18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

How believable is your BBN?

Kuhnert, P.M. [1] **and K. R. Hayes** [1]

[1] *CSIRO Mathematical and Information Sciences,*
Brisbane and Hobart Laboratories
Email: Petra.Kuhnert@csiro.au

Abstract:

Bayesian Belief Networks or BBNs are gaining prominence in ecology. They are a powerful and attractive tool for managing and understanding complex processes because they represent the process graphically, where each node in the network represents either the prior or conditional probability of the parameter of interest. Causal links are represented by arcs that join to nodes and indicate the dependencies between nodes in the network. The fundamental idea behind these causal relationships is Bayes' theorem, which provides a premise for combining the prior and conditional probabilities assigned to each node of the BBN to form posterior estimates of the quantities of interest that can be readily updated. Despite their popularity and wide spread use in solving complex, large scale ecological problems of importance there are a number of issues relating to the structure of the model and incorporation of expert information that modellers need to be aware of. This leads us to ask "How can we really believe the output from a BBN, given that a large proportion of the information feeding into the network often relies on expert opinion that may be inaccurate or carry hidden biases?"

Analysts are faced with three important tasks when producing a BBN: (1) identifying the variables pertinent to the problem at hand; (2) identifying the relationships between these variables; and (3) expressing these relationships as a series of conditional probabilities. Through this implementation a number of issues can arise resulting in an ill-defined BBN which inadequately reflects the underlying system processes nor captures and reflects the expert opinion defining these processes well.

Using a popular BBN software package, Netica™, we examine some of the potential pitfalls of BBNs through two specific examples. The first focuses on a fishery management problem, which examines whether management should remain passive or seek a more active approach to commercial fishing. The second investigates the quarantine risks associated with the importation of commodities and examines the probability that pests or diseases will enter a nation with imports of goods. We examine issues of discretisation (process of converting a continuous probability distribution to one that is discrete), scaling (process of transforming data), complexity (number of nodes and linkages) and network structure (nodes and linkages defining causal relationships) in the context of these two examples and show that each can have a dramatic impact on the posterior probabilities of each model investigated and therefore have the potential to impact management decisions if implemented. In our analysis, we show that although BBNs offer a powerful mechanism for capturing both expert information and empirical data (where available) and can also address issues of uncertainty through the elicitation of conditional probabilities, there are some aspects of BBNs that the modeller needs to consider. We highlight these issues and suggest ways in which to guard against these problems so that BBNs are used more appropriately for the modelling exercise being considered.

Keywords: *Bayesian Belief Networks; Conditional probabilities; Elicitation; Model structure; Model complexity; Prior information.*

1. INTRODUCTION

Bayesian Belief Networks (BBNs) are becoming a popular tool for capturing expert knowledge in order to assist decision makers in complex, data poor situations (Peterson et al. 2008; Hamilton et al. 2007; Pollino et al. 2007; Borusk et al 2004). They represent an attractive decision support tool as they are well suited to problems with small or incomplete datasets and they are not restricted by a minimum sample size. They can be constructed using expert opinion alone and although they do not rely on any empirical data, it can be easily incorporated (Wooldridge and Done, 2004). BBNs are well suited to cross-disciplinary collaboration (Pollino et al, 2007) because they can incorporate prior information from a diverse range of disciplines. Moreover, the process of building and populating the network forces the analyst(s) to think very carefully about the mechanisms, processes and context of their problem (Peterson et al, 2008) as well as any uncertainties that arise (Moskowitz and Sarin, 1983). With the ongoing use of expert opinion in modelling and the relative ease with which experts' prior beliefs can be captured and incorporated into a model like a BBN, we have to ask the question, "how believable is a BBN?", especially one that has been constructed without much empirical evidence to support its structure or the prior and conditional prior probabilities that form its backbone. How influential are the priors in these models and what happens if the expert presents a biased judgment on how these should be represented?

