ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

UNCERTAINITY MANAGEMENT TECHNIQUE TO SUPPORT BIOLOGICAL MODELING FOR CONSERVATION OF PRIORITY TREE SPECIES

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

BY

BEHAILU GETACHEW WOLDE

MARCH 2009

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

UNCERTAINITY MANAGEMENT TECHNIQUE TO SUPPORT BIOLOGICAL MODELING FOR CONSERVATION OF PRIORITY TREE SPECIES

BY

BEHAILU GETACHEW WOLDE

MARCH 2009

Name and Signature of Members of the Examining Board

_____ _____

_____ _____

_____ _____

DEDICATION

I dedicate this thesis to my elder sisters, Fekade Getachew and Addis Getachew, who costed everything they had to me.

ACKNOWLEDGMENT

Many thanks are extended to my advisor Dr. Rahel Bekele for her wonderful constructive comments and for refining my work, and to Ato Getachew Berhan from IBC, Addis Ababa for his useful support during my research work.

I would like to thank my brothers and sister, Biniyam Getachew, Guche Mekecha and Alemteshay Getachew for their support and encouragement during my study. Their unchanged love and encouragement was always driving me forward.

I am very grateful to the management and staff of Institute of Biodiversity Conservation, especially to Dr. Gemedo Dalle, Dr. Alishun Ahmed, and Ato Milikiays for their unreserved support in accessing the data and in providing relevant information to my research work.

I am also grateful to my class mates and staff members of Department of Information science and Faculty of Informatics. Special thanks to my friends Amanueal, Alemayehu and Mohammed for their kind help and lovely relationship that we had had.

Above all, I thank Almighty God who provides me with so many invaluable persons.

Behailu Getachew

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

LIST OF APPENDICES

Page

LIST OF ABBRIVIATIONS

BBNs	Bayesian belief networks
BDN	Bayesian decision network
BN	Bayesian network
BNJ	Bayesian Network tools in Java
CPT	Conditional Probability Table
ConservThreatStatus	Conservation Threat Status
DAG	Directed acyclic graph
GBS	Global biodiversity strategy
IBC	Institute of biodiversity conservation
IUCN	World Conservation union
MIT	Medical Institute of Technology
QBE	Query by Example
RD	Relative density
RDOM	Relative dominance
RF	Relative frequency
SQL	Structured Query Language
TPDA	Three phase dependency analysis

ABSTRACT

Bayesian belief networks (BBNs) are useful tools for modeling biological predictions and aiding species conservation and managing uncertainty in decision-making. This paper provides practical indications for predicting, building, testing, and eliciting BBNs. Primary steps in this process include preparing data for experiment and predicting of the hypothesized "causal(dependency) relationship or conditional independence" of major biological factors affecting the target tree species or biological outcome of interest.

A total of 1200 cases and 9 attributes were used for BN model prediction with 10-fold cross validation and building BBN model before elicitation process; and reinforcing the model after experts' opinion; testing and visualizing the model with instance examples to see the conditional probabilities of the predictive inference thereby evaluating the final application model have been conducted respectively.

To this end, the average prediction accuracy for the BN model is 75.76%, and this is a promising indication for the domain experts to make decision in their future endeavors. The paper also shows that the Bayesian network classifier has a potential to be used as a tool for prediction of biological modeling to forward about conservation actions in the field of forestry. In general, the whole research process can be a good input for further in-depth study and thus, making a good pragmatic analysis in the real world situations.

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND

Tree Species in Forest

According to Newton et al (2003**),** it is widely recognized that forests are the most biologically diverse terrestrial ecosystems and that pressures on forest biodiversity continue to increase throughout the world. Around 350 million of the world's poorest people depend almost entirely on forests for their basic needs and 2 billion people depend on wood for cooking and fuel; millions of others depend on trees for food and medicines. Trees are also the structural components of forests, providing a habitat for many other species and defining the characteristics of forest ecosystems.

However, information is limited on the distribution and conservation status of tree species. Preliminary surveys undertaken to date suggest that approximately 8,000 tree species are threatened with extinction[1] worldwide. The potential loss of nearly 10 per cent of all tree species is a major conservation issue, requiring international attention and widespread action (Oldfield et al., 1998).

In principle, the ideal conservation of biological resources is the preservation of all life forms everywhere; but this is practically impossible for two major reasons: first, people must use the resources as a means of livelihood; second, the necessary resources to implement management and conservation activities are limiting or often lacking. This calls for a systematic prioritization of the biological resources for conservation purposes. One principle for conservation management that seems to be useful is to direct effort to those resources that are most valuable in terms of socioeconomic, at most risk (highly threatened with extinction), and most effectively manageable (Namkoong et al, 2000).

[1] The definition of "extinction" means species that has been disappeared world wide before ten years ago (IUCN, 1994).

According to Taye et al. (2002), it is recognized that priority[2] tree species are the one to be considered for conservation, and conservation biologists are usually in charge of attention to those species. The basic reasons are: first, they should be conserved in order to preserve or enrich the biological species diversity, and secondly, they have great contributions in its economic and social values both at the community and national level.

Based on (IUCN, 1994) priority tree species for conservation are generally those most threatened with extinction, and having the characteristics of declining rapidly, restricting to small areas (endemics) or few remaining individuals.

1.2 THE PROBLEM AND ITS IMPORTANCE

According to (Marcot et al., 2006; Mead et al, 2006) quantifying the relationships between biological variables[3] (e.g. Important value index, regeneration status, logging intensity, relative density, relative frequency, relative dominance, sapling, seedling and degree of threat[4]) to evaluate or obtain from a set of possibly related observations remain fundamental problems of prediction /causal inference in forestry.

In addition to this, a great deal of relevant information does exist; however, much of it remains inaccessible to decision-makers because it resides only in the scientific literature or even in unpublished reports or observations. Collating the information and making it available to a wide audience is a challenging issue in the current existing situations and thus, doing an appropriate research, for instance shifting them to automating system, is inevitable to be happened (Newton et al., 2003).

In many literatures, for instance, (Newton et al., 2003), information about the biological status of the tree species and its economic importance is often lacking. As many trees are very long-

[2] Priority tree species are the target tree species that need attention for conservation. http://www.unep-wcmc.org/resources/. Hereafter I use Target tree Species on this paper.

[3] Biological "variables", "traits" or "attributes" can be interchangeably used in this thesis.

[4] Degree of threat means threat category or conservation threat status (Taye et al., 2002).

lived, it is often difficult to assess how rapidly a species is likely to become extinct and its causal relationship with the environments. This implies that there is great uncertainty about a specific species in order to make a good decision.

In line with the above premises, conservation biologists make management decisions for target tree species (i.e., endangered and threatened species) under severe uncertainty. Although frameworks for formal decision-making (Jeffrey 1992) have been applied in conservation contexts (Possingham 1997), the full suite of uncertainty is rarely considered (Regan et al. 2002). Failure to acknowledge and treat the sources of uncertainty can lead to poor management decisions.

In forestry or biology, conservation focuses on target tree species and of course, it is feasible and practical making a decision on the threatened species with extinction (Namkoong et al, 2000). Moreover, biological (or ecological) modeling without assuming uncertainties make the planning unrealistic and intractable (inflexible) or making a decision with few options for the future leads to uncertainties due to the fact of many tree species may have long-lived rotation periods (e.g., for some tree species, harvesting time may take about 100 years) Therefore, to make reliable optimized decisions and to be certain about it, we have to deal with uncertainties in the real world. This will, however, make the problem much more complicated to model. Often, availability of complex relationships will arise in the process of prediction with increased the number of attributes, and such complexities lead to uncertainties about the basis for observed causal relationships and dependencies in the forest ecosystems. Fortunately, the Bayesian network[5] approach can help us with this reasoning under uncertainty and causality (Schroth et.al. 1996).

[5] Belief network, Bayes Net and Bayesian belief network have the same meaning with Bayesian network (Cheng et al., 1998). However, Bayesian belief network and Bayesian network can be interchangeably used in this paper.

The current reports have also indicated that the Institute of Biodiversity Conservation (IBC), Addis Ababa, have indeed suffering from modeling and documenting about the causal relationship between the biological attributes that have a significant contributions for conservation planning. Hence, evaluating the conservation status or degree of threat of species is unlikely enhancing to predict its uncertainty (Taye et al, 2002).

