

School of Computer Science and Software Engineering
Monash University

Bachelor of Computer Science Honours (1608), Clayton Campus

Research Proposal - Semester 1, 2003

**Knowledge Engineering a Bayesian Network for an Ecological
Application**

Owen Grant Woodberry
ID: 12093009

Supervisors:
Ann Nicholson
Kevin Korb
Carmel Pollino

November 10, 2003

Contents

1	Introduction	3
2	Research Context	3
2.1	Bayesian Networks	3
2.2	Knowledge Engineering	4
2.3	The Application	4
2.4	Analysis of knowledge engineering process to date	5
3	Research Plan and Methods	7
3.1	Research Methods	7
3.2	Proposed thesis chapter headings	8
3.3	Timetable	9
3.4	Special facilities required	10
4	Deliverables	10

1 Introduction

Knowledge engineering is the process of building expert systems, such as Bayesian networks. It involves investigating an application domain, identifying important concepts and expressing these concepts as objects and relationships in the formal representation of the expert system. The knowledge engineer is, usually, either trained in the formal representation of the expert system or is an expert in the application domain, but often not in both. For this reason the knowledge engineering task can be seen to address two issues. The first, how can someone, especially someone who isn't an expert in the application domain, properly identify the important concepts? And second, how can someone, especially someone who isn't trained in the formal representation of the expert system, express the concepts that they have identified as important? To address these issues the field of knowledge engineering emerged, its objective, to formalise the process of building expert systems, ensuring they are created correctly and used to their maximum potential.

It is also important to consider this engineering process in the domain for which the system is to be used. This is to account for differences in the viewing of the concepts of the domain, by experts and users, and the differences in what the system is to be used for. This project will analyse a Bayesian network application in development, created by someone without formal training in the technology, identify any shortcomings and undertake improvements upon it. The Bayesian network to be considered is being developed for an ecological application.

2 Research Context

2.1 Bayesian Networks

Bayesian networks (BN), also known as belief networks, knowledge maps, probabilistic causal networks among other names, are tools for reasoning with probabilities. They consist of a graphical structure component composed of nodes representing random variables with qualitative relationships between them, representing causal influences, and a quantitative component composed of a conditional probability table (CPT) for each node representing the effects of its parent nodes on it [6] [2] [7] [5]. The power of Bayesian networks lies in its ability to calculate the probability of an event given a set of evidence from only a small set of probabilities defined in the CPTs. It does this by using Bayes rule and conditional independence relationships between variables to reduce the number and form of conditional probabilities required to represent the problem.

2.2 Knowledge Engineering

The process of knowledge engineering a Bayesian network can be broken up into three tasks. The first two tasks relate to defining the graphical structure of the network. The first task is identifying variables of importance, along with the possible states that they can take. And the second task is to define the causal relationships between these variables. The last step is to define the conditional probability tables for each node [4]. Along with properly defining the problem domain, an objective of the first two steps is to express the problem in its simplest yet sufficiently complete form. This is done to reduce the number and form of probabilities to be entered into the CPTs which is to be done in the third task, and is often considered the most difficult task.

2.3 The Application

This application is an ecological risk assessment (ERA) in which a BN is being used as a component. The overall objective of the ERA is to develop and test a generic framework to be used in the assessment of ecological risks associated with Australian irrigation activities. The case study being conducted at Monash University is specific to the Goulburn Broken catchment. The geographical areas of the catchment being considered are broken into regional and local scales. The local scale chosen was the Goulburn Weir/Lake Nagambie, and the regional scales are the Goulburn River reaches and major tributaries from Eildon to Seymour and Murchison to the Murray River. Time scales to be considered are 1, 5, 10 years, and 30 to 50 years; these were picked to properly reflect the life history of the fish species in the area. Phase 1 of the ERA, the problem formulation phase, identified native fish abundance and diversity in the catchment as being at risk due to irrigation activities in the area [3]. Phase 2 of the ERA is to identify and quantify these hazards, and to develop a model to measure the probability of increasing or decreasing the population abundance and diversity in response to varied management interventions. This phase is currently being undertaken at the water studies centre at Monash University, under Dr Carmel Pollino. Bayesian networks were selected as the best model methodology to do this. During Phase 1 of the ERA, a conceptual model was created to show possible factors influencing the native fish population and diversity, a simplified version of this model is shown in Figure 1, the double arrow links are included to demonstrate possible interaction between factors.

A Bayesian network is currently in production; it already has a preliminary graphical structure and CPTs.

