10

Application of Bayesian Belief Network models to food-safety science

G.C. Barker[#]

Abstract

We describe the use of Bayesian Belief Network methods for the representation of complex systems and indicate their role in analyses of food-borne hazards. We illustrate the method with two applications (i) an exposure assessment for non-proteolytic *Clostridium botulinum* in a minimally processed food product (ii) an analysis of the variability associated with the spore germination and growth lag times.
Keywords: Belief Network; Bayesian Network; food safety; variability

Introduction

Currently there is a strong desire for improved quantification and clearer expression of food-safety information. One focus for this drive concerns the integration of the data and the expertise that surrounds large-scale food production with information that relates to modern food-retailing practices, to changing consumer behaviours and to changing patterns of food-borne disease. There is an ongoing search for ways that information from these sources may be condensed into accountable expressions, and stated assurances, about the safety of manufactured food products. Recently opportunities have arisen which promote the transfer of developments and expertise, from other areas in which information dependency and uncertainty are dominant, into microbial food-safety assessments. We have explored the use of intelligent data-analysis techniques, and in particular Bayesian Belief Network methods (e.g. Jensen 1996), for the quantitative representation of food-safety information. The development is driven by requirements for improved decision support and for stronger, clearer, communications relating to food safety.

Mathematical modelling of complex systems has benefited from the rapidly improving performance of accessible computational resources. Modern desktop computer equipment can now store and process data in quantities that represent full food-borne hazard domains (material specifications, process descriptions, predictive microbiology, quality-control results, etc.) and can access remote sources (safety databases, behavioural surveys and health records etc.) at speeds which promote 'real-time' investigations, comparisons and responses. This progression has been accompanied by innovation in data-analysis techniques for large uncertain systems, e.g. neural networks, fuzzy algebras, data-mining and Bayesian Belief Networks. A host of software tools that implement these methods, and which can be tailored for use in safety assessment, are available although, in general, these are still at the research-

[#] Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK. E-mail: gary.barker@bbsrc.ac.uk

technique level. We have developed Bayesian Belief Network methods for use in microbiological food-safety assessment scenarios (Barker 2000). The developments have emphasized modular constructions in order to enable the greatest flexibility in application and to ensure a rapid pick-up rate for technological advances and innovations in other fields.

In this report we will introduce the concept of a belief network as a graphical scheme to organize information and events. We will develop this idea to show how network structures can be combined with appropriate models and data to quantify the events. This scheme builds into a full probabilistic representation, the domain model, which facilitates consistent use of probabilistic reasoning and Bayesian inference. We develop two examples: (i) a process model, for *Clostridium botulinum* in a potato product, that shows how knowledge of events may be combined subject to uncertainty; (ii) an explicit Bayesian inference that illustrates how the parameters of a model, for the lag times of a population of spores, can be estimated from observations.

Bayesian Belief Networks and Food-Safety Science

Belief networks are, simply, a sophisticated method for data analysis. A belief network converts large quantities of sometimes disparate, sometimes non-quantitative, information into a common representation that is amenable to consistent interpretation and clear expression. Bayesian Belief Networks are a variety of expert system; user implementations usually take the form of typical windows software applications. A belief network looks like a flow diagram, with labelled nodes and directed links, and it represents a complex system of interconnected variables that is called a model domain. The first stage of domain-modelling is relatively straightforward, qualitative, and often highly interdisciplinary. In practice food-borne hazard domains are modelled by a combination of process engineers, microbiologists, toxicologists and health professionals etc. who identify a list of quantities that are significant and the potential dependencies that exist between them. This expression of belief does not rely on knowledge of particular values or relationships, and a map, representing the direction of information flow in a domain, can usually be constructed relatively quickly. A domain model representing a food-poisoning hazard may contain ~100 nodes and will have a modular construction. At each point in a belief network the variables are expressed by probability distributions; this representation is an integral part of belief-network modelling and is particularly appropriate to food-safety assessment, where it is accepted that the majority of information sources are always, to some extent, uncertain. Uncertainty associated with food-safety assessments arises from many different influences, including the statistical errors from experiments, finite-sized samples and the natural variability of animal and human populations. Bayesian Belief Network modelling treats the different sources of uncertainty equitably and combines them, consistently, to reveal an overall level of confidence associated with the values of particular variables.