Most of the decisions and communications rely on manual work through referring the hard copy. In reality, this leads to various sources of errors to make decisions as to the conservation measures and it will also create a complexity in terms of efficiency and validity of estimating the degree of threat for the target tree species. It probably ensures that predicting/measuring a degree of threat without appropriate causal modeling like machine learning (Bayesian network) might be very unreliable and not confidential. In conclusion, collecting, identifying and publishing out the threatened species for preliminary conservation practices are not sufficient to make a good decision **(**Girma, 2002).

On similar problem domain, Samir (2001) recommended that it is more appropriate to further study on tree species with machine learning tools like (Bayesian network). His major reason was due to the presence of uncertainty of biological systems that would be stayed in the future.

It is obvious that automating and modeling the biological system often involves working with complex systems operating under conditions of uncertainty due to its dynamic change. However, over the past half century, Bayesian methods have emerged as a preferred method for reasoning with uncertainty due to their mathematical foundation. Although Bayesian theory does not solve all problems in probabilistic reasoning, it has given scientists a sound framework within which uncertainty can be represented and analyzed pragmatically. By

looking at systems probabilistically, the models constructed explicitly represent the uncertainty in the underlying system (Mead et al, 2006).

Hence, understanding and effectively conserving complex biological systems therefore require a multidisciplinary approach. A modeling approach such as that afforded by BBNs can represent the complexity of ecosystem and resource-conservation systems in hierarchical ways by decomposing or partitioning the problem into solvable steps, clearly representing value-laden concepts by empirical parameters, and combining knowledge from different domain experts (Cain et al. 1999).

Developing causal model for target tree species allow us making further inference and estimating realistic planning not only for a single institution but also for other institutions that have similar goals and objectives.

The outputs of the research can be used in designing a full-fledged prediction model in the future. It also paves the way to develop and implement a full-fledged machine learning classifiers. Thus, such a prediction model help experts provide quicker and better services to their research output and get support on evaluating conservation measures for threatened species.

The Bayesian network is expected to ensure as a sort of a reference and comparison model for the experts to provide a quality and optimal decision. Finally, the domain experts can have a chance to see the difference between biological modeling tools and thus, selecting the best statistical tools.

This study, therefore, attempts to formulate research questions that arise from the discussion of the problems.

1.3 RESEARCH QUESTIONS

The major research questions that have been emanated from this problem are, therefore:

- What are the major factors and its causal relationships that affect the occurrence of the target tree species?
- How can one build the Bayesian network (BN) for degree of threats/conservation threats?
- How good is the BN prediction model for degree of threats/conservation threats?

1.4 OBJECTIVES OF STUDY

1.4.1 General Objective

The general objective of this study is to investigate the application of Bayesian network as a causal (i.e., fundamental) model for the purpose of biological conservation, and reaffirming its potential to evaluate the degree of threats or conservation status of target tree species in Ethiopian contexts.

1.4.2 Specific Objectives

The specific objectives of the study include the following:

- To analyze and organize the domain knowledge acquired using appropriate tools.
- To examine/investigate the variables and the domain knowledge /skill used in biological modeling and conservation of target tree species.
- To investigate the extent of applications of Bayesian network model in relation to modeling the biological variables for target tree species.
- To build a Bayesian network model that predicts the degree of threats for target tree species.
- To make conclusions and recommendations for future work.

1.5 METHODS/APPROACHES

Briefly, this section informs about the data collection, analysis, modeling and evaluation in order to achieve the above mentioned research questions and objectives. Detailed explanations of this part are also presented in chapter 4 of this thesis work.

1.5.1 Data Collection

The relevant data for the study has been maintained in MS Access since 2002. The database consists of 25 tables, accessing the appropriate fields (attributes) was not easy to collect directly since it required going through each of the tables in the database. To make the process of acquiring the data very short and have good back ground knowledge as to the problem domain, the preliminary tasks were conducted as follows:

(i) Informal discussions were made with the domain experts in order to make easy understanding of the database information as a whole; there by understanding of the major biological attributes and where they were locating and how one could access them with simple query was performed.

(ii) For this study, five related tables were identified (see figure 4.1) for further data analysis.

1.5.2 Data Analysis

After properly collecting the data from the tables in the database with query by example (QBE) and importing a total of 6000 records into Ms Excel 2007 was done for data cleaning. The reason to select MS Excel was found to be easy and simple office tool that automatically could detect and filter out the missing and irrelevant data. After data cleaning, we remained with only 1200 or 20% dataset (i.e., cases) for further analysis. Then, the dataset was imported to MS Access to be used as input for discretization of the continuous data.

Discretization of the continuous attribute values were done first in Netica Application **(http://www.norsys.com)** (see figure 4.4 and 4.5) and then, domain experts also provided comments on discretized state values, however, they found it with little or no significant difference from their experiences. Following this, mapping each discrtized state values in to quantifiable nominal values were done based on experience of domain experts so as to make sound and match with the real world situations (see table 4.3).

Finally, parameters or the first row of the attributes (i.e., LoggingIntensity, IVI, RD, RDOM, RF, RS, SaplingNo, SeedlingNo, and ConservThreatStatus) were adjusted in a way the Bayesian belief network tool could take it without any semantic and syntax errors during model construction.

1.5.3 Building and Training the Model

In order to build the Bayesian belief network model, PowerConstructor system **(www.cs.ualberta.ca/~jcheng/bnsoft.html/)** and cross validation technique were used. In this method a training set was divided into two sets – an estimation set (which is used to estimate probability models) and a validation set (which is used to evaluate the performance of the estimated probability models).

Partitions of the dataset were done with 10-fold cross validation technique. Often, we chose this technique for two major reasons as indicated in Whitten et al. (2005) and Cheng et al. (1998).

These are: in the first place, extensive tests on numerous datasets, with different learning techniques, have shown that 10 is about the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up; secondly, it is the

standard evaluation technique in situations where only limited data (i.e., small sample size) is available.

We used 1200 records, among which, 120 data set were used for testing internally. Training and testing was done at the same time with BN tool. The estimated prediction accuracy was also conducted repeatedly ten times and generated its corresponding confusion matrix (see APPENDIX III). Then, the average prediction accuracy was computed for each ten run folds before and after eliciting the opinions of domain experts. The BN model prediction was found to be a good tool to further enhance the performance and prediction of the model. The result, of course, did not have a big difference before and after reinforcing the BN model.

Building BN model for biological network was implemented using BN tool before and after eliciting the opinions of domain experts respectively as indicated in figure 4.9 and 4.10.

1.5.4 Evaluating the Model

In order to evaluate the model a 10-fold cross validation technique, model structure validation with domain experts, and visualizing of the CPT for target variable “ConservThreatStatus’ with respect to direct cause effect relationship or dependency between attributes have been performed. For instance, two examples of instance classification and two samples of posterior (or conditional) probabilities of predictive/causal inference have been taken to make evaluation on the results of CPT values. The inference results that are predicted has made sound while comparing it with the real situations (i.e., experience of domain experts) (see figure 4.14).

In order to make both building and estimating the prediction accuracy of BN model, the software tools we have used for preparation of data for experiments as well as for experimental analysis are:

(i) Netica Application for discretizing purpose due to its favoring to visualizing while doing it,

(ii) BN PowerConstructor for building and constructing the network model,

(iii) BN PowerPredictor (**www.cs.ualberta.ca/~jcheng/bnsoft.html/)** for training and testing the BN model, and

(iv) BNJ **(http://bnj.sourceforge.net/)** for visualizing the CPT values in the table presentation format.

Besides to this, we have used MS Excel 2007 for data cleaning purpose and MS Acess 2007 has been used as an input storage for BN model to build and construct the biological attributes for the target tree species. To this end, all the BN tools that we have used for experiments are free or open source software.

1.6 SCOPE AND LIMITATIONS OF THE STUDY

Because of time and resource limitations,

(i) The research is focused on modeling major biological attributes for the target tree species.

(ii) The available resources are also not sufficient in the area of problem domain with its application of Bayesian network.

(iii) The research has dealt with only on the case of learning structure with complete data; records have also been taken for building and training the model only from the archived data source, database.

The target tree species' records that have been taken for data analysis are not considered at the national level, but only from the South West Ethiopia. Thus, the sample records that have been used in the study are not representative. Hence, it is beyond the scope of the current work.

1.7 ORGANIZATION OF THE THESIS

This thesis is organized in to chapters. Chapter one is dedicated to introduction where background, problems and its importance, research questions, objectives and methods/approaches, and scope and limitations are presented.

