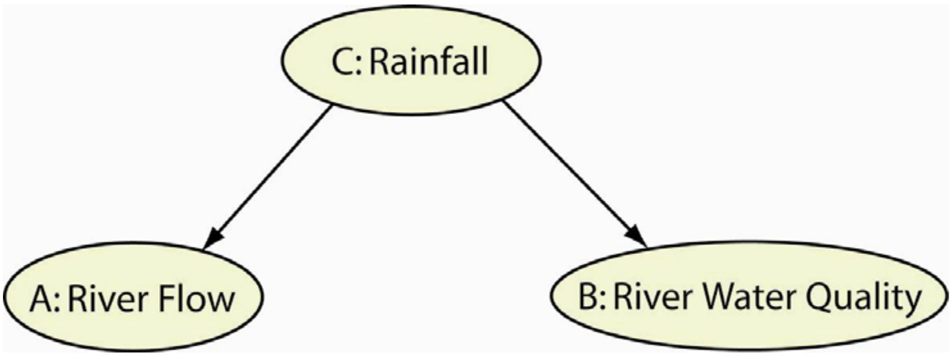


Figure A1.3 Diverging connection

In the example (Figure A1.3) the variables ‘River Flow’ (A) and ‘River Water Quality’ (B), are both dependent on ‘Rainfall’ (C). When rainfall is high, we know that river flow is also likely to be high, and water quality is probably going to be good because of the dilution effect of direct run-off. But suppose nothing is known about the state of rainfall, then evidence about either flow or quality will have an influence on the other. For example, if the quality is poor this might lead us to believe that it is more likely that flow is low and vice-versa; in other words A and B are *dependant* or d-connected.

However, if we have hard evidence that rainfall is ‘high’, then any evidence about water quality that we might have will not change our belief about river flow, which will be overwritten by our knowledge of rainfall. In this case ‘River Flow’ and ‘River Water Quality’ are said to be *conditionally independent* or *d-separated,* given the state of ‘Rainfall’.

Thus evidence may be transmitted through a diverging connection, *unless the connecting (parent) variable is instantiated*

Serial connection

There may be a situation where variable A connects to B which in turn connects to C, a so called serial connection. A will have an impact on B, which in turn will impact on C. Clearly evidence about A will influence the certainty of B and in turn the certainty of C. In the same way evidence about C will influence our belief about B and A. However, if the state of B is known, then the route between A and C is blocked, in which case A and C are d-separated.

An example is given in Figure A1.4 where the variable ‘Water Supply’ (A) has an impact on ‘Agricultural Production’ (B) which in turn has an impact on ‘Farmers Income’ (C). Suppose all the variables have states ‘high’ and ‘low’, and we have some evidence that ‘Water Supply’ is ‘low’. Then clearly this increases our belief that both ‘Agricultural Production’ and ‘Farmers Income’ will also be low. Evidence about A is thus transmitted through B to C. However, suppose that ‘Agricultural Production’ (B) is known to be in the ‘high’ state. In this case any information about ‘Water Supply’ (A) cannot be transmitted to the variable ‘Farmers Income’ (C) because it will be overwritten and replaced by the certain knowledge we have of B.

Thus in a serial connection evidence cannot be transmitted through the connecting variable if the state of the connecting variable is known

Figure A1.4 Serial connection

Converging connection

In the case where the links between variables A and B converge to C, then if nothing is known about C, the parents A and B are independent. But if the state of C is known, or even if we only have some vague indication about the state of C, then it is possible our beliefs about the state of A given B will be changed. An example of this type of converging connection is shown in Figure A1.5.

In this example, if the state of 'Recharge' is known, but there is no information about 'Water Resources', then nothing can be deduced about the state of 'Water Abstraction'; the variables A and B, are thus independent. If however, there is some evidence about the state of 'Water Resources', then from knowledge of 'Recharge' some inference about 'Water Abstraction' can be made. For example, if we knew that 'Recharge' was 'high' but 'Water Resources' was 'low', the implication is that 'Water Abstraction' is more likely to be in a 'high' rather than a 'low' state (to account for the low water resource). In this case A and B are dependent. Likewise if the state of D, 'Agricultural Production', a descendent of C, was known to be low, the same deduction would apply. It follows that in a converging connection evidence can only be transmitted between the parents A and B when the converging variable C, or one of its descendents, has received some evidence.

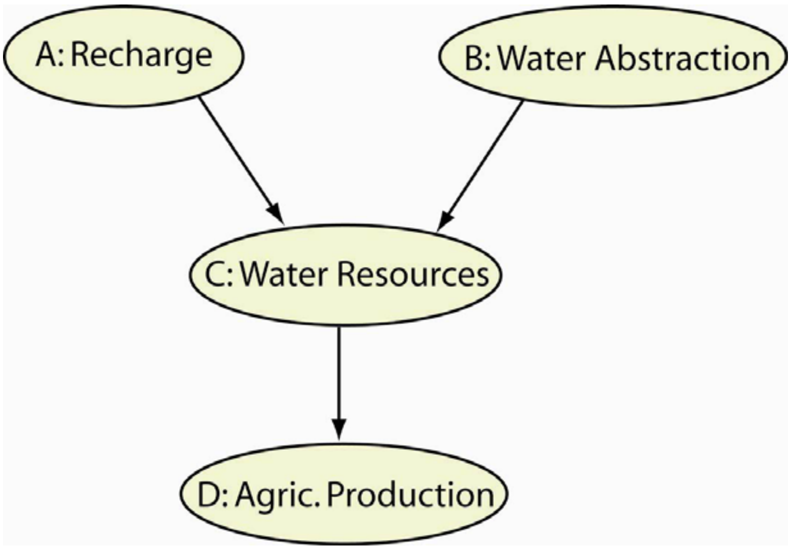


Figure A1.5 Converging connection

D-separation and d-connection: summary

In summary, two nodes A and B in a Bn are *d-separated* (independent) if, for all paths between A and B, there is an intermediate node C for which either:

1. the connection is serial or diverging and the state of C is known for certain; or

2. the connection is diverging and neither C (nor any of its descendants) have received any evidence.

