9

Bayesian networks and food security – an introduction

Alfred Stein[#]

Abstract

This paper gives an introduction to Bayesian networks. Networks are defined and put into a Bayesian context. Directed acyclical graphs play a crucial role here. Two simple examples from food security are addressed. Possible uses of Bayesian networks for implementation and further use in decision support are discussed.

Introduction

Bayesian networks are based on the Bayesian paradigm: prior information + evidence gives posterior information. Often such prior information is imprecise, for example, given by a probability distribution. Evidence then occurs in the form of data that are relevant for the information. Such data are then combined with the prior information to give posterior information. Posterior information hence essentially combines prior information with data. As such it is different from frequentist information, which largely ignores prior information, but it also differs from belief, which may ignore evidence.

In food-security issues, the idea of a network of events is in current practice (Van Der Vorst et al. 1998). The food network, either from seed and pesticides to legume or towards consumption meat is an important line of analysis. What happens at one stage within the network has an effect on subsequent steps. Also, an effect observed at one particular stage can have its origin at several steps prior in the network. Using a network approach we may be able to distinguish between such a forward-looking and a backward approach. To quantify expressions and consequences it is important to obtain quantitative expressions for risks and sizes of effects within food networks. A mechanism for propagation can then be formulated by applying Bayes' rule.

The study will focus on an introduction of a Bayesian-network approach towards food-chain issues. Examples of these include measuring the health effects of ingesting food still containing a small dose of a herbicide. Another example concerns the detection of the cause of some food-related hazard to possible sources of such an effect. This paper will apply a Bayesian network to precisely these examples. More complicated examples, also including policy and decision-making issues, can readily be constructed.

Bayesian networks date back to at least the early nineteen nineties. Very readable introductions to the topic is Jensen (1996; 2001), on which much of the current text is based. Other, more advanced texts include Lauritzen (1996), whereas hierarchical models, a typical consequence of the use of Bayesian networks, can be found in Gelman et al. (1995).

[#] Biometris, P.O. Box 100, 7600 AC Wageningen, The Netherlands. E-mail: alfred.stein@wur.nl

The objective of this paper is to give an introduction to Bayesian networks, with a clear eye towards examples from food security. The paper will be constructed as follows. We first define Bayesian networks. Next, we use statistical methods to estimate probabilities within these networks and we describe tools for doing so. Finally, we will discuss Bayesian networks in a few hypothetic examples.

Bayesian networks

As an introduction we will start with giving two simple examples of a Bayesian network. Such simple models can be useful as building blocks for further developing and applying networks in a more general setting.

Two simple examples

Example 1

A constituent applied in manufacturing two food items may or may not exceed a threshold level (*T*) of a contaminating substance (Figure 1). This may influence the quality of both a biologically produced product (*B*) or a conventionally produced product (*C*). In addition, the value of the threshold level *T* may be uncertain. Combination of *T* and the products *B* and *C* forms a (small) network, with causal relations. The probabilities of *B* exceeding the threshold level can be quantified using conditional probabilities, expressed as Pr(*B*|*T*). At some stage a control institute discovers that the conventionally produced product *C* does not exceed the safety level. The question then is whether a sound update of the probability that *B* does not exceed the threshold level can be made.

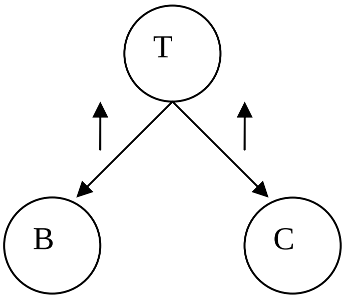

Figure 1. A network of the threshold example. The arrows on the links model the causal impact, the small arrows indicate the direction of the impact on the certainty

Example 2

A case is reported of a child having stomach problems (Figure 2). It is vomiting and is taken to a doctor, who confirms the ingestion of a possibly infectious material. The next question is then what a possible cause could be. Either the child has ingested this material at home, in which case it may be the only one in the vicinity to show the signals, or ingestion may be due to a substance located somewhere outside, in which case more children in the vicinity may show similar signals. This brings us to a simple graph, representing a Bayesian network with four nodes, with *E* representing the substance being ingested in the environment, *H* the substance being ingested at home, *1* the situation that only one child shows symptoms and *M* the situation that many children show symptoms. After some investigations it appears that only one child has stomach problems.