Analysts who choose to use BBNs face three important tasks: (1) identify the variables pertinent to the problem at hand; (2) identify the relationships between these variables; and, (3) express these relationships as a series of conditional probabilities. The latter probably being the most difficult. The variables and relationships between them determine the structure of the network, while the conditional probabilities elicited from experts and incorporated into a set of conditional probability tables (CPTs) describes the prior beliefs about the relationships between variables, conditioned upon this structure. Although this process seems relatively straight forward to implement, there are a number of issues that can arise:

1. *Discretistation* - process of converting a continuous probability distribution to one that is discrete and therefore has the potential to produce different discretisations depending on the choice of bins used (mandatory with Netica)
2. S*caling* – process of transforming (empirical) data e.g. taking logs, z-score, subtracting the mean, etc.;
3. *Structure* – nodes and linkages (or dependencies) defining the causal relationships of the BBN; and
4. *Complexity* – the number of nodes and linkages in the BBN, and the number of discrete states within each node,

These issues have the potential to result in an ill-defined BBN which does not properly reflect the information elicited from the experts about the underlying process of the problem investigated. Analysts who have had little experience with BBNs or who have a lack of understanding of the underlying theory or indeed of the subject domain that they are modelling fall into this category. Analysts seeking "realism" can, and often do, build large networks with twenty or more nodes and hundreds or even thousands of conditional probabilities that can be cumbersome to populate and maintain (Borsuk et al, 2004; Walton and Meidinger, 2006; Pollino et al, 2007). Hence, this relatively simple form of modelling can lead to errors in the structure of the model and unforeseen biases in the way in which information is elicited, incorporated and interpreted from the model.

In this paper we address the 4 issues identified above and illustrate their impact in two ecological examples using the popular BBN package, Netica™. The first focuses on a fishery management problem which examines whether management should remain passive or seek a more active approach. The second investigates the quarantine risks associated with the import of commodities and examines the probability that pests or diseases will enter a nation with imports of goods. Our investigations reveal that without proper consideration of the network structure, complexity and the translation of expert opinion into CPTs with the correct scaling and discretisation (for models that require discrete probability distributions) interpretations from the BBN can become severely biased and have the potential to mislead management in making decisions based on this process. Due to space limitations we omit an overview of the BBN methodology and direct readers to the paper by Lauritzen and Spiegelhalter (1988).

1. CASE STUDY 1: FISHERY MANAGEMENT PROBLEM

The Northern Prawn Fishery (NPF) is one of Australia's most valuable commercial fisheries, bringing in millions of dollars each year from the commercial trawling of nine species of prawns in the region. In addition to capturing prawns, byproduct such as squid, scallops, bugs and cuttlefish are also captured and kept as they are seen as a potential source of income by fisherman. The Australian Fisheries Management Authority (AFMA) has placed emphasis on determining abundance and reproductive rates for byproduct and their vulnerability to trawling (AFMA 2006) and this has the potential to place restrictions on the catch if rates are deemed too high. A problem associated with placing restrictions on byproduct catches is determining how many of each species are being caught. Although byproduct species are regularly caught during commercial prawn trawling activities conducted in the NPF, information about catches concerning byproduct is limited as the focus is generally on the target catch.

In an attempt at managing the two species of bugs within the NPF with very limited empirical information, we constructed a BBN using Netica™ that incorporated monitoring data, predictions from models and expert information. Due to the sensitivity of this project, we consider a simplified version of this model in this paper with fictitious priors as shown in Figure 1. The aim with this type of modelling is to determine whether to passively manage (i.e. continue with the "business as usual" approach) or actively manage, which may involve additional monitoring or a change in management strategy such as an increase in the minimum legal size. In order to determine the management position, we need to compare in any one year what is actually caught, y_c (via commercial catch records) to some "trigger" value, *ST* that measures sustainability. The trigger is determined using a fisheries catch equation (Quinn & Deriso, 1999) which incorporates an estimate of the biomass, natural mortality, *M* and the sustainable amount of fishing required. This sustainable level is determined by the probability of compliance, *C*, the optimal fishing mortality, *F* and the management strategy, *S* that is actually implemented. Here we only consider two management strategies, *Passive* or *Active*.