If none of these conditions apply the nodes A and B are *d-connected* (dependant).

Appendix II

Completion of Conditional Probability Tables using large data sets

When you have a large data set the task of completing the resulting CPT for each case is daunting, particularly if there are a large number of states per variable.

Fortunately, commercial Bn packages, such as HUGIN, provide an automated process to calculate the probability for each case, based on the data available, and enters these into the resulting CPT. The procedure in HUGIN is known as EM (Expectation – Maximisation) learning and can best be illustrated by describing an example.

Example of EM learning

This example is a network that predicts the change in domestic water demand in response to changing weather conditions. In the summer months in the UK, Water Companies are aware that domestic demand is dependent to a large extent on (a) the temperature and (b) the rainfall. Water consumption will tend to increase during warm dry periods, but be reduced when the weather turns colder and wetter. The network provides a prediction of the likely consumption given different combinations of temperature and rainfall based on a 7 year data set of rainfall, temperature and water demand.

The network, shown in Figure A2.1, has 3 variables:

1. Average Maximum Daily Temperature in ^{o} C (AvMaxTdegC)
2. Weekly Rainfall in MM (WeeklyRainMM)
3. Domestic Water Consumption in ML day^{-1} (MLDay)

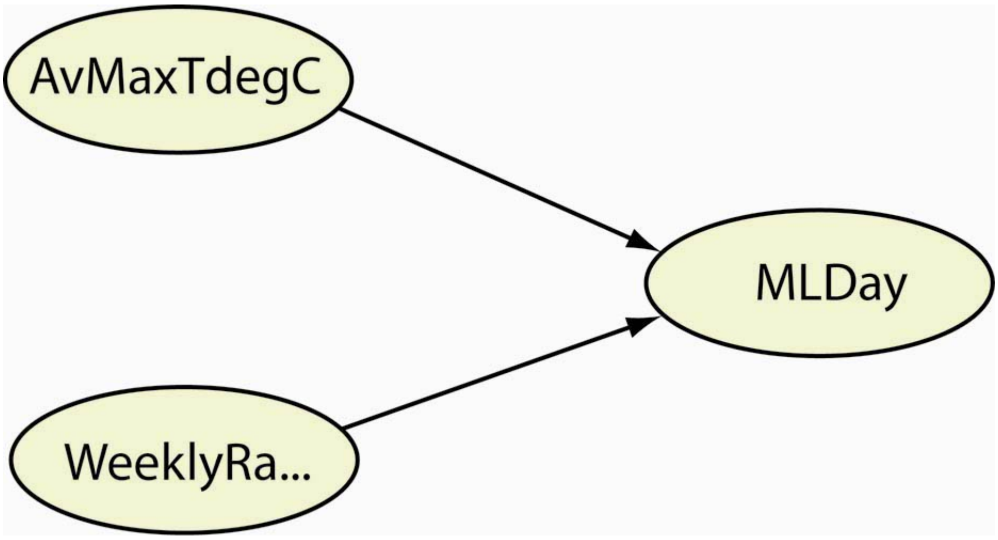


Figure A2.1 Domestic consumption network

The available data set has weekly rainfall, temperature and demand values covering a period of about 7 years. However, not all of this data is useful. In the winter months from October to March, the weather tends to be cool and wet and does not influence demand. For the construction of the CPT table, therefore, only data for the period from May to September is used. After making this adjustment the data set comprised 182 sets of information. The first 18 records are shown in Figure A2.2. The first column shows the average maximum weekly temperature, the second the weekly rainfall total, and the third the average weekly water consumption in mega-litres per day.

AvMaxTdegC	WeeklyRainMM	$MLday^{-1}$
10.21	41.1	211.30
10.22	22.7	212.16
13.92	0	220.65
12.28	9.5	220.47
13.56	0	227.66
15.03	0.6	229.80
14.20	4	220.86
14.36	25.9	225.53
15.54	9.1	213.13
21.67	0.6	243.43
16.47	33	218.07
16.91	0.4	230.79
20.64	0	262.21
20.62	0	273.82
16.80	20.8	225.12
18.00	13.5	225.95
18.05	16.6	222.25
18.34	1.2	230.86

Figure A2.2 Part of the consumption network data set

Once the structure of the network is defined and the relevant data obtained, the next step is to specify the states for each variable. In this instance the states are expressed as a series of intervals as shown in Figure A2.3.

Variable states

AvMaxTdegC	**WeeklyRainMM**	**$MLday^{-1}$**
0-15	0-5	210-220
15-20	5-10	220-225
20-30	10-15	225-230
	15-20	230-235
	20-30	235-240
	30-100	240-250
		250-300

Figure A2.3 The states for each variable; this gives a CPT with 126 cases (3*6*7)

With this combination of states, the CPT for the variable 'MLDay' requires the input of probabilities for 126 cases, clearly something that is impractical to do manually. Instead, the learning routine in the Hugin package can be used to generate and complete the table automatically using the data available. The procedure is as follows:

Step 1: The first step is to assemble the data in a tabular format that can be read by the software (e.g. 'txt' or 'csv'), similar to that shown in Figure A2.2.

Step 2: Initiate the learning wizard in Hugin. The first screen requests the name of the file containing the data.