At first sight it appears that the consistent assignment of probability distributions to all the domain variables may be overpowering. In practice this is not the case because

each node in a belief network is assigned a 'conditional-probability table' that computes the local probabilities, in terms of the distributions on nodes at the end of incoming links (parent nodes), using the combination law

$$p(A \cap B) = p(A|B)\, p(B)$$

where p(A|B) is the conditional probability, i.e. the probability of event A given that event B has already happened. Some conditional probabilities are quite complex, with matrix representations, because the events A and B may themselves represent sets of possibilities. In many cases the construction of such a table involves re-ordering datasets that already exist as predictive models or laboratory growth curves. Modelling-expertise centres on the construction of these tables but this operation is required only once for each node and there are strong opportunities for cross developments between applications. Several systematic approaches and table generators have been developed to support belief-network constructions.

The economy of the belief-network technique arises from the strong use of causality. It is far easier to estimate the probabilities of particular causal connections, which have previously been identified by domain experts, than it is to establish full distribution functions for the domain variables in isolation. The dependency structure encoded within the domain model, by subject experts, acts to reduce the complexity of the system. This process corresponds to a particular partition of the full joint-probability function. In general, for a domain spanned by variables $\{A_1,A_2,A_3, \ldots, A_n\}$ a belief network reduces the joint-probability distribution, $p(A_1,A_2,A_3, .., A_n)$, according to

$$p(A_1, A_2, A_3, ..., A_n) = \prod_{i=1}^{n} p(A_i \mid \{\mathrm{P}arent(A_i)\})$$

where probabilities are conditional on a limited set of parents $\{\mathrm{Parent}(A_i)\}$. Once the scheme has been implemented a belief network follows closely the 'programming' paradigm where fixed instructions act on multiple realizations of data. Simple operation of a belief network involves the addition of information, often input from keyboard or mouse as changes to the probability distribution of a variable on the edge of the domain, and the observation of associated changes in other variables. In this way a belief-network representation of a food-borne-hazard domain encodes and expresses, simultaneously, the results of very many integrated food-safety calculations.

An additional property of Bayesian networks is particularly relevant to food-safety operations and to associated decision-support tasks (e.g. Malakoff 1999). Bayes' theorem connects opposing conditional probabilities p(A|B) and p(B|A). Bayesian Belief Network software tools use sophisticated algorithms (e.g. Spiegelhalter et al. 1993), to evaluate all the opposing probabilities, quickly and consistently, even in large domains. Crucially this means that a Bayesian Belief Network can, as well as work forwards to evaluate the probabilities of particular states, also work backwards to indicate what input information is consistent with observed, or required, outputs. This process is Bayesian inference – evidence and prior information combining to give posterior information. The flexibility of Bayesian Belief Network methods ensures that different users can interpret information in their own context but, because of the commonality of the information content, there can also be a rational debate concerning the actions and decisions that surround food-safety issues.

The common data representation in a belief network facilitates a wide range of endpoint measures and indicators. The most common form of output measure associated with belief networks are ‘expected utilities’. Costs and benefits, exposures, prevalence and sensitivity can all be formulated in terms of the utility-function formalism. Within the context of food-safety assessment risk may be considered as an expectation value; most simply measuring the expected size of detriment or harm associated with food consumption. Thus we may write

$$\text{Risk} = \int p(x)\, D(x)\, dx$$

where the kernal, D(x), is a numerical expression of detriment, p(x) is the probability of the detriment arising and the integral is performed over all possible realizations (denoted by a variable x). Belief-network constructions facilitate this computation. There is a rapidly growing industry in the development and application of belief-network tools; these include faster manipulations, sophisticated displays and interactive support etc. Working applications in telecommunications, finance and medicine have stimulated developments in a wide range of disciplines including food safety (e.g. Barker, Talbot and Peck 2002; Nauta et al. 2003; Pouillot et al. 2003)