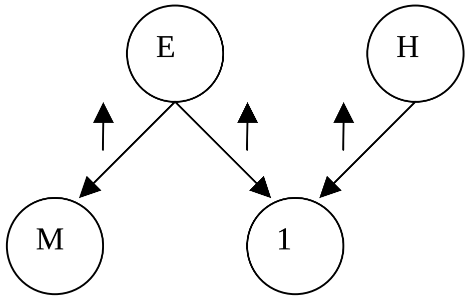

Figure 2. A network model for the one (*1*) vs. many (*M*) example, where illness is caused by either home ingestion (*H*) or environmental ingestion (*E*). Only *E* can cause many illnesses

General theory

We consider a network defined as a collection of events. In this paper these will be denoted by capitals. In addition, we consider causal networks, which consist of a set of variables and a set of directed links between these variables. Such a network is called a directional graph. Often, an event at the end of a link is termed a child (for an event *A* denoted as *ch*(*A*)), and at the source of the link a parent (denoted as *pa*(*A*)). Variables represent events. In a causal network a variable represents a set of possible states of affairs. We will assume that a variable is in exactly one of its states. For the moment we restrict ourselves to a finite number of states, although extensions towards infinitely many states could be made. Each event has associated to it a certainty, which is a real number.

Causal relations have both a start and an ending point, but they also have their strength. This is expressed by attaching numbers to links. To quantify connections in a network we use probabilities. In a network we come to the conclusion that an event A_1 causes with certainty *p* the event A_2. From this the following reasoning is applied: if we know that A_1 has taken place, then A_2 has taken place with certainty *p*. In a causal network, let A_1 be a parent of B_1. It is then natural to let the strength be given by the conditional probability $\Pr(B_1 | A_1)$. However, A_2 may also be a parent of B_1. Then the two conditional probabilities $\Pr(B_1 | A_1)$ and $\Pr(B_1 | A_2)$ alone do not give any clue on how the impacts from A_1 and A_2 interact: they may co-operate, or counteract in various ways, requiring a specification of $\Pr(B_1 | A_1, A_2)$. Also, a second child B_2 may exist, leading to effects from A_1 and/or A_2 on B_2 as well, possibly linked by evidence that has occurred for B_1.

A Bayesian network is only defined for networks that can be visualized as a directed acyclical graph (Figure 3). A directed acyclical graph contains nodes and directed arrows that are such that it is impossible for any state to return into it by following the direction of the arrows. In many applications one of the nodes in a network will receive evidence. This concept implies that a variable is fixed in a specific state. Evidence will hence influence all subsequent events.

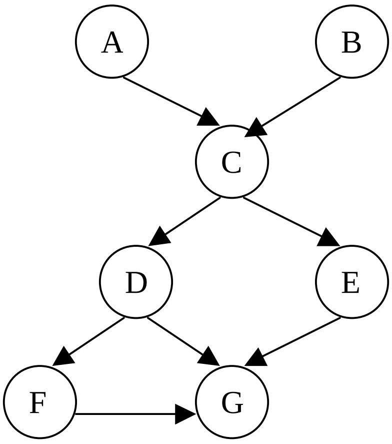

Figure 3. A directed acyclical graph (DAG). The probabilities to specify are Pr(*A*), Pr(*B*), Pr(*C*|*A*,*B*), Pr(*E*|*C*), Pr(*D*|*C*), Pr(*F*|*D*) and Pr(*G*|*D,E,F*).

We now come to the definition of a Bayesian network. A Bayesian network consists of the following:

- a set of *variables* and a set of *directed edges* between variables
- each variable has a finite set of states
- the variables together with the directed edges form a directed acyclical graph (DAG)
- to each variable A with parents $B_1,\dots,B_n$ there is attached a conditional probability table $\Pr(A|B_1,\dots,B_n)$.