Step 3: The next screen (Figure A2.4) is for data pre-processing. Here you are given the opportunity to select the variables you wish to include in the data set; in this case all three are selected. At this stage it is necessary to define the ranges of the states for each variable. This is done by highlighting the discretise option and entering the intervals for each variable in turn. In Figure A2.4 the intervals for the 'AvMaxTdegC' variable are shown.

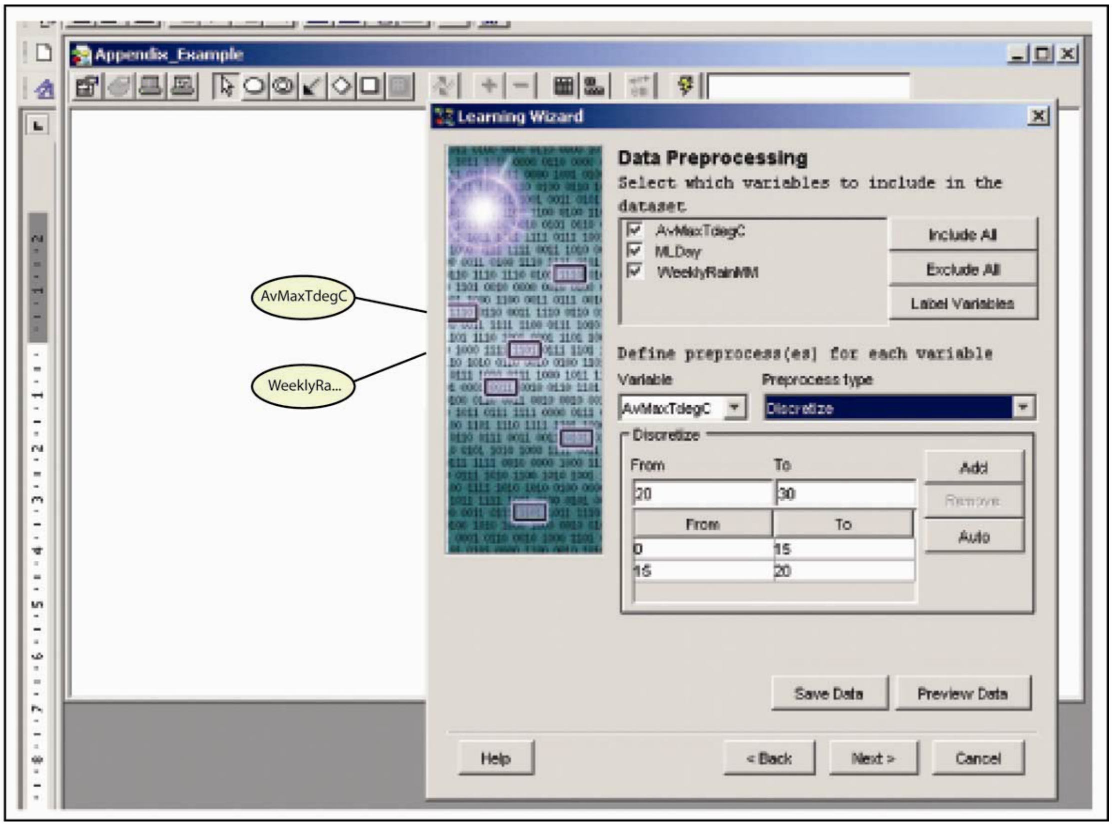


Figure A2.4 Pre-processing screen of the EM learning wizard in Hugin

Step 4: Once the states have been entered it is possible to view the resulting table, which now lists all the possible combinations of cases. Part of this file is shown in Figure A2.5.

Step 5: This step gives you the opportunity to specify the links between each variable. In this case we are sure about the links and have indicated where they should be placed. However, in some cases where you are not certain how the data links together, the software will suggest the strongest relationships based on the data provided.

Step 6: Here you are allowed to select the algorithm to be used to learn the structure of the data. The usual choice is the Necessary Path Condition (NPC) procedure; the level of significance (the probability of rejecting a true independence hypothesis) is normally set at .05 (Figure A2.6).