Chapter two presents review of literature on challenges and opportunities of modeling in biodiversity conservation.

Chapter three deals with review of literature on statistical approach to address uncertainty that comprise uncertainty, Bayesian probability, Bayesian network, causal Bayesian network and application of Bayesian network.

Experimental Modeling and Results are presented in chapter four while conclusions and recommendations have been forwarded in chapter five.

CHAPTER TWO

CHALLENGES AND OPPORTUNITIES OF MODELING IN BIODIVERSITY CONSERVATION

2.1 INTRODUCTION

Biological diversity (or biodiversity) refers to the variety and variability among living organisms and the ecological complexes in which they occur. Diversity can be defined as the number of different items, their relative frequency and their relative abundance. Thus, the term encompasses different genes, species, and the broad scale to ecosystems (GBS, 1992). The species is the taxonomic category ranking immediately below genus; it includes closely related, morphologically similar, individual organisms that play a particular ecological role. Species diversity refers to the variety of different species. Genes represent the basic unit of genetic inheritance within the species. Genetic diversity refers to the variety of genes. Ecosystem diversity, in addition to fostering species and genetic diversity, enhances our quality of life through recreation, aesthetic enjoyment, and spiritual enrichment opportunities (GBS, 1992).

According to Girma (2002), Institute of Biodiversity Conservation (IBC) has done a number of activities focusing on biodiversity conservation and research on biological resources as well as ecosystem management and biotechnology. Today, IBC is about 30 years old, and it is the oldest in Africa. One of the major components of the biological resources that IBC underlines is conserving forest resources. Having the realization of the importance of forest genetic resources, currently, IBC has established a department to cater for the conservation and sustainable use of forest genetic resources. In general, biodiversity conservation deals with vital and essential elements for the human well-being.

2.2. BENEFITS OF BIODIVERSITY CONSERVATION

The benefits of conservation for biodiversity are closely related to the economic values they represent and their impact on sustainable development. Species and their physiological processes, e.g. biomass production or biochemical processes, have always been considered on a material basis and as renewable capital for the primary production sector (e.g. agriculture, forestry or fisheries**)(**McNeely 1988). The availability of potential benefits in the biological diversity has a great opportunity to see its implications in the human well-being as well as its social values attached to it.

The potential benefits of biodiversity conservation can be categorized as shown in the following table:

Benefits	Short Descriptions
(a) **Biological Resources**	• Food for humans and for cultivated animals • Medicinal and pharmaceutical resources • Breeding stocks, population reservoirs • Resources not yet identified (future resources) • Wood products • Ornamental plants and animals • Potential agents for crop improvement or biological control
(b) **Ecosystem Services**	• Protection of water resources • Soils formation and protection • Nutrient storage and cycling • Pollution breakdown and absorption • Contribution to climate stability • Maintenance of ecosystems • Recovery from unpredictable events
(c) **Social Benefits**	• Research, education and monitoring • Recreation & tourism • Cultural values

TABLE 2.1: BENEFITS AND ITS DESCRIPTION OF THE BIODIVERSITY CONSERVATION (**GBS, 1992).**

2.3. CONSERVATION RELATED TO TARGET TREE SPECIES

According to (Namkoong et al, 2000), for two major reasons, it is practically impossible to conserve all the life forms everywhere unless we focus on the target tree species. First, people must use the resources as a livelihood; second, it is expensive to conserve with the available

skills and resources. In line with this assumption, as explained in (Taye et al., 2002), hundreds of target tree species have been identified due consideration for conservation. Samples of these target species have been taken from the South West Ethiopia **(**see APPENDIX I**).** According to (Taye et al., 2002; IUCN 1994), species evaluations usually consider two equally important criteria. First, socioeconomic criteria that refers to the actual and potential social, economic and ecological values/uses of the target species, and secondly, conservation status criteria that refers to the rate of biological attributes and performances of the species. This research work focuses on the conservation status criteria that have been known for measuring conservation threat. The reason is that these are the major biological attributes which are of importance in biological modeling for conservation of tree species as explained in (Taye et al., 2002).

Biological or biodiversity researches in particular with tree species (see section 2.4 below), however, indicated that there are a number of challenges and opportunities in the process of modeling of biological variables such as regeneration status, important value, relative density, relative frequency, relative dominance, disturbance of target tree species (i. e., logging intensity) and other related factors. These variables (or attributes) are believed to be the major biological factors that can serve as indicators for threaten the target tree species (Taye et al., 2002).

2.4. MODELING IN BIOLOGICAL RESEARCH

According to Marcot et al (2006) refining our understanding, quantifying relationships, generating causal about the relationships between biological (or ecological) predictor variables and response variable, and forecasting potential effects of management or conservation[6] actions are primary goals of biological research. Biological models and related

[6] Conservation is synonymous with management. (**Marcot et al., 2006).**

decision-support frameworks are simplifying abstractions of knowledge (Jones et al., 2002) that provide structure to what we know, and need to know, about a system of interest. Such abstractions are necessary to help define problems, convey biological concepts and relationships (either known or assumed), characterize potential system responses to conservation perturbations, and evaluate alternative conservation policies.

We contend that models are particularly effective when they represent complexity, causality and uncertainty in a clear and intuitive fashion. Any model, however, will be founded on limiting assumptions. Models are not intended to be perfect descriptions of reality and resultant predictions will always be imperfect (McCarthy et al., 2001). Nonetheless, models are contributed greatly to biological conservation of species when they have used and invoked further field research leading to new insights, model revisions, and more accurate predictions of the potential effects of conservation decisions. Such an approach to modeling fits well with the application of BBNs in biological modeling for conservation of species (Marcot et al., 2006).

Most problems in biological resource conservation are characterized by scant data and uncertainty about how biological systems function and respond to specific human activities (Starfield and Bleloch 1986). This presents two challenges for biological resource conservations:

(i) how to make good, science-based resource conservation/management decisions; and

(ii) how to best acquire the data needed to improve understanding. These are also related problems in biological resource modeling.

Uncertainty and the inherent complexity of biological conservation systems have been cited as a basis for legal challenges to the biological models credibility and associated biological resource-conservation decisions (Taylor et al. 2000).

More over, ecosystems are composed of heterogeneous, complex networks that exhibit nonlinear and transient (short-lived or temporary) behavior (Green et al., 2005). Multiple interactions occur within ecosystems among plants and abiotic (e.g., climatic, topographic) variation of species and system parameters (Olson et al., 1990a).

The next chapter reviews the Bayesian probability, Bayesian network, causal Bayesian network and application of Bayesian network.

CHAPTER THREE

STATISTICAL APPROACH TO ADDRESS UNCERTAINITY

3.1 INTRODUCTION

In order to address the above mentioned problem domain, appropriate statistical tools should be used. Before discussing the approach, first a brief concept of uncertainty has been tried to introduce in order to relate the statistical approach like Bayesian probability and Bayesian network.

3.2 UNCERTAINTY

Uncertainty is a lack of information or knowledge (Kanga and Kangas 2004) and is a property of our limitations in observing or understanding the system (Finkel 1996). Difficulties in estimating system parameters arise from bias and sampling errors due to imperfect sampling techniques, and from measurement errors. Limitation in obtaining sufficient information about a system's behavior prevent correct specification of causal relationships among system parameters and lead to incorrect specifications of the underlying model (Finkel 1996). Uncertainty about parameter estimates and causal relationships often can be reduced with additional research (Finkel 1996).

Uncertainty is one of the components of quantitative risk assessment from which it invokes its own treatment and interpretation in decision-making. Modeling uncertainty can involve eliciting expert judgment to determine probability distributions (Kanga and Kangas 2004). Under uncertainty the true levels of the decision are unknown (Kanga and Kangas 2004) because the expected outcome of the decision might not actually occur. Being cognizant of this fact, the recommended powerful knowledge representation and plausible statistical approaches that can handle probabilistic reasoning under conditions of uncertainty is Bayesian

network (Marcot et al., 2006; Cheng et al., 1998). Next consecutive sections are discussing about the issues related to Bayesian probability and Bayesian network with its applications.

3.3 BAYESIAN PROBABILITY THEORY

(i) Events

As explained in (Heckerman, 1995), the language of probabilities consists of statements (propositions) about probabilities of events. The probability of an event **a** is denoted **P (a).** An event can be considered as an outcome of an experiment (e.g., a coin of flip), a particular observation of a value of a variable (or set of variables), an assignment of a value to a variable (or set of variables), etc.