Process modeling – exposure assessment in a potato product

In suitable conditions the spores of *C. botulinum* can germinate to produce vegetative cells. In turn the growth of a population of cells may produce a dangerous neurotoxin. As part of a full exposure assessment we have considered the roll of refrigerated storage in preventing the production of toxin from spores of non-proteolytic *C. botulinum* that survive heat treatment during the manufacture of a potato product. The product is a relatively simple combination of raw potato flakes, starch and other minor ingredients and has an extended lifetime under refrigeration conditions. A combination of information and data, from a variety of sources that includes the manufacturer and established predictive microbiology, can be used within a Bayesian Belief framework to show that storage is particularly safe with respect to non-proteolytic *C. botulinum* hazards. The model concentrates on a simple end point, the toxicity of an individual retail unit of the product at the end of the storage period, which is related to an individual risk. The modelling technique could easily accommodate broader and more complex hazard analyses.

A skeleton graph that represents beliefs concerning toxin production during storage of the potato product is illustrated in Figure 1. On the left-hand side of the network ‘S5’ is a discrete variable that represents the total spore load in a retail pack of potato product prior to storage; ‘Toxin’ is a Boolean variable which indicates the presence, or otherwise, of botulinum toxin in a single pack of potato product after storage. Storage time is represented by the variable ‘tstore’; all the other variables represent components of a model that describes germination and growth of *C. botulinum* populations. As a fundamental part of the domain model we have associated, with each spore, a finite amount of botulinum neurotoxin if, and only if, it is allowed to germinate and grow. In practice this process has a very low probability within a pack of the potato product, prior to consumption, and has never been observed. In this initial domain we have not identified toxin production mechanisms, nor quantified the toxin synthesis, but have aligned the cell-growth phenomenon, directly and unambiguously, with the development of toxicity. This level of description includes an element of precaution.

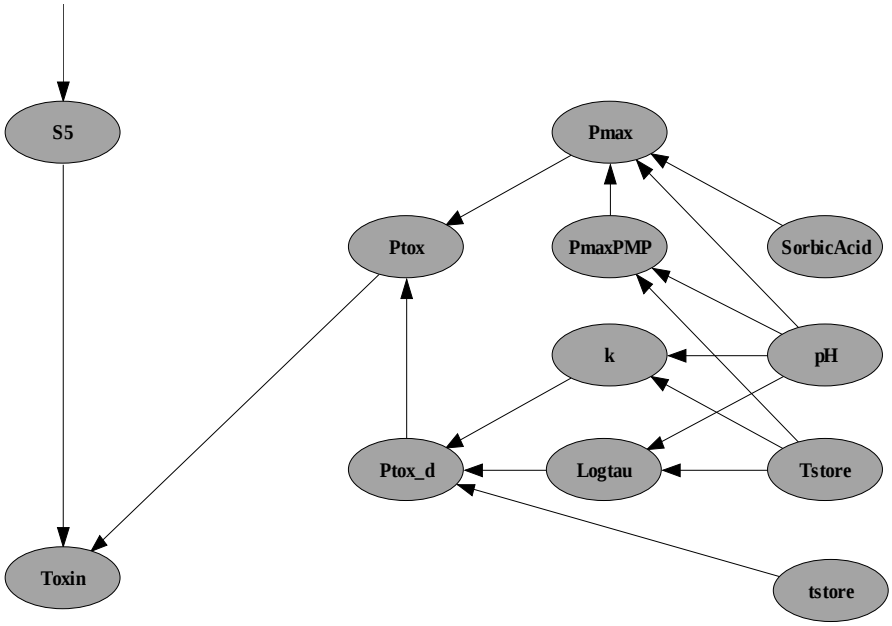

Figure 1. A Bayesian Belief Network component representing the germination and growth of non-proteolytic *Clostridium botulinum* during refrigerated storage

Underlying the graphical representation in Figure 1 each of the variables is represented by a probability distribution. These distributions represent the 'belief' or information content associated with different regions of the domain. Conditional probability tables that represent uncertain causal links have been constructed from established models. As a primary growth model we have used the probabilistic time-to-growth model for spores of non-proteolytic *C. botulinum* that is part of the USDA Pathogen Modelling Program (Whiting and Oriente 1997). The probability of growth is expressed as

$$P(growth) = P_{\max} /(1 + \exp(k(\tau - t)))$$

where P_{max}, τ and k are the parameters, of a monotonically increasing function of the incubation time t, that correspond to the maximum (long-time) probability of growth, to the mean time to growth and to a rate constant. In the initial domain model we have used restricted, two-dimensional, secondary polynomial models to represent the variation of the model parameters with the environmental conditions (pH and temperature T). Uncertainty associated with the growth-model determination is included in the domain in terms of a distribution for the time to growth parameter, τ. In the initial implementation this distribution is normal, centred on the value identified by the secondary model, with a variance of 0.1.