Priors for nodes *C* and *S* (Figs. 2a and 2b) were discrete and resulted from an elicitation process while the priors chosen for nodes *B* and y_c were based on empirical catch data from survey vessels and commercial log-book data respectively. Both of these parameters were assigned Normal distributions for log catch as shown in Figures 2c and 2d. Natural mortality and fishing mortality are defined as constants in this model. Natural mortality was based on a study of bugs by Courtney (1997), while *F* represented the optimal fishing mortality from a yield per recruit analysis under no size restriction. As highlighted earlier, the nodes defining the sustainability trigger, $ST = B + \log(SF) - \log(SF + M) + \log(1 - e^{-(SF+M)})$ and the sustainable level of fishing, $SF = C \times S + (1 - C) \times F$, represent deterministic functions of parameters causally linked to that node in the BBN. Note that the equation for ST is presented on the log-scale.

The very top node of the network represents the management position, *D*, either *passive* or *active*.

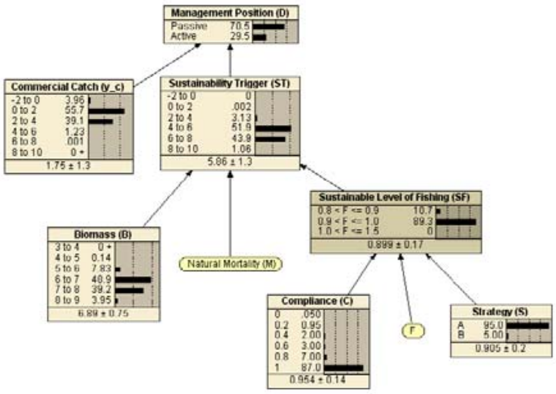

Figure 1. A conceptual model for managing a single spatial region in the NPF that was derived through stakeholder interaction at a workshop and input into a BBN framework.

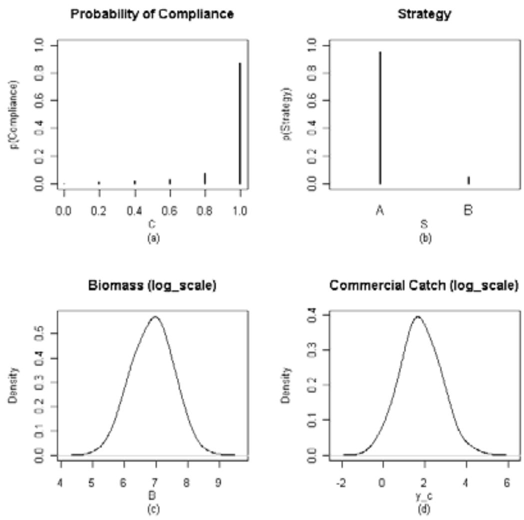

Figure 2. Priors used in the fisheries model for (a) compliance C, (b) strategy S, (c) biomass B and commercial catch, y_c. Note, biomass and commercial catch estimates are shown on the log-scale.

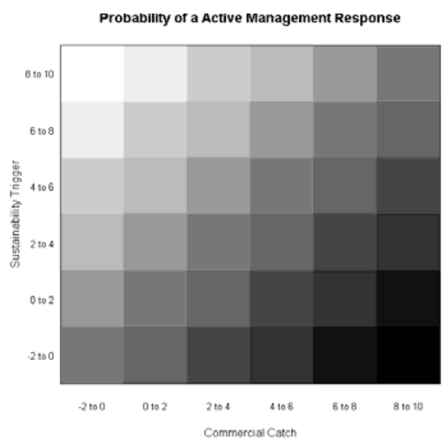

Figure 3. Conditional probability tables for actively managing a fishery. Black pixels indicate a probability of 1 while white pixels indicate a probability of 0.

Figure 3 shows the discrete conditional distribution for the probability of a particular management position given the sustainability trigger (y-axis) and commercial catch (x-axis). In this diagram, conditional probabilities are represented by a grey scale image, where 1 is represented by a black pixel and 0 is represented by a white pixel. As commercial catch increases with respect to the sustainability trigger, the probability of instigating a more active management position increases, linearly. Taking into account all of the priors and deterministic relationships, we see that the priors set up in this BBN lead to a low probability of actively managing the fishery: p(*Active*) = 0.295.