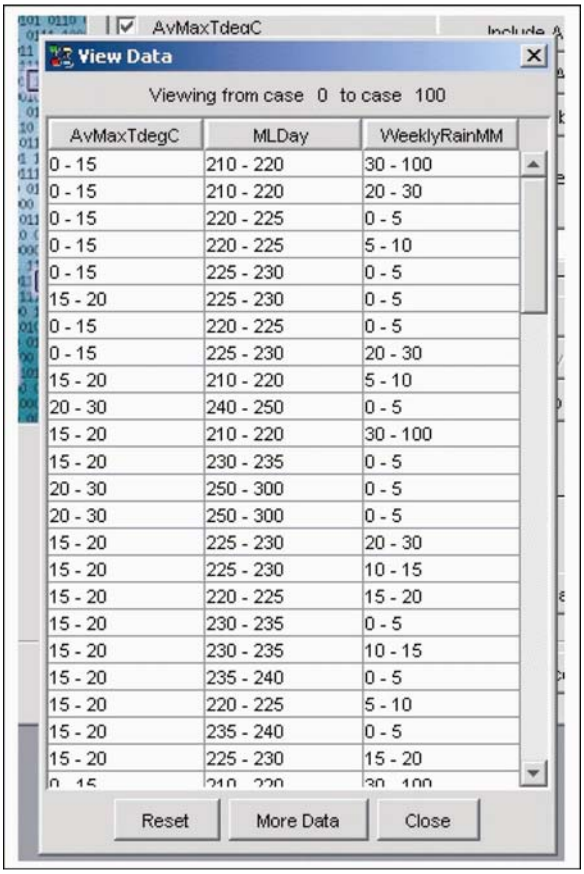


Figure A2.5 Some of the cases defined by the EM learning Wizard

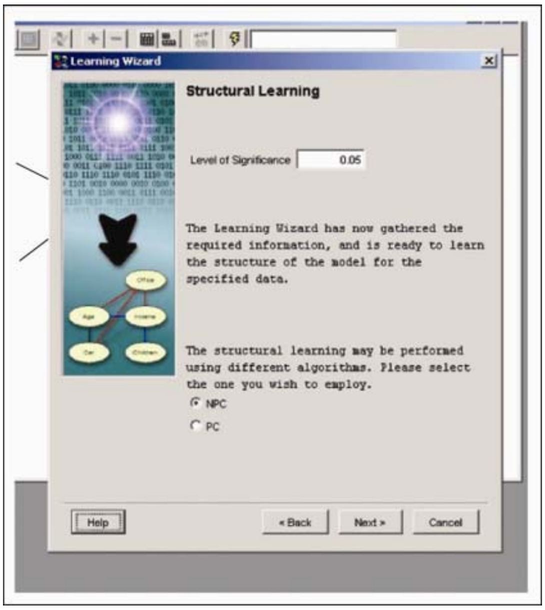


Figure A2.6 Selection of algorithm and significance level

Step 7: The next screen uses a slider to indicate the relative strength of the links.

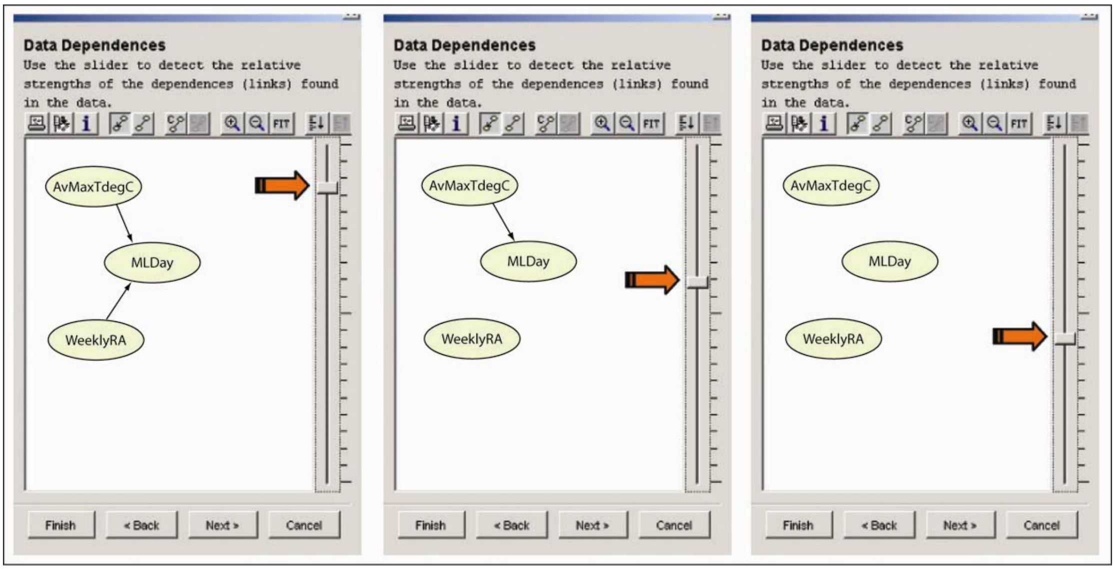


Figure A2.7 Moving the slider on the right downward indicates the relative strength of the links. In this case the link between Temperature and Demand is stronger than between Rainfall and Demand

Step 8: Next, the software provides the opportunity to enter any prior knowledge. In this case we have no idea what the probabilities might be so a ‘1’ is entered for each case to indicate no prior knowledge (Figure A2.8).

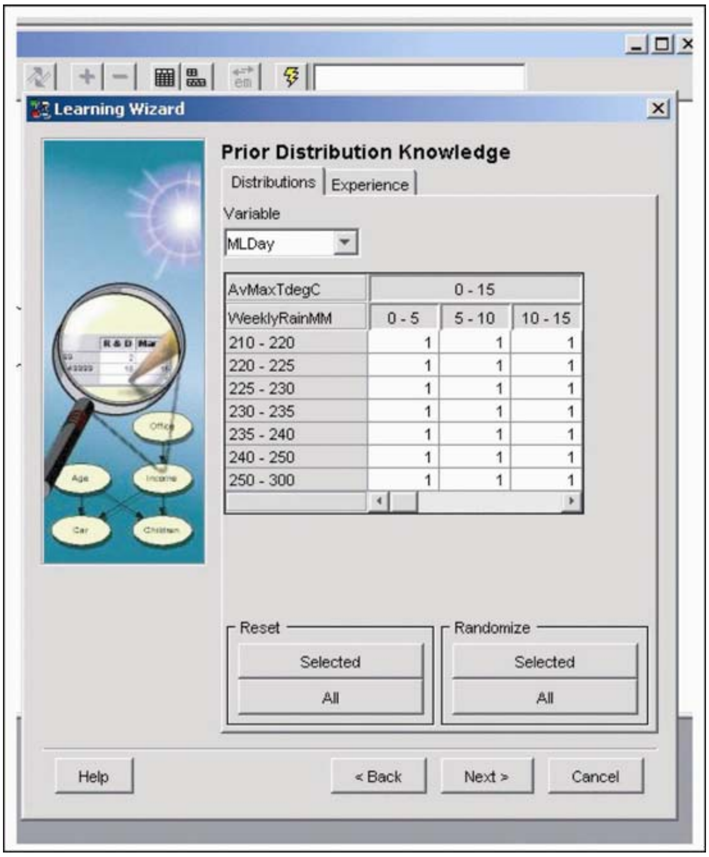


Figure A2.8 CPT for the variable ‘MLDay’ indicating no prior knowledge

Step 9: The final step before calculating the CPT is to specify the number of iterations required for the procedure. Typically a zero is entered here to indicate no upper limit to the number of iterations. The convergence threshold is also required. This must be above zero but is usually taken to be 1x10^{-4} (Figure A2.9). For more details of these values and how they are derived the reader is referred to the help menus of the software.