In this example the events *A* and *B* are called converging, *A* and *B* are both parents of the child *C*, whereas *C* is a parent of both *D* and *E*, and in addition *C* is diverging into *D* and *E*. The probabilities to specify are Pr(*A*), Pr(*B*), Pr(*C*|*A*,*B*), Pr(*E*|*C*), Pr(*D*|*C*), Pr(*F*|*D*) and Pr(*G*|*D,E,F*). If *A* receives evidence, then this will directly influence all subsequent probabilities. In particular, evidence on *A* can change belief concerning *B* because of their connection through *C*. It will not affect Pr(*C*|*A*,*B*), which is constant (and is part of the domain specification), but it may lead to a different posterior distribution.

To analyse such a graph, standard probability rules apply:

- the fundamental rule for probability calculations:

$$\Pr(A \mid B)\Pr(B) = \Pr(A, B) \tag{1}$$

- Bayes’ rule: $$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A)} \tag{2}$$

- Marginalization: $$\Pr(A) = \sum_i \Pr(A, b_i) \tag{3}$$

- Conditional independence: A and C are independent given B if $$\Pr(A \mid B) = \Pr(A \mid B, C) \tag{4}$$

Of great importance in a causal system is the chain rule. Let *BN* be a Bayesian network over $U = \{A_1,\dots,A_m\}$. Then the joint probability distribution Pr(*U*) is the product of all conditional probabilities specified in *BN*:

$$\Pr(U) = \prod_i \Pr(A_i \mid pa(A_i)) \tag{5}$$

Finally, we need the concept of *d*-separation. Two variables A_1 and A_2 in a causal network are *d*-separated if for all paths between A_1 and A_2 there is an intermediate variable *B* such that either

- the connection is serial or diverging and the state of *B* is known, or
- the connection is converging, and neither *B* nor any of B's descendants have received evidence.

Building models

A Bayesian network consists of variables and directed edges. At the basis of it are the parentless nodes. These nodes detail the prior information that is required and used within subsequent stages of the network. In further organizing a Bayesian model for a decision-support system, the next four steps can be identified:

1. The first thing to have in mind is that its purpose is to give estimates of certainties for events, which are not observable or only observable at an unacceptable cost. Therefore, the primary task in model building is to identify these events, the so-called *hypothesis events*.
2. The hypothesis events have to be organized into a set of variables. A variable incorporates an exhaustive set of mutually exclusive events. That is, for each variable precisely one of the events is true.
3. To come up with a certainty estimate we should provide some *information channels*. Therefore, the task is to identify the types of achievable information that may reveal something about the state of some hypothesis variable. This is done by establishing *information variables*, such that the piece of information corresponds to a statement about the state of an information variable. Typically, the information will be a statement that a particular information variable is in a particular state; but also softer statements are allowed.
4. Finally, we need to consider the causal structure between the variables. At this stage we need not worry about how information is transmitted through the network. The only thing to worry about is which events have a causal impact on other events.

We clearly need three types of variables: hypothesis variables, being variables with a state that is asked for but which are either impossible or too costly to observe, information variables, which can be observed, and mediating variables, which are introduced for a specific purpose. Mediating variables are either introduced to reflect properly the independence properties in the domain, to facilitate the acquisition of conditional probabilities or to reduce the amount of distributions to acquire for the network.

Conditional probabilities need to be acquired as well. For that purpose well-founded probabilities can be applied, such as frequencies or purely subjective estimates. If the amount of probabilities is too large for a reasonable estimation, simplifying assumptions can reduce it.

An important issue as well is the learning process. We can distinguish two processes here: batch learning and adaptation. In batch learning we consider *U* as the set of configurations over a universe of variables and we try to find the best probability distribution according to a relevant criterion (see appendix for more details). As an alternative we may consider adaptation. An example is the continuous updating of prior belief from a stream of cases. Another example is the second-order uncertainty, i.e., uncertainty on the conditional probability table $\Pr(A|\mathrm{pa}(A))$. Consider *A* to be a variable with parents $\mathrm{pa}(A)$. Adaptation consists of using the incoming cases to reduce the second-order uncertainty. Adaptation may take place by either trough-type variable or by fractional updating. More details are given in Jensen (1996).