Additional sorbic acid is used by the manufacturer to preserve the potato product. The PMP model for growth of non-proteolytic *C. botulinum* does not include the inhibitory effect of sorbic acid. However we have included, in the initial domain model, the effect of sorbic acid on the maximum probability of growth from a single

spore. The modification included in the domain model uses a limited form of a secondary growth model, for non-proteolytic (type B) *C. botulinum* (Lund et al. 1990). We have transformed the PMP maximum-probability parameter according to

$$P^{*}_{\max} = P_{\max} 10^{f([s])}$$

where f([s]) is a polynomial of [s] (only), extracted from the full expression given by Lund et al., and [s] is the concentration of undissociated sorbic acid. This conjunction of two secondary models is an economic use of distinct information sets that relate to the growth of non-proteolytic *C. botulinum* although, in general, the two models cannot be compared directly, e.g. they use alternate cell and spore inocula etc. Uncertainty associated with the approach is included into the model domain in terms of a distribution of P_{max} values that is uniform on the interval [P_{max}^{*}, $10P_{max}^{*}$]. We have used a very narrow distribution, centred on the manufacturers specified value, to represent belief concerning the pH of the potato product (Beta[1,15,5,7]; <pH>=5.1, σ^2=0.1). We have used a normal distribution of temperature, with mean 7C and standard deviation 1C, to represent the variation of temperature during storage. This distribution represents a significant constraint on post-manufacture storage of the potato product and does not include periods of lost control or abuse. This temperature distribution represents conditions that support an analysis of the intrinsic food safety associated with the potato product.

Subject to an assumption that spores of non-proteolytic *C. botulinum* germinate and grow independently the affirmative state of the output measure, representing exposure of a consumer to botulinum toxin at the end storage of a single pack of the potato product, has a probability which is a simple product of the probability of growth for a single spore and the number of spores occupying a single retail pack. We have implemented a Bayesian Belief Network, as an interactive computer application, that is a quantitative representation of the model domain. Our implementation of this Bayesian Belief Network employs Hugin runtime software (Hugin Expert A/S, Aalborg, Denmark). This tool provides direct inspection and graphical representation of all the information and data that constitute the domain. Importantly the Bayesian network ensures that all the component pieces of information are joined, consistently, into statistical representations of uncertain variables. In addition the belief network facilitates analyses and developments that support decision-making and improved understanding in relation to the safety and the manufacture of the potato product.

Figure 2 shows the effect of the added sorbic acid on the probability that a pack of the potato product, at the end of storage, will contain toxin that has arisen from non-proteolytic *C. botulinum*. This probability is the probability of the affirmative state of the ‘Toxin’ variable in the model domain, and the data in Figure 2 can be obtained by simple interrogation of the belief network without additional computations. The potato product is particularly safe with respect to the non-proteolytic *C. botulinum* hazard.