2. CASE STUDY II: IMPORT RISK ASSESSMENT

Between 1950 and 2002 the total volume of goods and commodities traded around the world increased exponentially (WTO, 2003). Imported commodities are an important source of non-native species, and the rate of biological invasions around the world has consequently increased in an equivalently exponential fashion (Levine and D'Antonio, 2003). Bio-invasion risks are managed around the world via a range of international and national regulations and guidelines that mandate or recommend a variety of risk assessment schemes (Hayes, 2003). In this case study we examine an attempt to quantify and improve the qualitative import risk assessment developed by the European and Mediterranean Plant Protection Organisation (EPPO). The EPPO risk assessment consists of a series of 34 unstructured questions relating to the probability of entry, establishment and spread of non-native species that may be associated with an imported commodity (EPPO, 2007). Figure 4 illustrates an early attempt by researchers to improve the consistency, transparency and uncertainty analysis of the risk assessment by converting it into a BBN. Note that for the purposes of brevity we only consider here the assessment of entry risk. The nodes in the BBN closely resemble the questions in the original risk assessment scheme, but the analysts have reduced the number of discrete values from five options (very unlikely, unlikely, moderately likely, likely, very likely) to three or four, presumably in an attempt to reduce the size of the CPTs. All of the CPTs were subsequently populated with simple estimates.

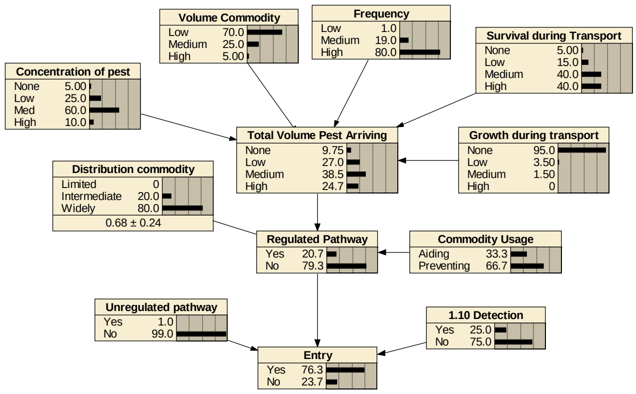

Figure 4. An initial BBN for the probability of entry based on the EPPOs quarantine risk assessment scheme.

As an illustration of the complexity of the network let us examine the node representing the total volume of pests arriving at a new location and the large number of conditional relationships that it seeks to represent. Close inspection of the CPT underlying this node reveals a large table of conditional probabilities (576 x 9), representing all possible combinations of the 9 conditioning variables (Concentration: 4 categories; Volume: 3 categories; Frequency: 3 categories; Survival: 4 categories; Growth: 4 categories). Note the node "unregulated pathway" quantifies the risk that vectors or pathways that are not assessed under the EPPO risk assessment scheme are active. Assuming there is very little chance of an unregulated pathway, this network suggests that the probability of pest entry is very high: p(Entry) = 0.763 (Figure 4).

3. ISSUES ANALYSIS

3.1 Discretisation

BBNs are quite sensitive to discretisation, particularly in BBN packages such as Netica™ which does not have the capacity to use continuous variables. The term "discretisation" used here refers to the process of taking a continuous distribution, selecting appropriate bins and allocating observations to intervals prescribed by those bins. For example, consider converting the continuous prior distributions shown in Figure 2(c) and 2(d) into discrete distributions, each having only two states, low (L) and high (H), where the cutoff for a low catch is 4 on the log-scale. (Note we also discretised the *ST* node into low and high catch for simplicity.) The commercial catch node now has a prior probability of a low catch of 0.99 while the biomass node has a prior probability of 0.14 for a low catch. Re-running the BBN reveals that the probability of moving to a more active management position, given the other parameters in the model and these newly formed priors changes from 0.295 to 0.025, a very unlikely result. If the cut-off for low catch was defined as 2 on the log-scale then the posterior probability of an active management response is 0.224. This demonstrates the sensitivity of

BBNs in relation to how a distribution is discretised and the number of states used in the discretisation process. Reducing the number of node states is an important, and recommended (Marcot et al, 2006) way to simplify CPTs in complex networks. In our example the change in the outcome does not change substantially, for more complicated models however, coarse discretisation could have dramatic effects.