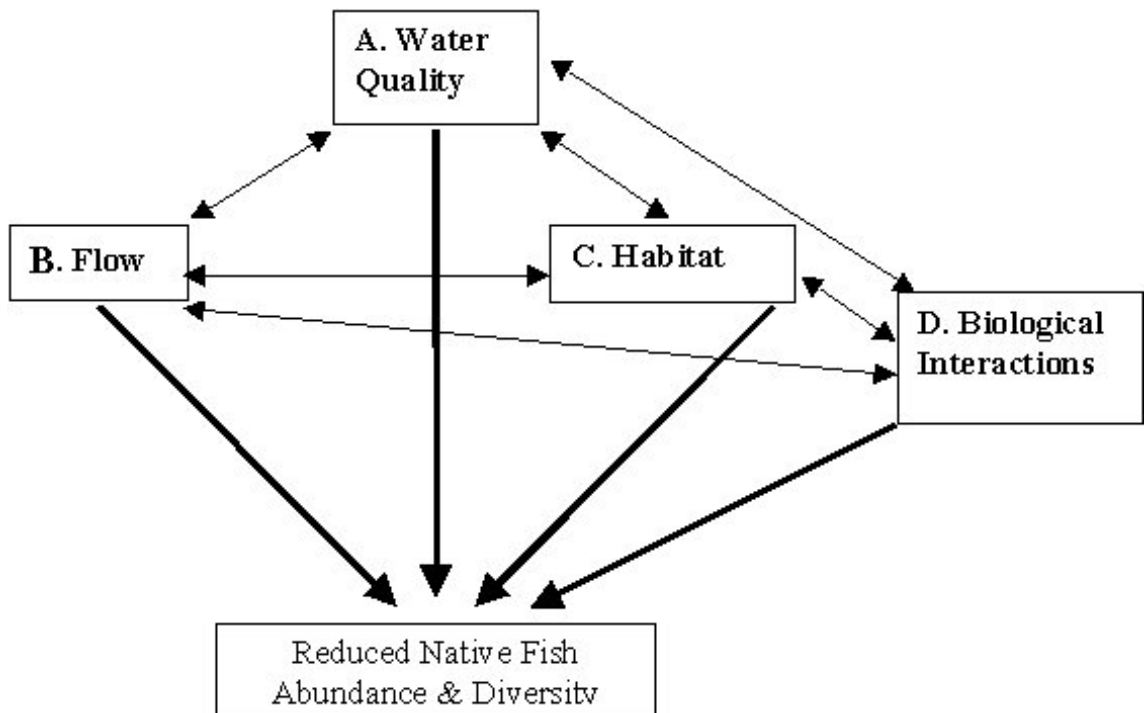


Figure 1: Conceptual Diagram

2.4 Analysis of knowledge engineering process to date

To aid in the process of knowledge engineering such a Bayesian network there are various guidelines in existence [?] (Cain, 2001). The guidelines provided by Jeremy Cain were consulted in the production of the Bayesian network to date.

To determine the graphical structure of the Bayesian network a workshop was held in which the stakeholders recommended improvements to the conceptual model, shown above, expanding it. This expanded conceptual model was used to identify important variables and their relationships. To determine the possible states that these variables could take existing literature was consulted to retain consistency with existing data. The resulting Bayesian network is shown in Figure 2.

This network is broken into four sub-networks or portions based on the four factors in the conceptual model, these can be identified by the nodes Water Quality, Overall Change in Flow Regime, Structural Habitat Quality and Competition, and their parent nodes. The two query nodes are Native Population Abundance and Native Population Diversity. Evidence is