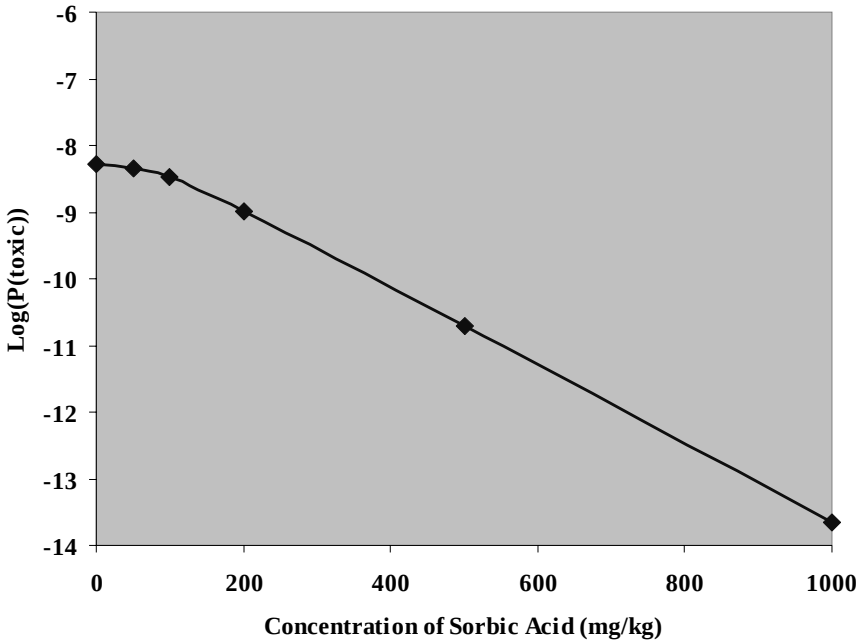

Figure 2. Probability of a toxic effect arising from non-proteolytic *Clostridium botulinum* in a potato product – variation with the concentration of added sorbic acid

Bayesian interpretation of spore variability

The development of a growing population of bacterial cells from an inoculum of dormant spores is an important component in several food-borne hazards. Of particular interest are minimally heat-processed foods that may harbour spores and, in the absence of controlled low-temperature storage conditions, may provide a suitable environment for the growth of pathogens such as non-proteolytic *Clostridium botulinum* or *Bacillus cereus* (Peck 1997). The microbial population kinetics for these hazards includes at least two distinct stages. Firstly a germination process in which individual phase-bright spores change into phase-dark spores that are not, initially, suitable for growth and division. Secondly a conversion of phase-dark spores into vegetative bacteria that have all the necessary machinery to develop into cells and then divide with a well-defined rate.

Natural populations usually include variability. That is, the individuals in the population have a range of values for each particular property or trait. Accounting for variability is important both when interpreting observations from measurements conducted on a population and when estimating the consequences of actions or events that take place within a population. In this discussion we examine the impact of the variability of the germination and growth processes, and in particular the variability of the component delay times, on the development of a population of spore-forming bacteria. In some respects population variability can be distinguished from other components of indeterminism, such as measurement uncertainty, that arise when considering large numbers of individuals. Separation of population variability and measurement uncertainty is an increasingly significant part of quantitative risk-assessment methodology (e.g. Vose 2000).

We consider individual spores that spend a time λ_G prior to germination and then a period λ_g prior to their conversion into a vegetative cell. These two periods are exclusive and exhaust the period that precedes the regular doubling regime (which has a period $T_2 = \ln 2/\mu$ where μ is a constant maximum growth rate). As a special case we may consider both delays to have shifted exponential distributions with

$$f_x(\lambda_x) = \mu_x e^{-\mu_x(\lambda_x - \lambda_x^0)} \qquad \lambda_x > \lambda_x^0$$

where x = g, G, μ_x is the distribution parameter and $f_x(\lambda_x) = 0$ for $\lambda_x < \lambda_x^0$. This distribution is consistent with observations of single spores for non-proteolytic *C. botulinum* (Webb et al. 2002). Within this framework the population lag time, for a large spore inoculum, is

$$t_{lag} = \lambda_G^0 + \lambda_g^0 + \mu^{-1} \ln((1+\mu/\mu_G)(1+\mu/\mu_g))$$

In contrast the single spore lag time, $t_{lag}(1)$, is simply the sum of the two components (distributions)

$$t_{lag}(1) = (\lambda_g^0 + \lambda_G^0) + (\lambda_g - \lambda_g^0) + (\lambda_G - \lambda_G^0)$$

Since the three terms on the right correspond to random variables this sum must be interpreted as a shorthand form of the conditional probability rather than an algebraic equality.

To examine spore variability with a Bayesian perspective we adopt a causal framework for uncertain information that is based on these equations. In a Bayesian interpretation this framework defines a space of models that is spanned by three parameters μ_g, μ_G and $\lambda_g^0 + \lambda_G^0$. These models develop two output variables, $t_{lag}(1)$, the unknown single-spore variability distribution and, t_{lag}, the observed value for the population lag time (the population growth rate is included in this model framework as a constant). The dependencies, of the parameters and the uncertain variables in this framework, are represented by the network structure in Figure 3. Each node of the diagram represents an uncertain quantity and each arrow represents a dependency of one quantity on another. In the Bayesian view this information structure does not represent a special case but, rather, defines a large space of possibilities that can be constrained by relevant information (in this case the observed population value).