3.2 Scaling

As in many cases of Bayesian hierarchical modelling, it is often wise to scale your data before incorporating it into a model as it leads to more stable estimates and faster convergence. We have found this also applicable to BBNs. We refer to "scaling" as a process of data transformation, where the data are transformed in some way (e.g. centred, divided by the standard deviation, log-transformed) for the sole purpose of making the estimation process more robust. Working with empirical distributions especially when they are heavily skewed as in the case of the biomass and commercial catch data, can lead to problems with the specification of priors in a BBN, particularly in packages like Netica™. We have found in instances where we have used the raw data and developed distributions accordingly, that the estimates are quite variable. If the BBN package cannot handle a prior presented as a continuous distribution then the discretisation issues presented above can complicate issues further as the choice of bins can lead to highly variable estimates at these nodes. Furthermore, if "uncertainty" is chosen in the Monte Carlo sampling option of Netica™, the estimates can be highly variable from run to run. For example, if we replace nodes B and y_c with log normal priors with evenly spaced bins of 0 to 1000g/ha, we find skewed results for the deterministic *ST* node. This leads to a management position weighted towards passively managing the fishery (p(*Passive*) = 0.898) and introducing more bins seems to exacerbate this further.

3.3 Network Structure

BBNs allow managers to explore the effects of variability and uncertainty (lack of knowledge) on nodes or decision variables within a network. This facet is widely touted as one of their key strengths (McCann et al., 2006). BBNs can easily represent the effects of variability within a system, but they are not so well suited to uncertainty analysis. A key aspect of uncertainty in this context is the structure of the model. Analysts rarely explore this however, because it requires the specification and parameterisation of alternative networks. This is time consuming and, in practice, the effects of this type of uncertainty are rarely explored, despite the fact that errors in the network's structure are far more important than numerical inaccuracies within it (Druzdzel and van der Gaag, 2000).

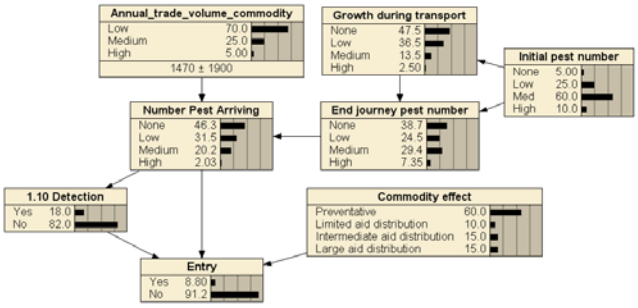

Figure 5. An alternative BBN for the probability of entry based on the EPPOs quarantine risk assessment scheme.

The structure of the EPPO risk assessment BBN (Figure 4) was designed to emulate the EEPO schema. The conditional independence between survival and growth and between volume of pest arriving and detection, however, is questionable. Figure 5 presents an alternative, arguably more plausible structure that explicitly allows for negative (or positive) density dependant growth effects during transit, and conditions the probability of detection upon the number of arriving individuals. Note the reversal of the probability of entry. This effect is partly due to the negative density dependence relationships (that might plausibly be positive), but is overwhelmingly due to the positive dependence between probability of detection and number of pests arriving with a commodity (which is very unlikely to be anything other than positive).

Finally it is important to note that the analyst may encounter other types of uncertainty in the structure of a BBN beyond the nodes and arcs. The "unregulated pathway" node in Figure 4, for example, is an example of "model domain" uncertainty (Morgan and Henrion, 1990) - in this case whether or not the pest will enter the nation through some pathway other than that which is being assessed by (in this case) the network. In this example it makes very little sense to be uncertain about the domain of the assessment, particularly in the manner expressed in Figure 4. Domain uncertainty here is expressed by a parent-less node whose child is the endpoint of the analysis (probability of entry), and is therefore very likely to be the most important node in the entire network. It might be better expressed by a separate network(s) that address other pathways.