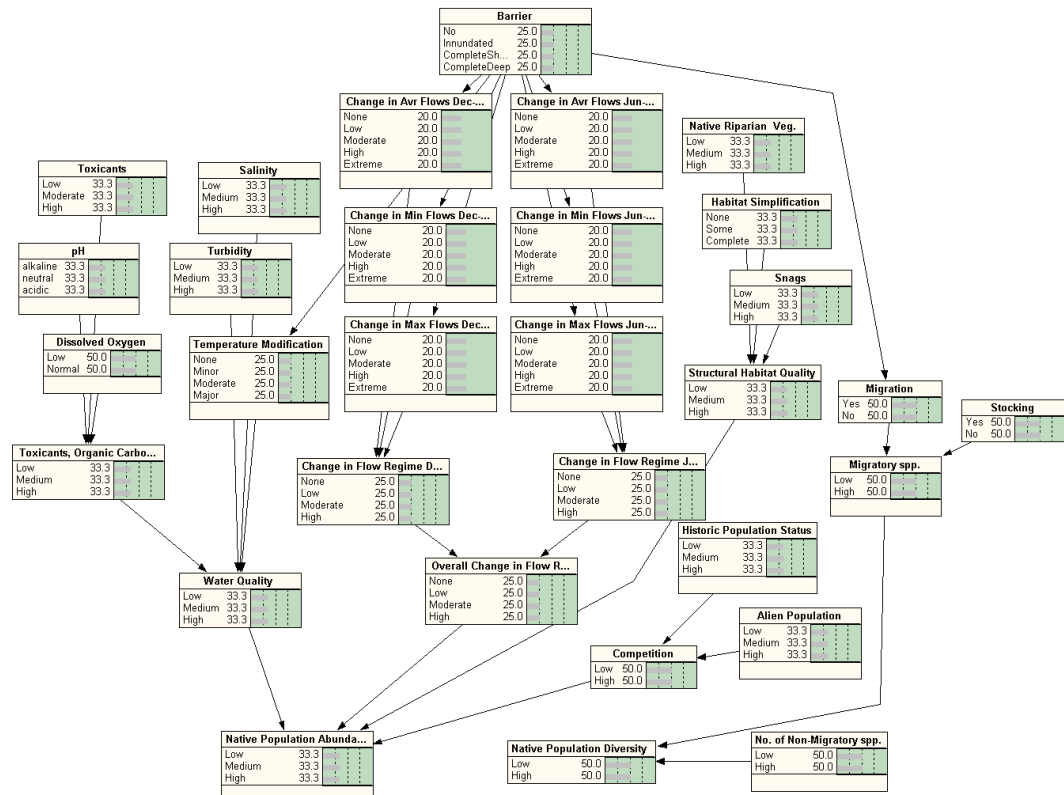


Figure 2: Bayesian Network to date

to be entered into some or all of the nodes; Barrier, Stocking, Temperature modification, Toxicants, pH, Dissolved oxygen, Salinity, Turbidity, Change in Average flows Dec-Feb, Change in Average flows Jun-Aug, Change in Minimum flows Dec-Feb, Change in Minimum flows Jun-Aug, Change in Maximum flows Dec-Feb, Change in Maximum flows Jun-Aug, Change in flow regime Dec-Feb, Change in flow regime Jun-Aug, Overall change in flow regime, Native Riparian Veg, Habitat Simplification, Snags, Historic Population Status, Alien Population, Migratory spp. and No. of non-migratory spp.

To determine the probabilities of the CPTs a generic network with generic CPTs was created for all the study sites, this was done based on a scoring system and assumptions based on expert knowledge. These CPTs were then further trained for each site using values from experimental data for each site entered into a case file. This data was compiled using data from various organizations, including the index of stream condition (ISC), the Victorian water data warehouse, expert opinion and other sources. An-

other workshop has been planned for later in the year to consult stakeholders and verify these CPTs. The result of this process is a number of Bayesian networks, with different CPTs, for each site. At this point in time there is no method for dealing with varying time scales.

3 Research Plan and Methods

3.1 Research Methods

This project will continue the analysis, started above, of the knowledge engineering process taken in the development of this Bayesian network to date. This analysis will include identification of the knowledge engineering tasks that have not been undertaken, could be improved or have not been done formally. Then after considering these issues and making any improvements required continue with the development of the Bayesian network.

This project will examine the effects of combining the experimental data with elicited CPTs. A preliminary look into these effects shows that there may be insufficient weight given to the elicited CPT values. This weighting could be defined as the 'experience or 'history of the CPT value. It is important to specify the confidence in these elicited CPTs, as they will be considered, by default, to have the same experience as a single data case, by any software tool that does not know the history of the CPT. This project will explore the effect of weighting these elicited values against experimental data values, determining relationships between the confidence in elicited values and the amount of data available.

This project will examine possible measures to incorporate networks from different sites into one global network. A way of doing this could be to include new site and/or type node, which would be a parent node to all site/type dependent nodes, this would also require labelling the experimental data cases with an additional site/type field.

This Project will examine possible measures to deal with varying time scales. A Bayesian network represents a single time step, to consider the changes over multiple time steps the network needs to be extended. This extension could be the addition of a new time node influencing the query nodes directly or could be the repetition of the BN for each time step, called a dynamic Bayesian model. To extend the model in either of these methods it will be necessary to define the time scales of changes and predictions.