In order to complete the specification of this information system each parameter must be described by a prior distribution that represents an initial state of belief. The assignment of prior distributions is a fundamental part of the Bayesian view. We have assigned uninformative distributions to express the absence of any prior knowledge concerning the three parameters μ_g, μ_G and $\lambda_g^0 + \lambda_G^0$. We have used the Hugin belief-network tool to evaluate posterior belief about the probability distribution of single-spore lag times. The posterior distribution for $t_{lag}(1)$ is obtained by combining the uninformative prior information about model parameters with specific information, or evidence, concerning the observed value of the population lag time. We have chosen an observed value t_{lag} = 430 min corresponding with the case investigated by Webb et al. (2002), i.e. non-proteolytic *C. botulinum* under good growth conditions. The Hugin tool evaluates the posterior distribution by applying an efficient information-

propagation algorithm, developed by Lauritzen and Spiegelhalter (Spiegelhalter et al. 1993), on the directed network structure shown in Figure 3.

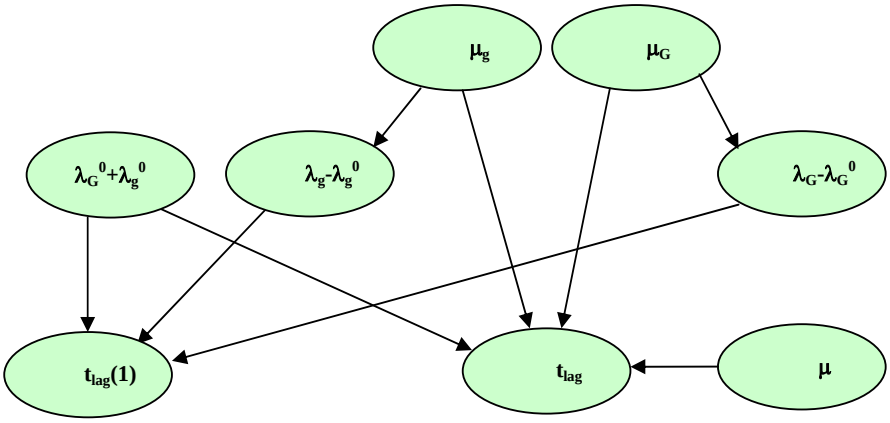

Figure 3. A directed graph structure showing the dependencies of uncertain quantities in a framework for Bayesian interpretation of spore variability

The posterior probability density for $t_{lag}(1)$ is shown as a dashed line in Figure 4. This distribution illustrates the outcome from a Bayesian combination of specific evidence, relating to the observed lag time for a population of spores, and prior ignorance concerning the values of the parameters for a model of spore variability. Figure 4 shows that the evidence acts as a constraint on the set of possible models that are described by the full ranges of the parameter values. For reference we have also shown in Figure 4, as a full line, the probability density for single-spore lag times that arises from the direct evaluation of the variability model using data from detailed single-spore experiments (Webb et al. 2002). The Bayesian posterior distribution shows, clearly, the impact of a single piece of evidence, but it also includes weight at small lag times that reflects a state of uncertainty relative to the well-defined model based on individual spore observations.

Using the Bayesian posterior distribution we may evaluate several quantities of interest. Based on the Bayesian posterior the dashes in Figure 5 show the distribution function for the cell population size that arises, after 500 min, starting from a spore inoculum with $S_0 = 5$. The distribution based on data from experimental observations is shown by the bars. The Bayesian posterior distribution leads to an increased probability that no spores will have germinated but, additionally, it indicates higher probabilities corresponding to relatively large cell populations (a point value calculation gives LogN = 1). The increased spread of the population size distribution reflects the limited information content in the Bayesian evaluation. Clearly in the presence of imprecise information the Bayesian posterior accounts for possibilities, of early germination and growth, which are excluded by more detailed data.