As a probabilistic network defines a probability distribution over a set of variables, **V**, in our context an event is a configuration, **x ε Dom(X)**, (i.e., a vector of values) of a subset of variables **X ⊆ V.** For instance, in figure 3.2(see page 22), the set of variables, V= {LoggingIntensity, RD, RF, RDOM, SeedlingNo, SaplingNo, IVI, RS, ConservThreatStatus} and the domain of the ConservThreatStatus $\langle HIGH, LOW, NORMAL \rangle$. are presented respectively. So, an event can be HIGH, LOW or NORMAL. From this we can conclude, HIGH ε{HIGH, LOW, NORMAL}.

Assume (see figure 3.1) we observe C **=yes** and **R = yes**. This evidence is given by the **"event" = (C = yes, R = yes),** and the probability, P (ε) denotes the probability of this particular piece of evidence, namely that both C = yes and R = yes are observed. Specifically, P(C=yes, R=yes).

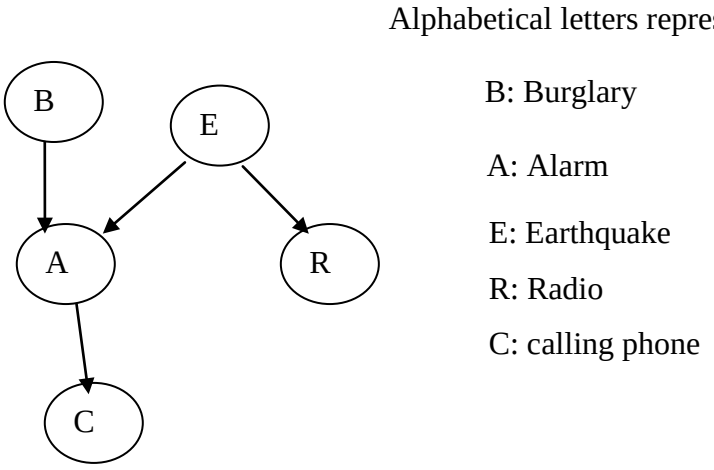


FIGURE 3.1: STRUCTURE OF A PROBABILISTIC NETWORK MODEL **(Pearl 1988**)

(ii) Axioms

Bayesian probability theory deals with events. If **a** is an event, then the probability of **a** is denoted by a real- valued number, **P (a).** While different people may give **P (a)** a different value there are nevertheless certain axioms which should always hold for internal consistency. These are the axioms of the probability theory (which can be *proved* to be valid when **P (a)** represents the frequentist (i.e., relative frequency of the events) approach) (Rahel, 2005; Heckerman, 1995; Bayes, 1763):

1. For any event, **a**, **$0 \leq$ P (a) ≤ 1**, with **P (a) = 1** if and only if **a** occurs with certainty.
2. For any two mutually exclusive events **a** and **b** the probability that either **a** or **b** occur is

P (a or b) ≡ P (a $\cup$ b) = P (a) + P (b).

In general, if events $a_1 \ldots a_n$ are pairwise incompatible, then

$$P(\bigcup_{i}^{n} a_i) = P(a_1) + ... + P(a_n) = \sum_{i}^{n} P(a_i) = 1$$

3. For any two events a and b the probability that both a and b occur is

P (a and b) ≡ P (a, b) = P (b|a) P (a) = P (a|b) P (b),

P (a, b) is called the joint probability of the events **a** and **b.**

(iii) Prior and Conditional Probabilities

As explained in many literatures, for instance, (Rahel, 2005; Bayes, 1763), the unconditional or prior probability associated with a proposition A is the degree of belief accorded to it in the absence of any other information. It is written as P (A). For instance, if the prior probability that any tree/shrub species has high threat in conservation status is 0.5, then we can write "P (threats=high) =0.5". If other information is known other than the prior, then the probability of A becomes conditional given this new information.

For instance, if both regeneration status (RS) and important value index (IVI) are high, then what is the probability of threat of the species in a certain random areas? In such cases, information is known other than the prior, then, P (threat of species|RS=HIGH, IVI=HIGH).

The basic concept in the Bayesian treatment of uncertainty is that of conditional probability: Given event b, the conditional probability of event **a** is **x**, written as

P (a|b) = x.

This means that if **b** is true and everything else known is irrelevant for **a**, then the probability of **a** is **x**.

In figure 3.1(see page 20) Assume that the alarm sounds in eight of every ten cases when there is an earthquake but no burglary. This fact would then be expressed as the conditional probability P (A = yes |B =no, E = yes)=0.8.

The formula of conditional probabilities can be defined as follows:

$$P(A \mid B) = \frac{P(AnB)}{P(B)} \text{ , which holds whenever P (B) > 0;} \quad (3.1)$$

Or

$$P(B \mid A) = \frac{P(AnB)}{P(A)} \text{ , which holds whenever P (A) > 0} \quad (3.2)$$

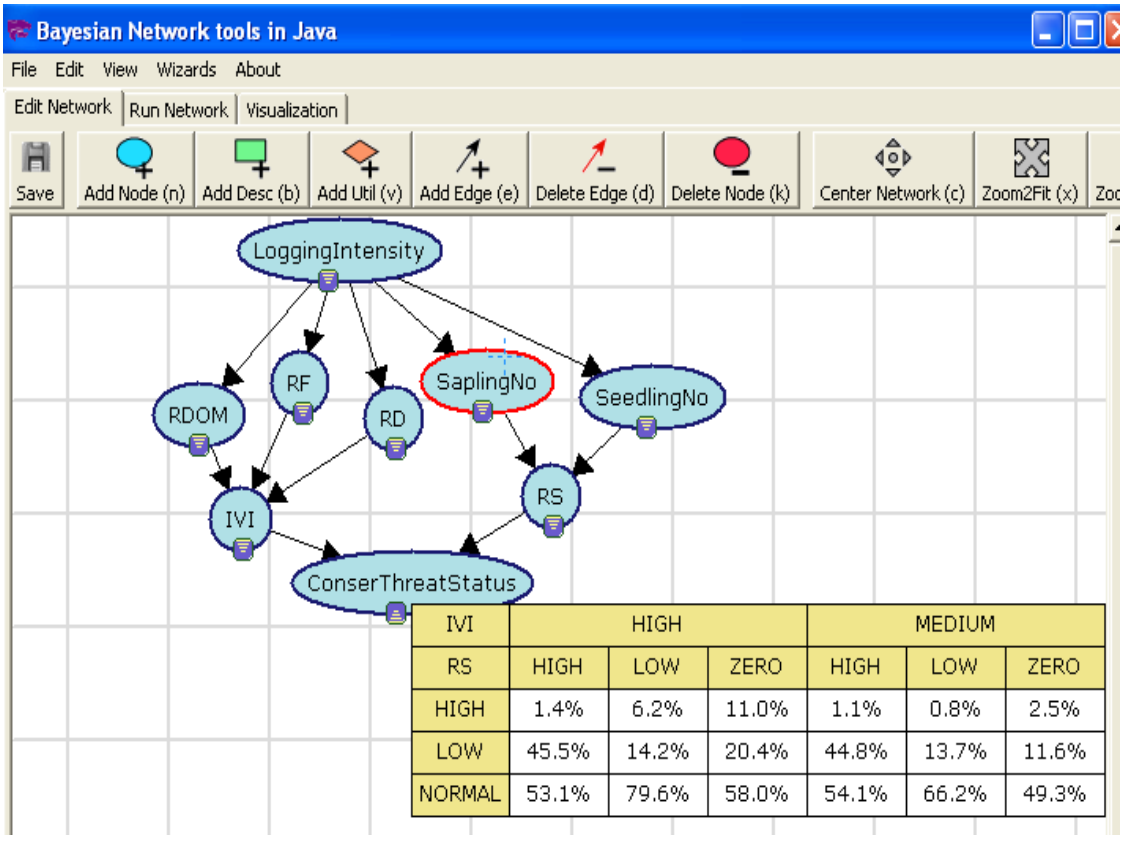
We can also get the **product rule** from equation (3.1) and (3.2) as stated below:

$$P(A|B) = P(A|B) * P(B) = P(B|A) * P(A) \quad (3.3)$$

Similarly, we may define the joint distribution P (A, B, C) as follows:

$$P(A, B, C) = P(A|B, C) * P(B|C) * P(C) \quad (3.4)$$

The following figure shows the screen print of the CPT table that has been computed after learning parameters for each attributes, and it is visualized with *BNJ Tool as follows:*



IVI	HIGH			MEDIUM		
RS	HIGH	LOW	ZERO	HIGH	LOW	ZERO
HIGH	1.4%	6.2%	11.0%	1.1%	0.8%	2.5%
LOW	45.5%	14.2%	20.4%	44.8%	13.7%	11.6%
NORMAL	53.1%	79.6%	58.0%	54.1%	66.2%	49.3%

FIGURE 3.2: DEMONISTRATION OF PRIOR AND CONDTIONAL PROBABILITIES USING BNJ

From the figure 3.2, the unconditional probability of "ConservThreatStatus' when it is HIGH, can be computed as, P (ConservThreatStatus=HIGH)

=1.4%+6.2%+11%+1.1%+0.8%+2.5%=0.23., that is, adding the first row gives the unconditional (or summing out- because the variables other than high conservation status are summed out) probability of high conservation threat status.