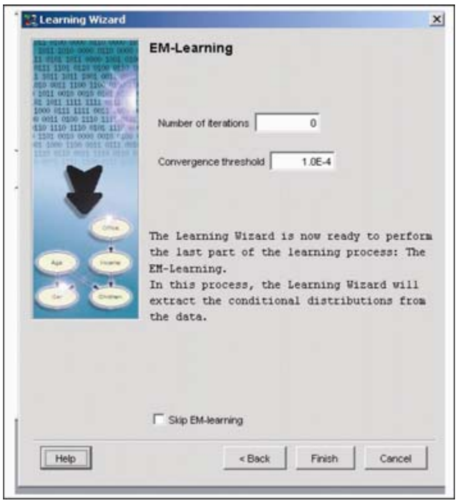
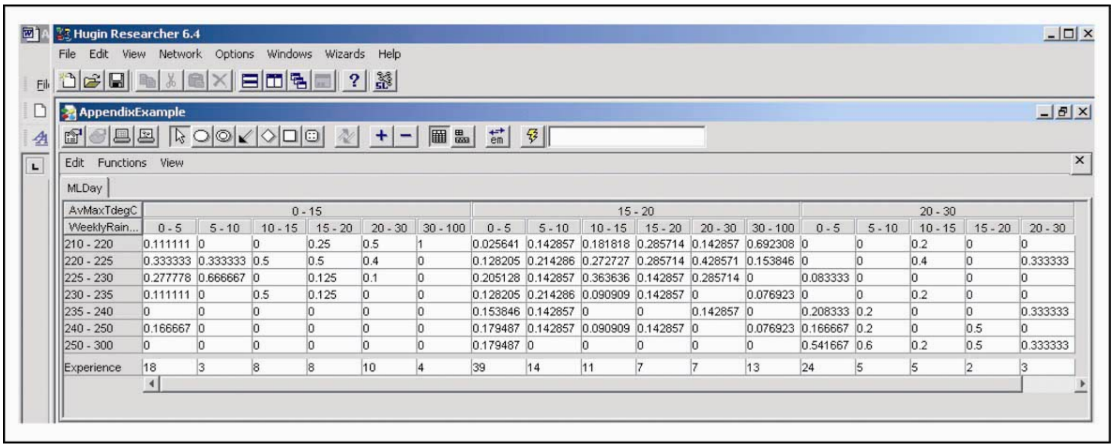


Figure A2.9 Selection of number of iterations and convergence threshold

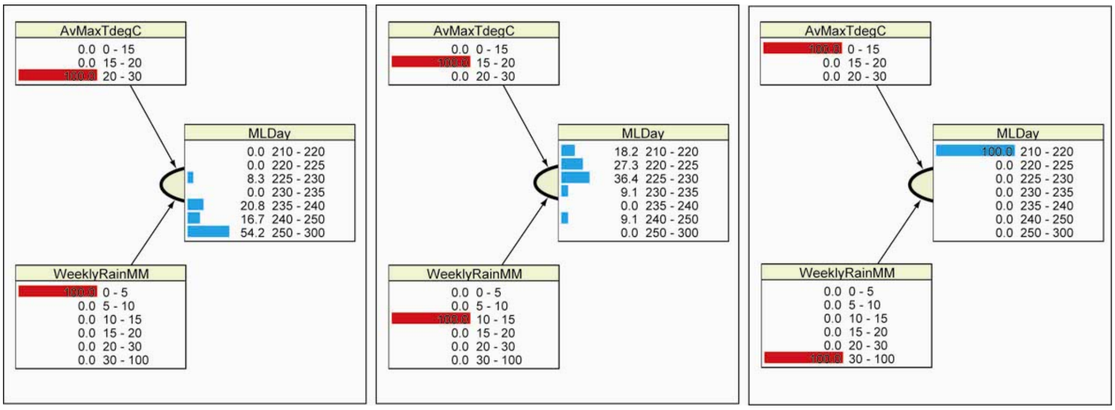
Step 10: The final stage is the generation of the CPTs; the CPT for the variable MLDay is shown in Figure A2.10.