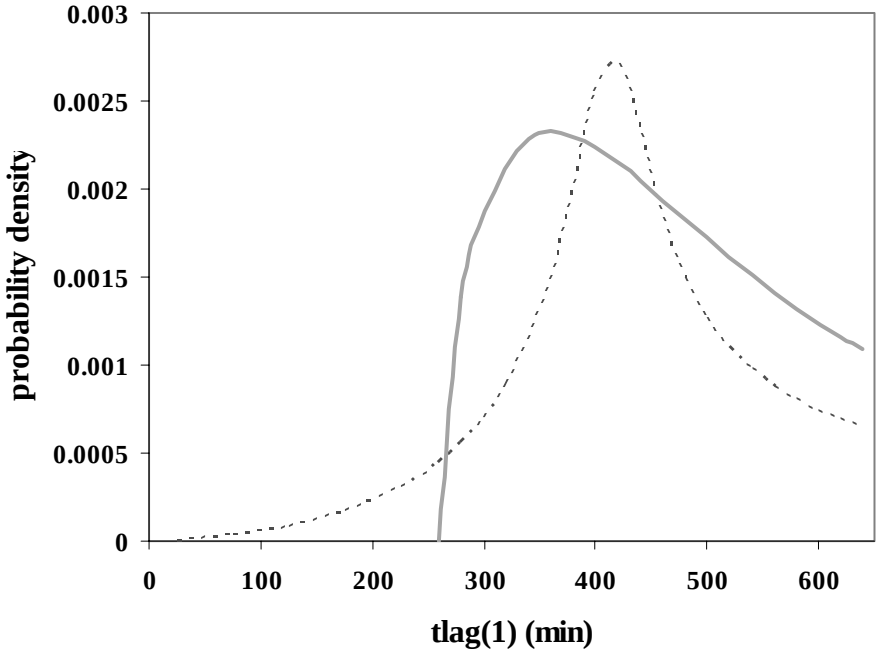

Figure 4. The variability distribution for the single-spore lag time with μ= 0.01, μ_G = 0.02, μ_g = 0.003 min^{-1} and $\lambda_G^0 + \lambda_g^0 \sim 250$ min. The broken curve shows the corresponding Bayesian posterior distribution based on an observation of the population lag time

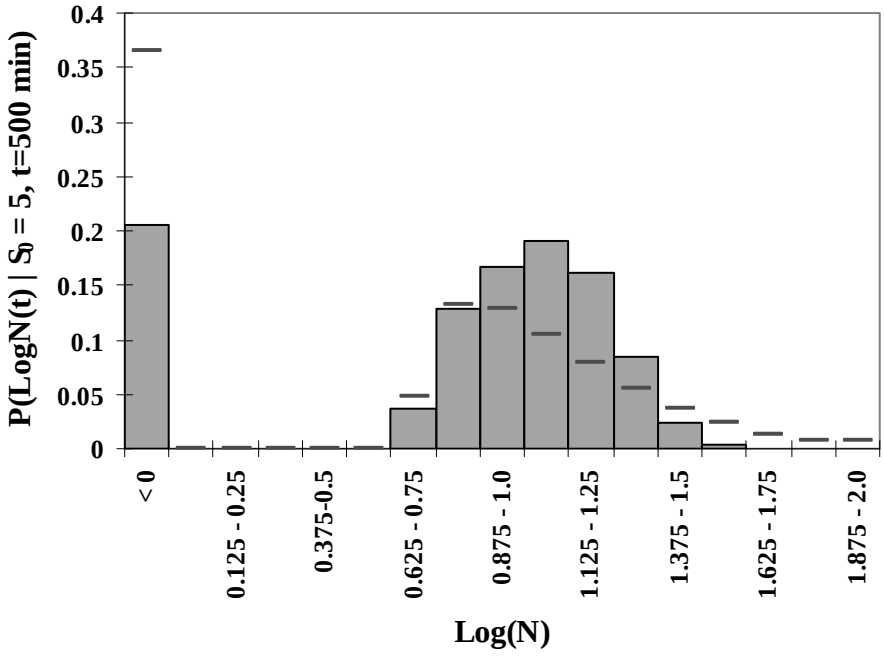

Figure 5. The probability distribution for the logarithm of the population size at t = 500 min given that the inoculum contained 5 spores (μ= 0.01, μ_G = 0.02, μ_g = 0.003 min^{-1} and $\lambda_G^0 + \lambda_g^0 \sim 250$ min). LogN(t) < 0 indicates that no cells have been produced. The superimposed horizontal bars show the corresponding distribution obtained by using a Bayesian posterior for the single-particle delay time

Although the numerical example above concerns growth and germination in particular, good, conditions the Bayesian scheme can be extended readily to other situations for which limited information, in the form of an observed population lag time and a maximum growth rate, is available.