The two examples revisited

Example 1

For the quantitative modelling we need three probability assessments: Pr(*B*|*T*), Pr(*C*|*T*) and Pr(*T*). The model in Figure 1 reflects that only knowledge of threshold exceedance is relevant for biological or conventional manufacturing of a food product. We should then attach a certainty to *T* based on whatever knowledge may be available. We let Pr(*T*) be equal to 0.7.

Since both ways of manufacturing may lead to exceedance in the final product, we put the probability for conventional manufacturing to 0.8 if the substance was too high in raw material and to 0.1 if the substance has not been used. For the biological manufacturing process these figures are equal to 0.4 and 0.2, respectively (Table 1). To calculate the initial probabilities for *B* and *C*, we first use the fundamental rule (rule 1), resulting into a table for the joint probabilities (Table 2). The probabilities for *B* and *C* are obtained by marginalizing *T* out of this table (rule 3) leading to Pr(*C*) = (0.59,0.41) and Pr(*B*) = (0.34,0.68). The information that the product manufactured by the conventional production process did not exceed critical health levels is now used to update the probability that it was the biological production process. For this, Bayes' rule (rule 2) is used:

$$\begin{aligned}\Pr(T \mid C = y) &= \frac{\Pr(C = y \mid T)\Pr(T)}{\Pr(C = y)} \\ &= \tfrac{1}{0.59}(0.56,0.03) \\ &= (0.95,0.05)\end{aligned}$$

To update Pr*(B*), first we use the fundamental rule (rule 1) to calculate Pr*(B,T*) as is shown Table 3. Finally, Pr*(B*) is calculated by marginalizing *T* out of Pr(*B,T*) (rule 3), with as a result that Pr*(B*) = (0.39,0.61). This is the quantitative effect of the information that it was not the conventional production process.

Table 1. Conditional probabilities for manufacturing processes *C* (conventional) and *B* (biological) given that a threshold *T* is or is not exceeded.

	T = y	*T = n*		*T = y*	*T = n*
C = y	0.8	0.1	*B=y*	0.4	0.2
C = n	0.2	0.9	*B=n*	0.6	0.8
	Pr(*C*|*T*)			Pr(*B*|*T*)	

Table 2. Joint probabilities for manufacturing processes *C* (conventional) and *B* (biological) with exceedance of threshold *T*.

	T = y	*T = n*		*T = y*	*T = n*
C = y	0.56	0.03	*B=y*	0.28	0.06
C = n	0.14	0.27	*B=n*	0.42	0.24
	Pr(*C,T*)			Pr(*B,T*)	

Table 3. Update of the probabilities *Pr(B)* using the fundamental rule (3) upon receiving evidence that *C* is applied

	T = y	*T = n*
B = y	0.38	0.01
B = n	0.57	0.04

As an alternative, we may first calculate Pr(*B*,*T*) and Pr(*C*,*T*), leading to two joint probability tables, both containing the variable *T*. Evidence on *C* then arrives in the form of $\Pr^*(C)$ = (0,1), where $\Pr^*$ denotes updated values following addition and propagation of evidence, i.e. denoting posteriors. Then:

$$\Pr^*(C,T) = \Pr(T \mid C)\Pr^*(C) = \frac{\Pr(C,T)}{\Pr(C)}\Pr^*(C)\,.$$

This means that the joint probability table for *C* and *T* is updated by multiplying by the new distribution and dividing by the old one. Multiplication consists of annihilating all entries with *C = n*. The division by Pr(*C*) only has an effect on entries with *C = y*, therefore the division is by Pr(*C = y*). Next, calculate $\Pr^*(T)$ from $\Pr^*(C,T)$ by marginalization and use $\Pr^*(T)$ to update $\Pr^*(B,T)$:

$$\Pr^*(B,T) = \frac{\Pr(B,T)}{\Pr(T)}\Pr^*(T)\,.$$

Finally, $\Pr^*(B)$ is calculated by marginalizing $\Pr^*(B,T)$.

Example 2

Let the two probabilities for *E* and *H* be equal to Pr(*E*) = (0.15,0.85) and Pr(*H*) = (0.2,0.8). The remaining probabilities are listed in Table 4. Note that because of the particular structure of the problem both univariate and multivariate probabilities occur. First the prior probabilities for *M* and *1* are obtained.