AvMaxTdegC	0 - 15						15 - 20						20 - 30				
WeeklyRain...	0 - 5	5 - 10	10 - 15	15 - 20	20 - 30	30 - 100	0 - 5	5 - 10	10 - 15	15 - 20	20 - 30	30 - 100	0 - 5	5 - 10	10 - 15	15 - 20	20 - 30
210 - 220	0.111111	0	0	0.25	0.5	1	0.025641	0.142857	0.181818	0.285714	0.142857	0.692308	0	0	0.2	0	0
220 - 225	0.333333	0.333333	0.5	0.5	0.4	0	0.128205	0.214286	0.272727	0.285714	0.428571	0.153846	0	0	0.4	0	0.333333
225 - 230	0.277778	0.666667	0	0.125	0.1	0	0.205128	0.142857	0.363636	0.142857	0.285714	0	0.083333	0	0	0	0
230 - 235	0.111111	0	0.5	0.125	0	0	0.128205	0.214286	0.090909	0.142857	0	0.076923	0	0	0.2	0	0
235 - 240	0	0	0	0	0	0	0.153846	0.142857	0	0	0.142857	0	0.208333	0.2	0	0	0.333333
240 - 250	0.166667	0	0	0	0	0	0.179487	0.142857	0.090909	0.142857	0	0.076923	0.166667	0.2	0	0.5	0
250 - 300	0	0	0	0	0	0	0.179487	0	0	0	0	0	0.541667	0.6	0.2	0.5	0.333333
Experience	18	3	8	8	10	4	39	14	11	7	7	13	24	5	5	2	3

Figure A2.10 The final CPT for the MLDay variable. Note the number of 'experiences', or the number of values used to calculate each combination of cases, is displayed in the bottom line. The more experiences, the more reliable will be the probabilities, and the more confidence you can have in the result.

Finally, Figure A2.11 shows the results of the exercise. These examples show that in case (a) with average maximum weekly temperatures between 20-30 degrees and zero rainfall, there is a 54% chance that consumption will be between 250-300 ML day^{-1}. At the other extreme, case (c) indicates that consumption is almost certain to be reduced to between 210-220 ML day^{-1} when conditions are wet and cold. In case (b) with moderate temperatures between 15-20 degrees and a weekly rainfall of 10-15 mm, consumption is most likely to be between 225-230 ML day^{-1}.



(a) Warm and dry (b) Intermediate (c) Cold and wet

Figure A2.11 Some examples of the output using an automatically generated CPT

Appendix III

Completion of Conditional Probability Tables using Stakeholder or Expert opinion

In cases where information is scarce, or when it is difficult to quantify a process, it may be necessary to complete CPTs using input obtained from sources such as stakeholders or experts. There are two types of situation where stakeholder or expert opinion may be sought:

(1) Where local knowledge can substitute for a lack of measurements. This might include historical information about the hydrology, population, incomes etc., which is not recorded, but can be used to help build a CPT. Local residents or organizations may be a source of such information. Stakeholder opinion can also be used to help when trying to quantify social issues, such as the value of an amenity, or acceptability of price changes etc.

(2) Where no data exists, but where expert knowledge of academics, professionals, or others, may be able to provide informed estimates based on theoretical calculations or informed judgement. A good example is provided by the Danish compensation network for which the impacts of different actions on farm economics were quantified using the opinion of a recognized expert in this field.

Both of these information types are subjective and as such open to criticism, but in the absence of data their use is valid, provided the accompanying limitations and uncertainties are acknowledged. A good description of how to complete CPTs using this type of information is given in Cain (2001) .

Cain distinguishes between two situations. The first is where the CPT to be completed is small (less than 10 combinations). In this case the probabilities can be entered directly into the table for all combinations. The second, perhaps more common situation, is where the CPT is larger as a result of either more parents or more states for each of the variables. Remember, a child node with 3 states, having 2 parents each with 3 states, will already have a CPT of 27 combinations. To fill this in directly requires the stakeholder(s) to provide 27 probability estimates, something that is asking a lot of a non-expert; even an expert might find the task a challenge. In this case Cain suggests using an EPT (Elicited Probability Table). In an EPT the probabilities for only a sub-set of the combinations are entered, the rest are calculated using interpolation factors. The procedures are best described using a few simple examples.

Example 1: Direct completion of a small CPT

For this we can use a simplified version of the example in Appendix II. Suppose we want to know how much water is used domestically under different weather conditions, but there is no data. One solution might be to consult an experienced water engineer to provide direct input into our CPT; because we wish to fill in all the combinations, the CPT should be kept as small as possible. Suppose there are 3 variables ‘Temperature’, ‘Rainfall’ and ‘Water Consumption’, with each variable being restricted to two states to minimize the number of combinations. For simplicity the states are described in qualitative terms. The CPT will look like Table A3.1. Because the table is small our engineer will have no problem directly entering the probabilities for each case, based on his experience. Incidentally, it would of course be possible to place threshold values on all the states, so that our definition of ‘Hot’ for instance might be an average temperature above 15°C, and 'High' consumption could be above 350 ML day^{-1}.

		'Water Consumption'	
'Temperature'	**'Rainfall'**	*High*	*Low*
Hot	*Dry*	1	0
Hot	*Wet*	0.7	0.3
Cold	*Dry*	0.4	0.6
Cold	*Wet*	0	1

Table A3.1 Example of direct input into a CPT. In this case the stakeholder or expert enters each of the probabilities into the table directly, based on their knowledge and/or experience

Example 2: Completion of an EPT (Elicited Probability Table)

It is not always possible to construct simple CPTs; sometimes they will inevitably be large and complex. In these cases, where entry needs to be manual, direct completion is not practical; instead it is possible to complete a small subset of all combinations and to complete the rest using interpolation factors. Cain terms such tables Elicited Probability Tables (EPTs). As an example let us take the network shown in Figure A3.1.