Similarly, the conditional probability of P (ConservThreatStatus=HIGH|IVI=HIGH, RS=HIGH) can be computed as:

$$= \frac{P(ConservThreatStatus = HIGH \wedge IVI = HIGH \wedge RS = HIGH)}{P(IVI = HIGH, RS = HIGH)} = \frac{0.014}{0.014 + 0.455 + 0.531} = 0.014.$$

Hence, for each variable, we can specify a table of conditional probability distributions, one for each configuration of states given its parents. Figure 3.2 shows these tables of conditional distributions, such as *P (LoggingIntensity), P (RS|SeedlingNo, SaplingNo), P (IVI|RD, RF, RDOM), and P (ConservThreatStatus|IVI, RS).*

(iv) Bayes' Theorem

As explained in (Bayes, 1763), true Bayesians actually consider conditional probabilities as more basic than joint probabilities It is easy to define P (A|B) without reference to the joint probability P (A, B). To see this note that we can rearrange the conditional probability formula to get:

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

but by symmetry we can also get:

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

It follows that:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$, **Bayes' Rule**. (3.5)

Bayes' rule refers to updating our belief about a hypothesis A in the light of new evidence B. That means, our posterior belief P (A|B) is calculated by multiplying our prior belief P (A) by the likelihood P (B|A) that B will occur if A is true.

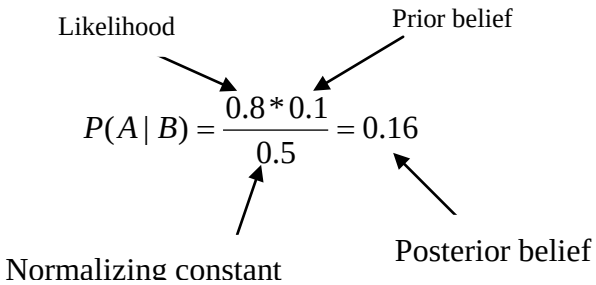
The power of Bayes' rule is that in many situations where we want to compute P (A|B) it turns out that it is difficult to do so directly, yet we might have direct information about P (B|A). Bayes' rule enables us to compute P (A|B) in terms of P (B|A).

For example, suppose that we are interested in diagnosing cancer in patients who visit a chest clinic.

- Let A represent the event "Person has cancer"
- Let B represent the event "Person is a smoker"

If we have prior information, P (A) =0.1 on the basis of past data (10% of patients entering the clinic turn out to have cancer). We want to compute the probability of the posterior event P (A|B). It is difficult to find this out directly. However, we are likely to know P (B) by considering the percentage of patients who smoke – suppose P (B) =0.5. We are also likely to know P (B|A) by checking from our record the proportion of smokers among those diagnosed. Suppose P (B|A) =0.8.

Using Bayes' rule, we can compute as follows:



Thus, in the light of evidence that the person is a smoker we revise our prior probability from 0.1 to a posterior probability of 0.16. This is a significance increase, but it is still unlikely that the person has cancer.

The denominator P (B) in the equation (3.5) is a normalizing constant which can be computed, for example, by marginalization whereby

$$P(B) = \sum_i P(B, A_i) = \sum_i P(B|A_i) \cdot P(A_i)$$

Hence we can state Bayes rule in another way as:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{\sum_i P(B|A_i) \cdot P(A_i)}$$

In any problem domain, there is a list of random variables to represent the knowledge domain needed. For instance, the list of random variables is {A1, A2,…, An} and that **parents** (A_i) denotes the set of parents of the node A_i in the BBN. Then, like conditional probability distributions (or conditional probability table), the joint probability distribution for {A1, A2… An} can be defined by chain rule as follows:

$$P(A_1, A_2, ..., A_n) = P(A_1)P(A_2 / A_1)....P(A_{n-1} / A_n)P(A_n)$$

$$P(A_1, \ldots, A_n) = \prod_{i=1}^{n} P(A_i | \text{parents}(A_i)) \qquad (3.6)$$

As a result, this implies that the joint probability P (A_1, …, A_n) can be represented as the conditional probabilities P (A_i | parents (A_i)). Problem domain usually allows to identify a

subset parents (A_i) μ $\{A_1, \ldots, A_{i-1}\}$ such that given parents (A_i), A_i is independent of all variables in $\{A_1, \ldots, A_{i-1}\} \setminus$ parents (A_i).

(v) Independence and Conditional Independence

As stated in many Bayesian references, for instance (Cheng et al., 1998), the conditional independence relationships encoded in the Bayesian network state that a node X_i is conditionally independent of its ancestors given its parents Π_i . Therefore,

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid \Pi_i)$$, where parents Π_i, Π_i is independent of all variables in $\{X_1, \ldots, X_{i-1}\} \setminus \Pi_i$.

In general, dependence between two events is when the probability of an event depends on the knowledge of the other event. Where as, in the absence of any dependency of one event with the other event, it is called independent probability.

The conditional probability of A given B is represented by P (A|B). The variables A and B are said to be independent if P (A) = P (A|B) (or alternatively if P (A, B) =P (A) P (B) because of the formula for conditional probability).

Once we know the joint probability distribution encoded in the network, we can answer all possible inference questions about the variables using marginalization (see figure 3.3).
The following example as shown in figure 3.3 demonstrates how the conditional independence is used to infer or predict the marginal probability via chain rule.

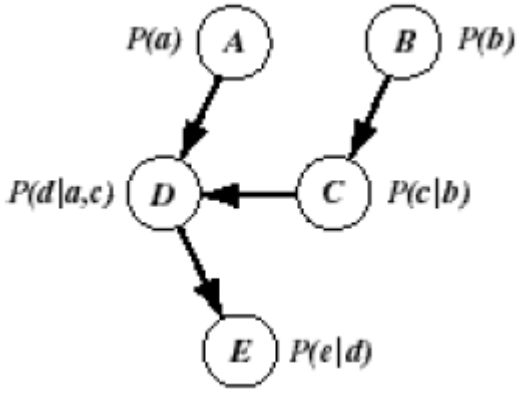


FIGURE 3.3: EXAMPLE OF CONDITIONAL INDEPENDENCE : P(a, b, c, d, e) =P(a)P(b)P(c|b)P(d|a, c)P(e|d)

From the figure 3.3, it can be realized that the conditional independence would make the ordering of the random variables based on the causal relationships of attributes (i.e., A, B, C, D and E) in the problem domain.

In any case, any three events A, B and C, then

P (A|C) = P (A|B, C), such that A *and B are conditionally independent given C.*

In many real life situations variables which are believed to be independent are actually only independent conditional on some other variable.

3.4 BAYESIAN NETWORKS

The Bayesian belief network is a powerful knowledge representation and reasoning tool under conditions of uncertainty. A Bayesian belief-network is a directed acyclic graph (DAG) with a conditional probability distribution for each node (Pearl, J., 1988; Neapolitan, R.E., 2004; Spirtes, P., et al. 2000). The DAG structure of such networks contains nodes representing domain variables, and arcs between nodes representing probabilistic dependencies. On constructing Bayesian networks from databases, we use nodes to represent database attributes. The figure that is shown below is example of DAG.