The prior probability Pr(*M*) is obtained by calculating Pr(*M*,*E*) and marginalizing *E* out. The result is Pr(*M=y*,*M=n*) = (0.2775,0.7225). The calculation of Pr(*1*) follows the same scheme, only the product now is Pr(*1*,*E*,*H*) = Pr(*1*|*E*,*H*) Pr(*E*,*H*). Since *E* and *H* are independent, we have Pr(*1*,*E*,*H*) = Pr(*1*|*E*,*H*) Pr(*E*)Pr(*H*). The result is given in Table 5. Marginalizing *E* and *H* out of Pr(*1*,*E*,*H*) by applying rule 3 yields Pr(*1*) = (0.286,0.714). We have now established joint probability tables (Tables 4, 5) for two of the clusters, (*M*,*E*) and (*1*,*E*,*H*), with the variable *E* in common.

Table 4. Initial probabilities for the second example with events *E* that substance occurred in the environment, *H* that substance only occurred at home, *M* that many children show an effect, and *1* that only one child shows an effect

	E = y	*E = n*		*E = y*	*E = n*
M = y	1	0.15	*H = y*	(1,0)	(0.8,0.2)
M = n	0	0.85	*H = n*	(1,0)	(0,1)
	Pr(*M*|*E*)			Pr(*1*|*E*,*H*)	

Table 5. Joint probabilities *Pr(1,E,H)* as a vector of two elements, where the first elements corresponds to {*1* = True} and the second element to {*1* is False}

	E = y	*E = n*
H = y	(0.03,0)	(0.136,0.034)
H = n	(0.12,0)	(0,0.68)

Next evidence emerges, namely that only one child is ill, hence *1 = y*. The evidence *1 = y* is used to update Pr(*1*,*E*,*H*) by annihilating all entries with *1 = n* and dividing by Pr(*1 = y*) = 0.286. Since the result is a probability table with all entries summing to one, there is no need to calculate Pr(*1*). After all entries have been annihilated, we simply normalize the table by dividing by the sum of the remaining entries. The distributions Pr(*E*) and Pr(*H*) are calculated by means of marginalization of Pr(*1*,*E*,*H*), applying rule (3) and dividing each entry by 0.272. This leads to Table 6, from which we obtain Pr(*E = y*) = 0.525 and Pr(*H = y*) = 0.580. Next we use $\Pr^*(E)$ to update Pr(*M*,*E*), yielding the posterior $\Pr^*$(M,E):

$$\Pr^*(M,E) = \Pr(M \mid E)\Pr^*(E) = \Pr(M,E)\frac{\Pr^*(E)}{\Pr(E)},$$

leading to Table 7. Next *M = y* is used to update the distribution for (*M*,*E*) (see Tables 8 and 9). For example, we get $\Pr^{**}(E = y) = 0.88$. We next have to calculate $\Pr^{**}(H)$ = Pr(*H*|*M = y, 1 = y*). The result must reflect the explaining-away effect: since illness is explained by localization of the contaminant within the house, the probability of *H = y* should decrease to its initial value. The calculation follows the same pattern. First, a message on $\Pr^{**}(E)$ is sent from (*M*,*E*) to (*1*,*E*,*H*):

$$\Pr^{**}(1,E,H) = \Pr^*(1,E,H)\frac{\Pr^{**}(E)}{\Pr^*(E)}.$$

By marginalizing we get $\Pr^{**}(H = y) = 0.296$. The reason why the probability for the localization of the contaminant within the house does not drop to the prior probability of 0.2 is that it is possible that contaminated substances were present in more than one house, and not just in the environment.

Table 6. Probability distribution of Pr(*M*,*E*) after receiving information that only one child is infected

	E = y	*E = n*
H = y	(0.105,0)	(0.475,0)
H = n	(0.420,0)	(0,0)

Table 7. Updated probabilities Pr*(*M*,*E*)

	E = y	*E = n*
M = y	0.525	0.0713
M = n	0	0.4037

Table 8. Updated probabilities Pr*(*M*,*E*) using the information that *M* = *y*

	E = *y*	*E* = *n*
M = *y*	0.88	0.12
M = *n*	0	0

Table 9. Updated probabilities $\Pr^{**}(1,H,E)$ after receiving evidence that *M* = *y*.