Discussion

The belief-network method provides efficiency and consistency for food-safety computations and belief-network technology facilitates the consistent expression of multiple indicators relating to food-hazard domains. The flexibility of the belief-network approach allows these models to be contrasted and compared with established safety assessments and analyses. The models developed above have relatively simple implementations, but these could be extended, and developed, to provide non-expert-user interfaces, and external tools, that facilitate realistic interactions with complex knowledge bases.

A belief network, in its graphical representation, is a powerful tool for communications between model builders and subject experts. The directed graph reflects the causal beliefs associated with the domain. Additionally the undirected graph reflects the information flows that occur within the domain. However, ultimately, the graphical representation summarizes the dependencies between the variables within the domain and, therefore, it is a compact representation that shows the level of organization that exists within a complex system of events.

Conclusion

We have shown that belief networks are powerful tools which have practical implementations. In particular we have shown that modelling of complex processes, subject to information uncertainty, and Bayesian inference, of undetermined parameters, can be approached with Bayesian Belief Network techniques. Although, currently, most applications of belief-network methods are in communications, engineering and information technologies there are clearly many applications in food-safety science that can benefit from this approach. Food-safety assessment is, historically, an area of disparate views and fragmented information but belief-network modelling promises opportunities for unification and organization.

Acknowledgement

The author wishes to acknowledge many useful discussions with Prof. M. W. Peck and Dr. Jozsef Baranyi.

References

Barker, G.C., 2000. *Uncertainties and priorities in food safety assessments*. Food Technology International. Sterling Publications, London: 45-47.

Barker, G.C., Talbot, N.L.C. and Peck, M.W., 2002. Risk assessment for *Clostridium botulinum*: a network approach. *International Biodeterioration and Biodegradation*, 50 (3/4), 167-175.

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

Lund, B.M., Graham, A.F., George, S.M., et al., 1990. The combined effect of incubation temperature, pH and sorbic acid on the probability of growth of non-proteolytic, type B *Clostridium botulinum*. *Journal of Applied Bacteriology*, 69 (4), 481-492.

Malakoff, D., 1999. Bayes offers a new way to make sense of numbers. *Science*, 286 (5444), 1460-1464.

Nauta, M.J., Litman, S., Barker, G.C., et al., 2003. A retail and consumer phase model for exposure assessment of *Bacillus cereus*. *International Journal of Food Microbiology*, 83 (2), 205-218.

Peck, M.W., 1997. *Clostridium botulinum* and the safety of refrigerated processed foods of extended durability. *Trends in Food Science and Technology*, 8 (6), 186-192.

Pouillot, R., Albert, I., Cornu, M., et al., 2003. Estimation of uncertainty and variability in bacterial growth using Bayesian inference: application to *Listeria monocytogenes*. *International Journal of Food Microbiology*, 81 (2), 87-104.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., et al., 1993. Bayesian analysis in expert systems. *Statistical Science*, 8, 219-283.

Vose, D., 2000. *Risk analysis: a quantitative guide*. 2nd edn. Wiley, Chichester.

Webb, M.D., Stringer, S.C., Pigott, R.B., et al., 2002. Germination and outgrowth of single spores of *Clostridium botulinum*. *In: Proceedings of the 2nd International Conference on analysis of microbial cells at the single cell level, Vejle, Denmark, 2-4 June 2002*. [http://www.ifr.ac.uk/science/Posters/outgrowth.pdf]

Whiting, R.C. and Oriente, J.C., 1997. Time-to-turbity model for non-proteolytic type B *Clostridium botulinum*. *International Journal of Food Microbiology*, 36 (1), 49-60.