3.4 Complex versus Simple

There are clear advantages in simplifying a BBN: the elicitation process is much simpler and requires *less* from the experts and the model is also much simpler to interpret. However, although the model is less complex, it also may miss key pathways that could potentially lead to a different end result. Furthermore, when experts construct simple models or parameterize simple CPTs, they are more likely to omit hidden beliefs and assumptions about the nature of the variables' relationships than if more detail were represented. In more complex models, there can be more transparency, and greater capacity to examine all plausible pathways. Take for example the fisheries management model in Figure 1. One way to simplify this model is to strip the network back to 3 nodes: D, y_c and ST. We could either choose to elicit a continuous probability distribution that is discretised in Netica™ or alternatively, opt for a discrete specification of priors for these nodes. We examine the latter here so we can concentrate on the complexity of the model. In this example, we categorise commercial catch and sustainability into low, moderate and high values with probabilities assigned to each that reflect the continuous prior distributions shown in Figure 2. The result of this specification is that the posterior probability of actively managing is now 0.38, considerably higher from the estimate produced from the model in Figure 1.

4. DISCUSSION AND CONCLUSIONS

BBNs are becoming an important modelling tool for ecological applications where data is limited. They offer a very powerful and convenient way of incorporating information – whether in the form of expert and/or experimental data – by way of prior and conditional probabilities that easily accommodates variability. The ease with which these models can be updated also makes them a very attractive model to use when experimental data become available as decisions can be made relatively quickly. Despite these features, there are a number of issues that modellers need to be aware of when constructing and evaluating a BBN and these were demonstrated using two case studies. In summary we recommend that modellers follow a few key measures when putting together a BBN for a problem:

1. **Implementation and choice of software:** Think carefully about the software you intend on using for the BBN implementation as some BBN packages do not deal with continuous probability distributions easily and require methods for discretisation.

2. **Representation of priors:** Consider carefully how priors and conditional probability distributions will be structured in the BBN.

    - If a continuous distribution is represented as a discrete measure, ensure that the states chosen are an adequate representation of the continuous process that has been elicited. The discretisation should therefore become an integral part of the elicitation process to ensure the information elicited is being represented adequately in the network.

    - Parameters that can lead to the elicitation of highly skewed distributions should be treated carefully as they can lead to spurious results in the model. Scaling of such variables using a log-transformation for example may need to be considered at the elicitation phase.

3. **Network structure & complexity**: There are clear advantages of simplifying a network, however in doing so ensure that the network adequately describes the key processes in the system being studied. Importantly, do not parameterise the BBN before exploring other plausible model structures via influence diagrams (usually available as unparameterised networks in commercial software platforms) or methods, such as qualitative modelling and cognitve maps (Ramsey and Veltman, 2005; Dambacher et al., 2002), that enable rapid analysis of alternative model structures. Avoid pursuing a "best" model and wherever possible maintain alternative plausible models (Pascual et al., 1997).

4. **Choice of BBN:** As BBNs are heavily reliant on expert opinion it is difficult to determine whether the model developed is "good" in the sense that the structure of the model and its associated dependencies accurately reflect the underlying system processes and the conditional probabilities and priors populating the model are accurately represented exhibiting little or no bias. It is our view that

    - BBNs need to undergo rigorous development through a series of workshops containing a broad range of experts;

    - workshops should consider (and document) at least two alternative models for analysis, validation and interpretation until it can be shown that one model outperforms the other(s);

    - obtain experimental data for nodes of the network to assist in further refinement of the models and validation which would lead to a final model that could be used in a decision support context.

ACKNOWLEDGMENTS

We acknowledge Dave Milton for initiating the fisheries problem and the many helpful discussions with Min Zhu, Quinton Dell and Mark Tonks relating to this project. We thank Willem Roelofs for providing access to the EPPO risk assessment BBN and the 2 anonymous reviewers who reviewed this manuscript.

REFERENCES

AFMA (2006) Northern Prawn Fishery Strategic Plan 2001-2006, Northern Prawn Fishery Advisory Committee, Canberra ACT.

Borusk M.E., Stow C.A. & Reckhow K.H. (2004) A Bayesian network of eutrophication models for synthesis prediction and uncertainty analysis, *Ecological Modelling*, 173: 219-239.