	E = *y*	*E* = *n*
H = *y*	0.176	0.12
H = *n*	0.704	0

Discussion

A Bayesian network serves as a model for networks occurring in food security, and the relations in the model reflect causal impact between events. The reason for building these computer networks is to use them when taking decisions. The probabilities provided by the network are used to support some kind of decision making. For that we can distinguish forward-looking and backward-looking approaches. Although the examples presented in this paper are very elementary only, they serve as building blocks for these types of decisions.

If decision making is being considered seriously, several aspects of the decision-making process come into view. We need to decide upon a proper loss function as well as on actions that have to be taken. We may make a distinction here into intervening actions, which force a change of state for some variables in the model, and non-intervening actions, whose impact is not a part of the model. Intervening action may also force the direction of causality to be inverted. It is beyond the scope of this paper to extend much beyond this notion. The interested reader is referred to Jensen (1996; 2001) for an introduction and to Lindley (1971) for more theoretical considerations.

A Bayesian network is also useful for illustrating the effects of different scenarios. For that purpose, stochastic simulations are useful. It can then be analysed what the effects are *if* some actions are taken. For example, the effects of lowering a threshold on a chemical constituent of a food may be illustrated.

Finally, Bayesian networks come into view as well when a hierarchical model is developed and considered appropriate. Hierarchical models are models in which the parameters are supposed to be derived from a distribution that is in turn characterized by parameters that may come from a distribution, etc. up to the required level. It goes beyond the scope of this overview as well to go deep into this issue. The interested reader is referred to Lauritzen (1996).

References

Gelman, A., Carlin, J.B., Stern, H.S., et al., 1995. *Bayesian data analysis*. Chapman & Hall, London. Texts in statistical science series.

Jensen, F.V., 1996. *An introduction to Bayesian networks*. UCL Press, London.

Jensen, F.V., 2001. *Bayesian networks and decision graphs*. Springer, Berlin.

Lauritzen, S.L., 1996. *Graphical models*. Oxford University Press, Oxford.

Lindley, D.V., 1971. *Making decisions*. John Wiley, London.

Van Der Vorst, J.G.A.J., Beulens, A.J.M., De Wit, W., et al., 1998. Supply chain management in food chains: improving performance by reducing uncertainty. Paper presented to the seventh international special conference of IFORS: 'Information systems in logistics and transportation', Gothenburg, Sweden, 16–18 June 1997. *International Transactions in Operational Research*, 5 (6), 487-499.

Appendix: the learning process

Let U be the set of configurations over a universe of variables and let Pr() be the true distribution over U taken from a database of cases. Further, let M^* be a candidate Bayesian network for Pr() and let $\mathrm{Pr}^*()$ be the distribution determined by M^*. Two distance measures between Pr() and $\mathrm{Pr}^*()$ are generally applied: the Euclidean distance

$$Dist_Q(\mathrm{Pr}, \mathrm{Pr}^*) = \sum_{x \in U} (\mathrm{Pr}(x) - \mathrm{Pr}^*(x))^2$$

and the cross entropy distance

$$Dist_L(\mathrm{Pr}, \mathrm{Pr}^*) = -\sum_{x \in U} \mathrm{Pr}(x) \log \frac{\mathrm{Pr}(x)}{\mathrm{Pr}^*(x)} .$$

Both distance measures are based on proper scoring rules. As a size measure we consider

$$Size(M^*) = \sum_i Sp(A_i) ,$$

where $Sp(A_i)$ is the number of entries in Pr(A|pa(A)), and as an acceptance measure

$$Acc(\mathrm{Pr}, M^*) = Size(M^*) + k \cdot Dist(\mathrm{Pr}, \mathrm{Pr}^*) ,$$

where a value for k is free to be chosen. As a general search method, we choose a threshold t for $Dist$ (Pr,Pr^*) and a k for Acc(Pr,M^*). Among the models with a distance to Pr less than t, we choose then one of a minimal Acc(Pr,M^*).