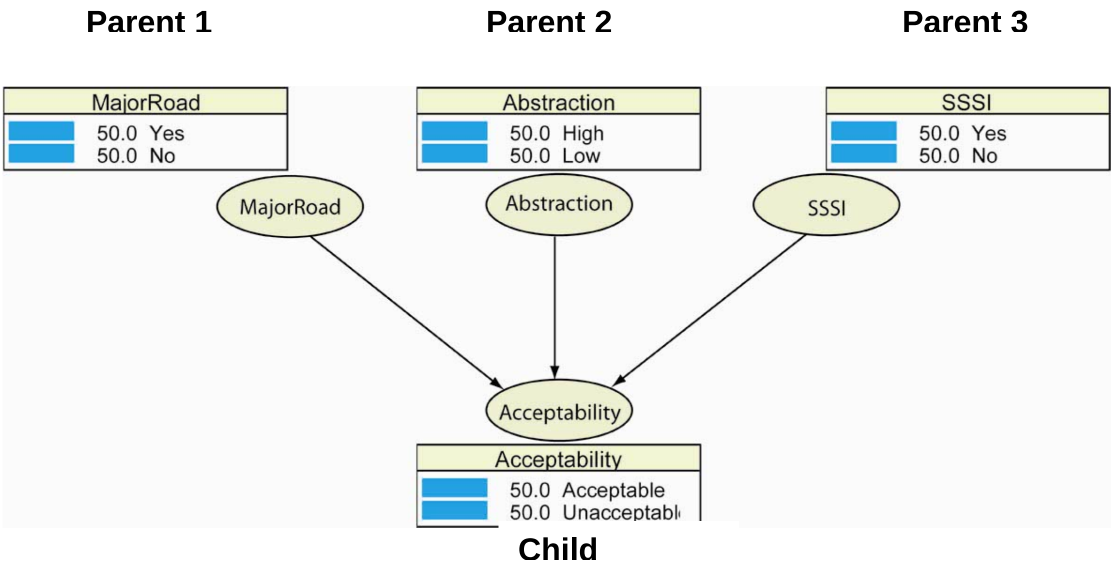


Figure A3.1 Network to illustrate EPT completion

This network investigates the extent to which development of an amenity area (perhaps an area of natural beauty) is acceptable to the stakeholders. In this example the impact of three possible actions are considered; the construction of a major road, abstraction from a river, and the declaration of the site as a SSSI (site of special scientific interest). All the variables are given two states, and it is assumed that the three parents do not affect the degree of change that the others have on the child (i.e. they are NOT modifying parents: see Step 3.3 in the guidelines).

The child has two states, one of which (Acceptable) is more desirable than the other; this can be termed the "success" state. Each parent is also taken to have two discrete states, one of which is more likely to give rise to the success state of the child node. This can be called the "positive" state. For example, for Parent 1 (Road

Construction) the state 'No' is positive from the point of view of improving the site as a public amenity, whereas 'Yes' would be negative, since this might be considered to detract from the area.

When the design of the network is complete, and we are ready to complete the CPTs, a table should be drawn up as shown in Table A3.2. These represent questions that can be put to the stakeholders.

	Parent 1	**Parent 2**	**Parent 3**	**Child State: score out of 10**
Question 1:	Positive state of P1 *(No)*	Positive state of P2 *(Low)*	Positive state of P3 *(Yes)*	Elicited state & score out of ten
Question 2:	Negative state of P1 *(Yes)*	Negative state of P2 *(High)*	Negative state of P3 *(No)*	Elicited state & score out of ten
Question 3:	Negative state of P1 *(Yes)*	Positive state of P2 *(Low)*	Positive state of P3 *(Yes)*	Elicited state & score out of ten
Question 4:	Positive state of P1 *(No)*	Negative state of P2 *(High)*	Positive state of P3 *(Yes)*	Elicited state & score out of ten
Question 5:	Positive state of P1 *(No)*	Positive state of P2 *(Low)*	Negative state of P3 *(No)*	Elicited state & score out of ten

Table A3.2: Elicited probability table (EPT) for network in Figure A3.1

The questions represent only a sub-set of the 16 possible combinations that exist in the CPT (i.e. 2 x 2 x 2 x 2 combinations). They have been chosen to cover the extreme range of probabilities that might be encountered in the CPT, and be the easiest for the stakeholder(s) to answer.

The EPT is formally structured, as follows. The first line (question 1) is such that the parents are all in their positive states. This question asks the stakeholder what state they consider the child would be in, if all the parent nodes were in their positive state. In the second line (question 2) the parents are all in their negative states. For all the other lines, each parent in turn is "switched" from its positive state to its negative state. This is done one parent at a time so that, after the first two, each line only ever has one state that is negative.

Before beginning the interview, you should run through, with the stakeholder, the sort of questions you are going to ask. Explain that you will ask the questions in sequences with each sequence being linked to the same child node. For the first question, ask the stakeholder to imagine that a major road is not being constructed in the area, (positive P1), that there will be no abstractions from the river (positive P2), and that the region will be protected by being declared a SSSI (Positive P3). Ask the stakeholder to give a score out of 10 for the acceptability of this set of actions, with 10/10 indicating total acceptance. The likelihood is that the response to this will be a score of 10/10 for the success state (Acceptable) of the child. The response to the next question (all negative states for the parents) is likely to elicit a 0/10 score. Succeeding questions will provide intermediate scores depending on the importance the stakeholder attaches to each parent.

These questions can be posed to a single representative stakeholder, or preferably to a large number. The responses from each can then be averaged to give the final table.

Let us assume that the stakeholders provide the responses shown in Table A3.3; this leaves us to interpolate the scores for the remaining 11 combinations.

	Parent 1	Parent 2	Parent 3	Child State: score out of 10 that child is in positive state
Question 1:	Positive state of P1 *(No)*	Positive state of P2 *(Low)*	Positive state of P3 *(Yes)*	10
Question 2:	Negative state of P1 *(Yes)*	Negative state of P2 *(High)*	Negative state of P3 *(No)*	0
Question 3:	Negative state of P1 *(Yes)*	Positive state of P2 *(Low)*	Positive state of P3 *(Yes)*	6
Question 4:	Positive state of P1 *(No)*	Negative state of P2 *(High)*	Positive state of P3 *(Yes)*	7
Question 5:	Positive state of P1 *(No)*	Positive state of P2 *(Low)*	Negative state of P3 *(No)*	9

Table A3.3: Stakeholder responses to the EPT

Table A3.4 shows all the combinations of the three parents, including those for which no values have been provided by the stakeholders (i.e. Combination states 4, 6 and 7). These can be calculated using interpolation factors.