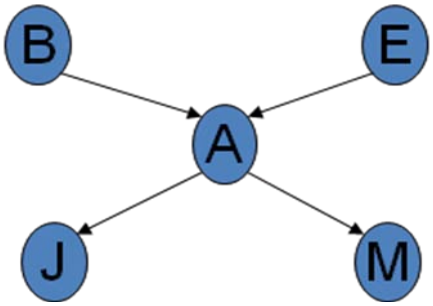


FIGURE 3.4: EXAMPLE FOR A BAYESIAN NETWORK

Some concepts and definitions related to Bayesian network are stated in (Cheng et al., 1998), and are briefly presented as follows:

Definition 1: A *directed graph* **G** denoted as (**V, E**), where ***V*** is a finite, non-empty set whose elements are called vertices or nodes, and **E** is a set of ordered pairs of distinct elements of ***V***. Elements of **E** are also called the edges or arcs. If **(x, y)** $\in$ **E**, we can say that there is a directed edge from **x** to **y** and that **x** and **y** are incident to the edge. It is denoted by an arrow from **x to y**, and we can say that **x** and **y** are incident to the edge.

x and **y** are said to be adjacent or neighbors if there is an edge from **x** to **y** or from **y** to **x**. If the start of the arrow is at **x** and the end point at **y**, then **x** is called the *parent* of **y** and **y** is called the *child* of **x**. Similarly **x** is called the ancestor of **y** and **y** the descendent of **x**. The set of edges connecting the nodes **x** and **y** is called the *path* from **x** to **y**.

A *directed cycle* is a path from a node to itself. A *simple path* is one with no subpaths which are directed cycles.

Definition 2: In Bayesian network learning, we often need to find a path that connects two nodes without considering the directionality of the edges on the path. To distinguish it from the directed path that connects two nodes by the arcs of a single direction, we call this kind of paths **adjacency paths** or **chains**.

Definition 3: A directed graph that contains no directed loops or cycles is called a *directed acyclic graph (DAG)*.

An example of a Directed Acyclic Graph (DAG) is shown in figure. 3.4.

Definition 4: Let *U* be a finite set of discrete value variables. Let *P(.)* be a joint probability function over the variables in U, and let X,Y,Z be any three subsets of variables in U. X and Y are said to be *conditionally independent* given Z if *P*(x/y,z)=P(x/z), if P(y,z)>0. Conditional independence means that knowledge of Z makes X and Y independent of each other.

3.5 LEARNING BAYESIAN NETWORK

In a Bayesian network, the graph (i.e., Directed Acyclic Graph (DAG)) is the structure of the Bayesian Network, and the conditional probability distribution is called the parameter. Both the parameter and the structure can be separately learned from the data. While learning the parameter, we assume that we know the structure of the DAG, but in structure learning we start with only a set of random variables with unknown relative frequency distribution. Learning structure can also be done with missing data items and hidden variables as discussed in (Neapolitan et al., 2004).

Similarly, Cheng et al (1998), learning a Bayesian network from data involves the tasks of structure learning, that is, identifying the graphical structure of the network, and parameter learning, that is, estimating the conditional probability distributions to be associated with the network's digraph.

Learning Bayesian network from data takes four different cases in terms of whether the structure of the network is known and whether the data is complete. These are:

- Unknown network structure and complete data
- Known network structure and complete data

- Unknown network structure and incomplete data
- Known network structure and incomplete data.

Learning with complete data indicates that the training data contains no missing values, while, learning with incomplete data indicates that some piece of information in the data are not known (Rahel, 2005; Cooper and Herskovits, 1992; Spirtes et al, 2000). Note that each case has its own learning algorithm.

There are two major methods for learning the structure of a network from complete data, namely:

(i) Dependency analysis methods; this suggests learning the BN structure by identifying the conditional independence relationships among the nodes. These algorithms are referred as *CI-based* algorithms or constraint-based algorithms (Spirtes et al., 2000; Cheng et al., 1998). The algorithm is known as Three-Phase-Dependency-Analysis (TPDA). The algorithm has three subsequent phases termed drafting, thickening and thinning. Explaining how each phases of the algorithm (i.e., mathematical details of each phase) working is essential, however, we skip it as the current study interests to see its application.

(ii) search and scoring method; this suggests that the best BN is the one that best fits the data, and leads to the *scoring based* learning algorithms, that seek a structure that maximizes the Bayesian, or entropy scoring function (Heckerman, 1995; Cooper et al., 1992)). Discussing more about this method may not give sense since the study does not use it for the experiments.

As explained in (Rahel, 2005; Cheng et al., 1998), if the network structure is already defined (i.e., if the structure is already known), the algorithm needs to estimate only the parameters (CPT) - using techniques such as Maximum Likelihood Estimation and Bayesian estimation;

but for unknown structure learning with complete data, the learning algorithm is given the set of variables in the model and needs to select the arcs between them to estimate the parameters.

According to Rahel (2005), unknown structure learning with complete data is helpful when we want to exploit all of the benefits of a Bayesian network model, to present the expert some suggestions of what attributes show a cause-effect relationship, and when situations don't allow us to get the domain expert.

Once we have a complete dataset, we can develop a Bayesian Network both in the case when a network structure is known and when it is not. There are some publicly known algorithms to develop Bayesian Network from complete data. One such algorithm is Three-Phase-Dependency-Analysis (TPDA). Hence, dependency analysis method to learn the Bayesian network structure from complete data has been adopted in the current study, too.

In dependency analysis method, identification of conditional independence relationships among the variables is done using information theory (like mutual information). One can find the conditional independence relationships among the attributes and use these relationships as constraints to construct a BN (Spirtes et al., 2000; Cheng et al., 1998).

If two nodes are dependent, then the knowledge of the value of one node will give us some information of the value of the other node. This information gain can be measured by using mutual information. Therefore, the knowledge of mutual information can tell us about the dependency relation between two nodes.

Discussions on unknown structure and partial data or known structure and partial data are not the focus of this study. Hence, in next sections, both unknown and known structures with complete data are briefly presented.

3.5.1 Unknown Network Structure and Complete Data

Bayesian learning algorithm with unknown structure takes a database table as input and constructs a Bayesian network structure as output. Since node ordering is not given as input, this algorithm has to deal with two major problems (Rahel, 2005):

(i) How to determine if two nodes are conditionally independent, and

(ii) How to orient the edges in a learned graph.

As explained in Cheng et al (1998), the algorithm has four phases: ***drafting*** (creating draft (arcs) based on mutual information of each pair of nodes in decreasing order), e.g.,

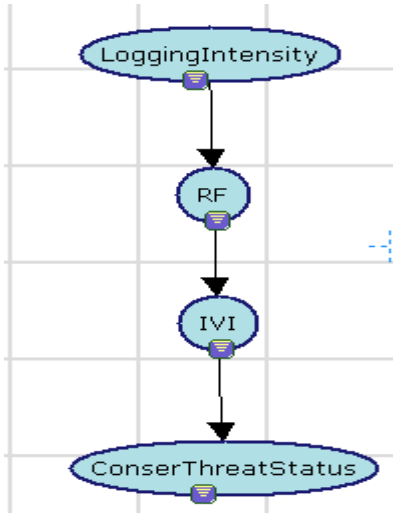


FIGURE 3.5: EXAMPLE OF DRAFTING

[Source: Adopted from the experimental results]

In figure 3.5, LoggingIntensity is highest in mutual information as compared to other nodes in the network. LoggingIntensity, RF, IVI and ConservThreatStatus are in dcreasing order of mutual information that is calculated by the program internally.

thickening (adding the missed arcs when a pair of nodes are not conditional independence)**,** ***e.g.,***

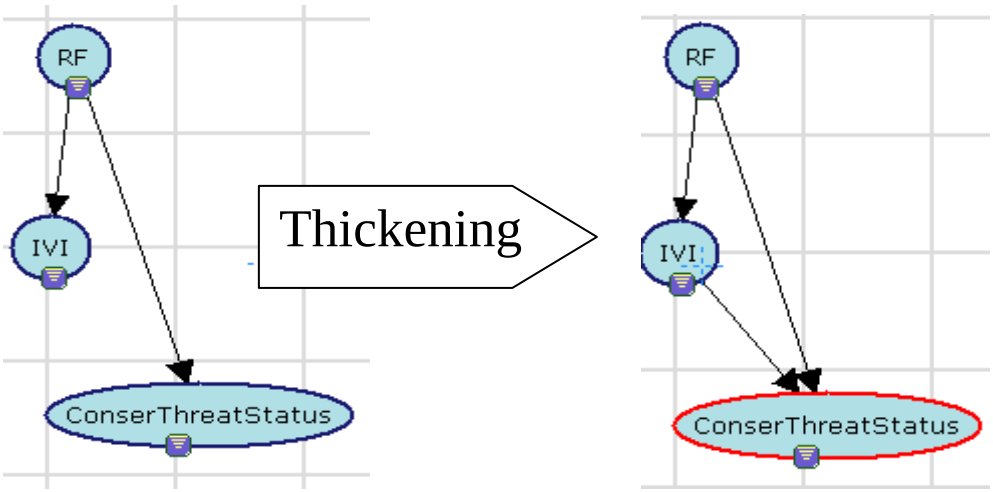


FIGURE 3.6*:* EXAMPLE OF THICKENING

[Source: Adopted from the experimental results]