Courtney, A. (1997) A study of the biological parameters associated with yield optimization of Moreton Bay Bugs, Thenus spp., Fisheries Research and Development Corporation (FRDC), Technical Report, Southern Fisheries Centre, Department of Primary Industries, Deception Bay.

Dambacher J.M., Li H.W., and Rossignol P.A. (2002) Relevance of community structure in assessing indeterminacy of ecological predictions. *Ecology*, 83(5): 1372-1385.

Druzdzel M.J., and van der Gaag L.C. (2000) Building probabilistic networks: "Where do the numbers come from?"Guest editors introduction *IEEE Transactions on Knowledge and Data Engineering*,12(4):481-486.

European and Mediterranean Plant Protection Organisation (EPPO) (2007) Decision Support Scheme for Quarantine Pests. Secretariat of the EPPO, Paris, France, 34pp. Available online: http://www.eppo.org/QUARANTINE/Pest_Risk_Analysis/PRA_intro.htm (accessed February 2009).

Hamilton G., Fielding F., Chiffings A., Hart B., Johnstone R., and Mengersen K. (2007) Investigating the use of a Bayesian network to model the risk of lyngbya majuscule bloom initiation in Deception Bay, Queensland Australia, *Human Ecological Risk Assessment* 13, 1271-1287.

Hayes K.R. (2003) Biosecurity and the role of risk-assessment, pp. 382-414 in Ruiz GM and Carlton JT, (Eds), Bioinvasions: Pathways, Vectors, and Management Strategies. Island Press, Washington, D.C., USA.

Lauritzen S.L., and Spiegelhalter D.J. (1988) Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistics Society*, B50: 157-194.

Levine J.A., and D'Antonio C.M. (2003) Forecasting biological invasions with increasing international trade. *Conservation Biology*, 17(1): 322-326.

Marcot B.G., Steventon J.D., Sutherland G.D., and McCann R.K. (2006) Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. *Canadian Journal of Forest Research*, 36: 3063-3074.

McCann R.K., Marcot B.G. and Ellis R. (2006) Bayesian belief networks: applications in ecology and natural resource management. *Canadian Journal of Forest Research*, 36: 3053-3062.

Morgan M.G. and Henrion M. (1990) Uncertainty: A guide to dealing with uncertainty in risk and policy analysis, Cambridge University Press, Cambridge.

Moskowitz H and Sarin RK (1983) Improving the consistency of conditional probability assessments for forecasting and decision making. *Management Science*, 29(6): 735-749.

Pascual M.A., Kareiva P., and Hilborn R. (1996) The influence of model structure on conclusions about the viability and harvesting of Serengeti Wildebeest. *Conservation Biology*, 11(4): 966-976.

Peterson D.P., Rieman B.E., Dunham J.B., Fausch K.D. and Young M.K. (2008) Analysis of trade-offs between threats of invasion by nonnative brook trout (Salvelinus fontinalis) and intentional isolation for native westslope cutthroat trout (Oncorhynchus clarkii lewisi). *Canadian Journal of Fisheries and Aquatic Sciences*, 65: 557-573.

Pollino C.A., Woodberry O., and Nicholson A.E. (2007) Parameterisation of a Bayesian network for use in ecological risk assessment. *Environmental Modelling and Software*, 22: 1140-1152.

Quinn T., and Deriso R. (1999) Quantitative Fish Dynamics, Oxford University Press, US.

Ramsey D., and Veltman C. (2005) Predicting the effects of perturbations on ecological communities: what can qualitative models offer? Journal of Animal Ecology, 74: 905-916.

Walton A., and Meidinger D. (2006) Capturing expert knowledge for ecosystem mapping using Bayesian networks. *Canadian Journal of Forest Research Research*, 36: 3087-3103.

Wooldridge S., and Done T. (2004) Learning to predict large scale coral bleaching from past events: A Bayesian approach using remotely sensed data, in situ data, and environmental proxies. *Coral Reefs*, 23: 96-108.

World Trade Organisation (WTO) (2003) International Trade Statistics, 2003. World Trade Organisation, Geneva, 232 pp. Available online: http://www.wto.org/english/res_e/statis_e/its2003_e/its2003_e.pdf (accessed February 2008).