This project will examine different Bayesian network evaluation methods, identifying suitable methods for the application and apply one or more of them. Evaluation methods are useful to grade the Bayesian network and

identify errors. Such an evaluation method is to test the network using cases, by withholding values for particular nodes in the case file, the predictive accuracy of the network can be determined.

Most evaluation methods developed are used to check the values of the quantitative component, the CPTs, of the BN. This project will trail an evaluation program [1] to check the qualitative component, the structural component, of the BN.

In addition to the points mentioned above this project will identify where possible support tools would be helpful and, time permitting, implement some.

3.2 Proposed thesis chapter headings

1. Introduction
2. Background
 - (a) Bayesian Networks
 - (b) Knowledge Engineering
 - (c) Ecological Risk Assessment
3. Combination of Data Sources
 - (a) Review of Methods
 - (b) Implementation
 - (c) Results
4. Measures to Incorporate Site BNs
 - (a) Review of Methods
 - (b) Implementation
 - (c) Results
5. Measures to Incorporate Temporal Component
 - (a) Review of Methods
 - (b) Implementation
 - (c) Results
6. Evaluation Methods
 - (a) Review of Methods
 - (b) Testing

- (c) Discussion
- 7. Matilda
 - (a) Introduction
 - (b) Testing
 - (c) Discussion
- 8. Conclusion
- 9. Bibliography
- 10. Appendix A Sample Data
- 11. Appendix B Support tools

3.3 Timetable

Date	Week No.	Activity
3/3	1/13	Project Selection
15/3	2/13	Identification of the Knowledge Engineering Process to Date
16/4	7/13	Prepare Research Proposal Draft
23/4	7/13	Research Proposal Draft
30/4	8/13	Research Proposal Due
5/5	9/13	Commence Literature Review
26/5	12/13	Prepare for Interim Presentation
5/6	13/13	Interim Presentation
4/6	13/13	Prepare Literature Review Draft
11/6	1/6	Literature Review Draft
16/6	2/6	Commence Combination of Data Sources
23/6	3/6	Thesis Chapter 3 - Combination of Data Sources
30/6	4/6	Commence Measures to Incorporate Site BNs
7/7	5/6	Thesis Chapter 4 - Measures to Incorporate Site BNs
14/7	6/6	Commence Measures to Incorporate Temporal Component
21/7	1/13	Thesis Chapter 5 - Measures to Incorp. Temporal Component
30/7	2/13	Literature Review Due
4/8	3/13	Thesis Chapter 2 - Background
11/8	4/13	Commence Evaluation Methods
18/8	5/13	Thesis Chapter 6 - Evaluation Methods
25/8	6/13	Commence Matilda
1/9	7/13	Thesis Chapter 7 - Matilda
8/9	8/13	Thesis Chapters 1 and 8 - Introduction and Conclusion
10/9	8/13	First Draft of Thesis
20/10	13/13	Prepare for Final Presentation
30/10	14	Final Presentation
1/11	14	Final Draft of Thesis

3.4 Special facilities required

The facilities offered honours students at Monash University, Clayton campus are sufficient for this research project.

4 Deliverables

A complete and functional Bayesian network that is to be applied as part of an ecological risk assessment. This is being done working in collaboration with the creator of the initial network, Dr Carmel Pollino, from the Water studies Centre (Monash University).

Implementation of all or some of possible support tools identified in the duration of this project.

References

- [1] T. Boneh. Support for graphical modelling in Bayesian network knowledge engineering: a visual tool for domain experts. Master's thesis, Dept. of Computer Science, University of Melbourne, 2003.
- [2] Eugene Charniak. Bayesian networks without tears. *Artificial Intelligence Magazine*, 12:50–63, 1991.
- [3] P. Cottingham, R. Beckett, P. Breen, P. Feehan, M. Grace, and B. Hart. Assessment of ecological risk associated with irrigation systems in the goulburn broken catchment. Technical report, ACT, Cooperative Research Centre for Freshwater Ecology, 2001.
- [4] M.J. Druzdzel and L.C. van der Gaag. Building probabilistic networks: Where do the numbers come from? Guest editors introduction. *IEEE Trans. on Knowledge and Data Engineering*, 12(4):481–486, 2001.
- [5] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, Englewood Cliffs, New Jersey, 2003.
- [6] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, Ca., 1988.
- [7] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Series in Artificial Intelligence. Prentice Hall, Englewood Cliffs, New Jersey, first edition, 1995.