In figure 3.6, IVI is not conditional independence, but it is directly influence the target variable, ConservThreatStatus, specifically: P (ConservThreatStatus|IVI). As a result, IVI should be added (i.e., **IVI** ⟶ **ConservThreatStatus**).

thinning (removing unwanted arcs when any two nodes are conditional independence), e.g.,

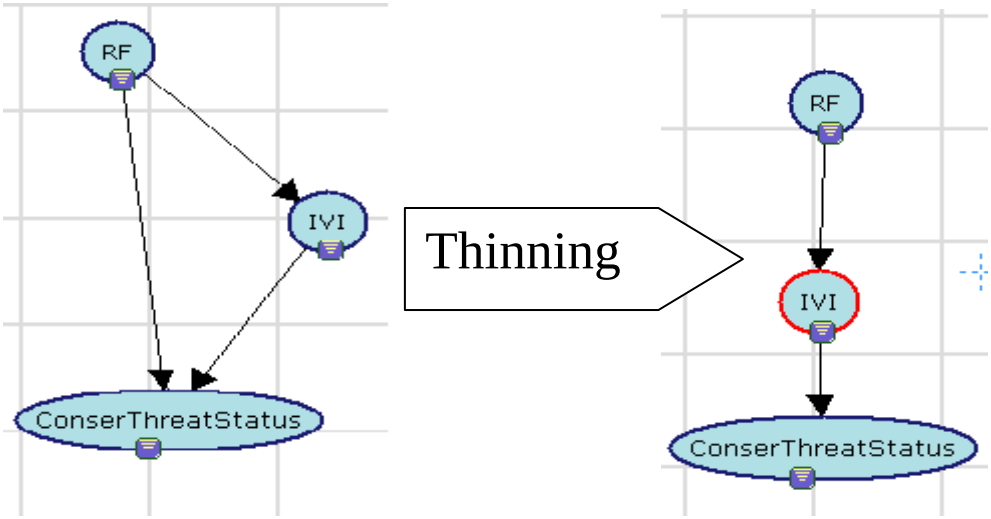


FIGURE 3.7: EXAMPLE OF THINNING

[Source: Adopted from the experimental results]

In figure 3.7, since RF and ConservThreatStatus are conditional independence given IVI in the entire network. Specifically: **P (ConservThreatStatus| RF, IVI) =P ConservThreatStatus| IVI).** Hence, **RF** ⟶ **ConservThreatStatus** is unwanted arc and it should be removed.

*and **orienting edges*** (adjusting misdirected arcs)., e.g.,

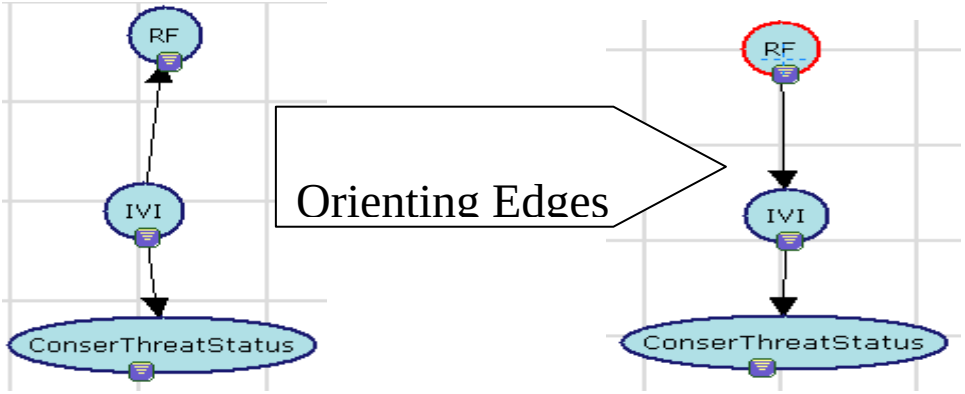


FIGURE 3.8: EXAMPLE OF EDGE ORIENTATION

[Source: Adopted from the experimental results]

In figure 3.8, since RF is directly influence IVI, that is, RF is the parent of IVI. Hence, the misdirected arc should be adjusted like, **RF ⟶ IVI.**

3.5.2 Known Network Structure and Complete Data

The algorithm takes as input both a table of database entries and a node ordering and constructs a Bayesian Network structure as output.

The first three phases of this algorithm are the same as the TPDA algorithm described in the previous section. However, the last phase (orienting edges) described above, is not implemented in this algorithm, since the direction of the arcs are decided by the node ordering provided. According to Rahel (2005), the main features involved in these three phases are the following:

(i) When direct causal effect relationships (i.e., dependency) between attributes are available, it uses them as a basis for generating a draft.

(ii) In thickening, the algorithm will try to add an arc only if it agrees with the domain knowledge.

(iii) In thinning, the algorithm will not try to remove an arc if it is already specified by

domain experts.

3.6 CAUSAL BAYESIAN NETWORK

As explained in (Heckerman et al., 2000), Bayesian networks use directed acyclic graphs to represent causalities. Nodes represent variables, and directed edges represent direct probabilistic dependences. For instance, the temporal order of the nodes could offer an intuitive(or normal) notion of the cause-effect relationships. In the next sections, basic issues related to causality and flows of information in causal network are presented.

3.6.1 Causality

Causality plays an important role in the process of constructing probabilistic network models. There are a number of reasons why proper modeling of causal relations is important or helpful, although it is not strictly necessary to have the directed links of a model follow a causal interpretation (Pearl, 1988).

According to (pearl 1988), a variable X is said to be a direct cause of Y if setting the value of X by force, the value of Y may change and there is no other variable Z that is a direct cause of Y such that X is a direct cause of Z.

To correctly represent the dependence and independence relations that exists among a set of variables of a problem domain it is very useful to have the causal relations among the variables be represented in terms of directed links from causes to effects. That is, if X is a direct cause of Y, we should make sure to add a directed link from X to Y. If it is done the other way around (i.e., Y $\rightarrow$X), we may end up with a model that does not properly represent the dependence and independence relations of the problem domain.

That said, however, one does not have to construct a model where the links can be interpreted as causal relations, it just makes the model much more intuitive, eases the process of getting

the dependence and independence relations right, and significantly eases the process of eliciting the conditional probabilities of the model (pearl 1988).

3.6.2 Flow of Information in Causal Networks

As mentioned above, the DAG of a probabilistic network model is a graphical representation of the dependence and independence properties of the joint probability distribution of the model. As explained in (Rahel, 2005), knowing the flow of information from the DAG is helpful to know the conditional independence as well as the causal relation ship of the attributes. In doing this, it is convenient to consider each possible basic kind of connection that can exist in a DAG (see figure 3.9).