State combination	Parent 1	Parent 2	Parent 3	Child State: score out of 10 that child is in positive state
1	Positive *(No)*	Positive *(Low)*	Positive *(Yes)*	10
2	Positive *(No)*	Positive *(Low)*	Negative *(No)*	9
3	Positive *(No)*	Negative *(High)*	Positive *(Yes)*	7
4	**Positive *(No)***	**Negative *(High)***	**Negative *(No)***	**?**
5	Negative *(Yes)*	Positive *(Low)*	Positive *(Yes)*	6
6	**Negative *(Yes)***	**Positive *(Low)***	**Negative *(No)***	**?**
7	**Negative *(Yes)***	**Negative *(High)***	**Positive *(Yes)***	**?**
8	Negative *(Yes)*	Negative *(High)*	Negative *(No)*	0

Table A3.4 The complete CPT showing missing values

Interpolation factors are obtained in the following way:

Interpolation factors are obtained for each 'switch' in the state of a parent from positive to negative. They are calculated in relation to the difference between the highest probability (all parents in the positive state), and the lowest (all parents in the negative state). Using the combination numbers in Table A3.4, this can be expressed as P_1-P_8. When one of the parents is switched from a positive to a negative state, the probability of the child being in the success state is reduced. The interpolation factor simply quantifies this reduction, for each parent, as a proportion of P_1-P_8. Thus:

Interpolation Factor (IF) for Parents

IF3 (Parent 3) = $(P_2-P_8) / (P_1-P_8)$ = (9-0)/(10-0) = 0.9
IF2 (Parent 2) = $(P_3-P_8) / (P_1-P_8)$ = (7-0)/(10-0) = 0.7
IF1 (Parent 1) = $(P_5-P_8) / (P_1-P_8)$ = (6-0)/(10-0) = 0.6

Interpolation factors calculate the way the probability of the state of a child changes when a parent switches from a positive to negative state. Thus, to calculate the child state in Table A3.4 for combination state 4, which has not been given by the stakeholders, all we need to do is to multiply state combination 3, by the interpolation factor associated with parent 3 (i.e. IF3). The only difference between state 3 and state 4 is that parent 3 has switched from positive to negative.

Thus for the three unknown probabilities:

$P_4 = [(P_3-P_8) \times IF_3] + P_8 = [(0.7 - 0) \times 0.9] + 0 = 0.63$
$P_6 = [(P_5-P_8) \times IF_3] + P_8 = [(0.6 - 0) \times 0.9] + 0 = 0.54$
$P_7 = [(P_5-P_8) \times IF_2] + P_8 = [(0.6 - 0) \times 0.7] + 0 = 0.42$

When these interpolation factors are applied, the resulting CPT is shown in Table A3.5

State combination	**Parent 1**	**Parent 2**	**Parent 3**	**Child State: score out of 10 that child is in positive state (Acceptable)**
1	Positive *(No)*	Positive *(Low)*	Positive *(Yes)*	10
2	Positive *(No)*	Positive *(Low)*	Negative *(No)*	9
3	Positive *(No)*	Negative *(High)*	Positive *(Yes)*	7
4	**Positive *(No)***	**Negative *(High)***	**Negative *(No)***	**6.3**
5	Negative *(Yes)*	Positive *(Low)*	Positive *(Yes)*	6
6	**Negative *(Yes)***	**Positive *(Low)***	**Negative *(No)***	**5.4**
7	**Negative *(Yes)***	**Negative *(High)***	**Positive *(Yes)***	**4.2**
8	Negative *(Yes)*	Negative *(High)*	Negative *(No)*	0

Table A3.5 The CPT with the calculated probabilities for the child being in an Acceptable state shown in red. The probabilities are given as a score out of 10.

In Table A3.5 the figures in the right hand column show the probability that the child variable is the positive state (e.g. for state combination 6 the probability of the combination being acceptable to stakeholders is 54%). Table A3.5 only shows the 8 probabilities that indicate the child is in the positive (Acceptable) state. The remaining 8 probabilities, that the child is in the negative (Unacceptable) state, is obtained by simply subtracting the positive values from 10. The final appearance of the CPT in the network, with all 16 combinations completed, is shown in Table A3.6.

Acceptance	**Road (1)**	Yes				No			
	River (2)	High		Low		High		Low	
	SSSI (3)	Yes	No	Yes	No	Yes	No	Yes	No
Yes		**6**	**0**	**4.2**	**5.4**	**7**	**6.3**	**10**	**9**
No		**4**	**10**	**5.8**	**4.6**	**3**	**3.7**	**0**	**1**

Table A3.6 Completed CPT for the variable 'Acceptability'; note all columns add up to 10

In his appendices Cain goes on to describe how to construct EPTs for a range of different circumstances including:

1. When one or more of the parents is a modifying variable
2. When one or more parent is a continuous variable
3. When the child has 3 or more states

Readers are referred to these appendices in Cain's report, which is included on the CD attached to these guidelines.

To facilitate the completion of EPT tables Cain also developed a calculator, which he has kindly made freely available, and which we have also included on the attached CD.

References

Cain, J., 2001. Planning improvements in natural resources management. Centre for Ecology and Hydrology, Wallingford, UK, 124pp pp.