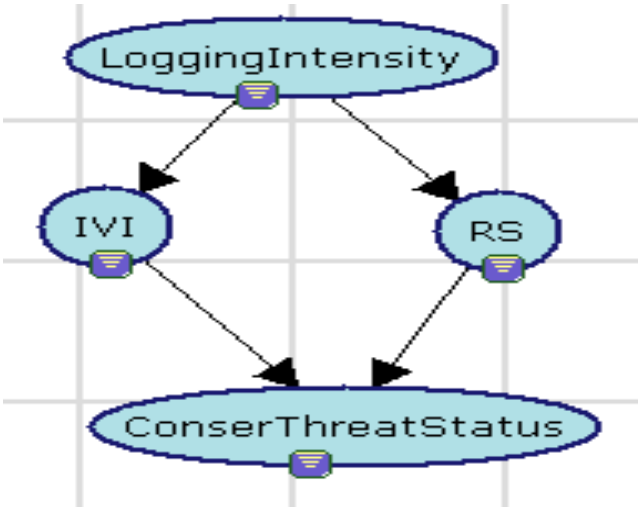


FIGURE 3.9: EXAMPLES OF CAUSAL NETWORKS
(Taken from the results of the experiment using BNJ)

For example, in figure 3.9, we see three different kinds of connections in the network:

- two serial connections

  LoggingIntensity→RS→ConservThreatStatus, and

  LoggingIntensity →IVI →ConservThreatStatus,

- one diverging connection

  IVI ←LoggingIntensity →RS, and

- one converging connection

IVI → ConservThreatStatus ←RS

We have discussed briefly each of these three possible kinds of connections as shown below:

(i) Serial Connections

Definition: (Serial connection) Information may be transmitted through a serial connection X →Y →Z unless the state of Y is known. E.g., LoggingIntensity →RS →ConservThreatStatus

Consider figure 3.9, a situation where LoggingIntensity directly causes RS and RS in turn directly causes ConservThreatStatus, i.e., RS is the direct causal effect of LoggingIntensity and ConservThreatStatus is also a direct causal effect of RS. Whereas as, the LoggingIntensity is independent of ConservThreatStatus. This implies, any knowledge that there is a LoggingIntensity is irrelevant to any hypothesis (or belief) that the ConservThreatStatus is either to be high or low. Or if we know the true state of RS, further knowledge of LoggingIntensity is irrelevant to ConservThreatStatus, i.e., ConservThreatStatus is conditionally independent of LoggingIntensity given RS.

In other words, the probability, say P,

P (LoggingIntensity|ConservThreatStatus)=P(LoggingIntensity), which shows LoggingIntensity and |ConservThreatStatus are independent probability..

Or

P (ConservThreatStatus| RS) =P (ConservThreatStatus| LoggingIntensity, RS), such that ConservThreatStatus and LoggingIntensity are conditional independence given RS.

(ii) Diverging Connections

Definition: (Diverging connection) Information may be transmitted through a diverging connection X ←Y →Z unless the state of Y is known. E.g. IVI ←LoggingIntensity →RS

For second example, consider a situation where LoggingIntensity directly causes RS and LoggingIntensity also directly causes IVI. Although knowledge of RS is relevant to IVI (if

RS is true then it is more likely that IVI is true which in turn means that it is more likely that RS is true), once we know the true state of LoggingIntensity, then further knowledge of RS is irrelevant to IVI, i.e., RS is conditionally independent of IVI given LoggingIntensity .

(iii) Converging Connections

Definition: (Converging connection) Information may only be transmitted through a converging connection X → Y ←Z if evidence on Y or one of its descendants is available.

E.g., IVI → ConservThreatStatus ←RS

Finally as a third example, consider the situation where IVI and RS are two independent direct causes of ConservThreatStatus, i.e., IVI and RS are independent. But if we learn something about the true state of ConservThreatStatus, then IVI and RS are no longer irrelevant to each other (if ConservThreatStatus is believed to be true and IVI is false, then it is more likely that RS is true), i.e., RS is not conditionally independent of IVI given ConservThreatStatus.

The three connections explained above show up all the forms in which evidence may be transmitted though a variable. It is observed that one can decide for any pair of variables in a causal network whether or not they are dependent once knowing the evidence entered into the network.

3.7 APPLICATION OF BAYESIAN NETWORK

3.7.1 General Applications

Bayesian networks have had considerable applications in many fields both in academia and industry. The major application area in both fields has been diagnosis, which lends itself very naturally to the modeling techniques of Bayesian networks. In the academic field, Nikovski (2000) applied it to problems in medical diagnosis, Hansson and Mayer (1989) in heuristic

search, Ames et al. (2003) in watershed management, Cain et al., (1999) in natural resource management, Breese and Heckerman (1999) in intelligent trouble shooting systems, and in "Computer-Assisted Learner Group Formation Based on Personality Traits", Rahel (2005) aimed at investigating the use of Bayesian networks for predicting performance of a student (i.e., 79.85% accuracy).

Industrial application of Bayesian technology spans several fields including medical and mechanical diagnosis, risk and reliability assessment, and financial risk management. An example of medical diagnosis is the Heart Disease Program developed by the MIT laboratory for Computer Science and Artificial Intelligence. This program assists physicians in the task of differential therapy in the domain of cardiovascular disorders (Long, 1989).

3.7.2 Related Works

Literature on the use of Bayesian Belief Networks in biological modeling for target tree species is limited. Even though similar works in the domain area are not sufficient, adopting from ecological modeling for the current study is also applicable as biological and ecological concepts are overlapped in many situations. For instance, in "Guidelines for developing and updating Bayesian belief networks for Ecological modeling and conservation ….", Marcot et al. (2006) underlined that Bayesian belief networks (BBNs) are statistical tools used in ecology to depict the influence of habitat or environmental predictor variables (or causal variables) on or ecological[7] -response variable (or target variable). In such cases, BBNs predict the probability of ecological response to varying input assumptions such as habitat and population-demography conditions. BBNs serve well as part of a risk-management framework by explicitly displaying the "causal web" of interacting factors and the probabilities of multiple states of predictor and response variables. The findings, in this paper, indicated that

[7] The word Ecological can be used interchangeably with Biological (Marcot et al (2006)).

there was a prediction accuracy of 61% cases based on actual or observed data for habitat suitability associated with species characteristics.

According to McNay et al. (2006), BBNs can be used to represent the primary relationships between biological variables in predictive simulation models, and can be used as a 'post-processing' step to summarize results from the model projections, allowing nonmodelers to explore and interpret results.

As explained in (Olson et al. , 1990a; Marcot et al., 2006;), biological conservationist may also want to use a tool, such as BBNs, that can represent uncertainty in terms of probabilities of different potential outcomes or system responses, given initial conditions and human activities. Because of their probabilistic basis and their ability to explicitly represent and quantify the expected utility of alternative conservation decisions and strategies, BBNs lend themselves well to representing uncertainty of understanding and their implications to possible conservation decisions (Kuikka et al. 1999). Hence, Bayesian approaches can be used to assess the relative plausibility of parameter values and hypothesis and weight them according through explicit consideration of uncertain or subjective information, and lead to a systematic approach to sensitivity analysis.

BBNs also can contribute indirectly to sound decision-making by representing probabilities of biological (or ecological) responses to natural events and conservation actions within larger decision-support frameworks. For example, dynamic landscape models can be used to generate inputs to BBNs that, in turn, predict outcomes of alternative simulations in meaningful ways that can aid a resource decisions process (McNay et al. 2006).

According to McCann et al (2006), BBN can be viewed as useful and limitations with respect to the issues that are presented as follows:

Description of issue	Useful Characteristics of BBNs	Limitations about using BBNs
Complexity : requires a multidisciplinary approach to account for-multiple interactions among plants, socioeconomic and cultural considerations	Flexible use of information, rapid and flexible modeling environment.	BBNs require a fully specified probabilistic model and often require elicitations of expert judgment; nodes in the models should be empirically observable, quantifiable, or defensible; feedback functions and temporal relationships are not possible or are poorly handled. Continuous variable must be discretized. Etc.
Uncertainty: Information is often insufficient to adequately define and predict ecosystem characteristics.	Predicted on an established normative theorem that can explicitly represent uncertainty; can provide support for development of field experiments to reduce uncertainties in risk analysis.	Usually do not explicitly represent bias or error. Models can be easily developed, entirely from expert judgment, with an unknown degree of bias and inaccuracy.
Decision-Making: Decisions must be made in the face of complexity and uncertainty.	Permits identifications of factors, or interactions between factors, that are most influential on model outcomes.	Decision makers must not assume that all relevant uncertainties regarding knowledge or objectives have been incorporated into the decision rules.

Table 3.1: Useful characteristics of BBN and limitations about their application in Biological Modeling.

The next chapter discusses the experimentatal result, which is the foremost element of the current work.

CHAPTER FOUR

MODELING AND EXPERIMENTAL RESULTS

4.1 MODELING AND PREPARATION OF DATA FOR THE EXPERIMENTS

In preparing data for the experiments, the following major steps are to be followed for further discussions and analysis of the study. These are:

(i) Flow of Information in BN Modeling Process

(ii) Data collection and Understanding

(iii)Transforming Data with Query By Example(QBE)

(iv) Data Preprocessing

4.1.1 Flow of Information in BN Modeling Process

BN modeling process shows us how to acquire the knowledge from the beginning to the final BN model of the application domain that we need to achieve. In this study, therefore, the major flow of information in the modeling process was presented diagrammatically in figure 4.1:

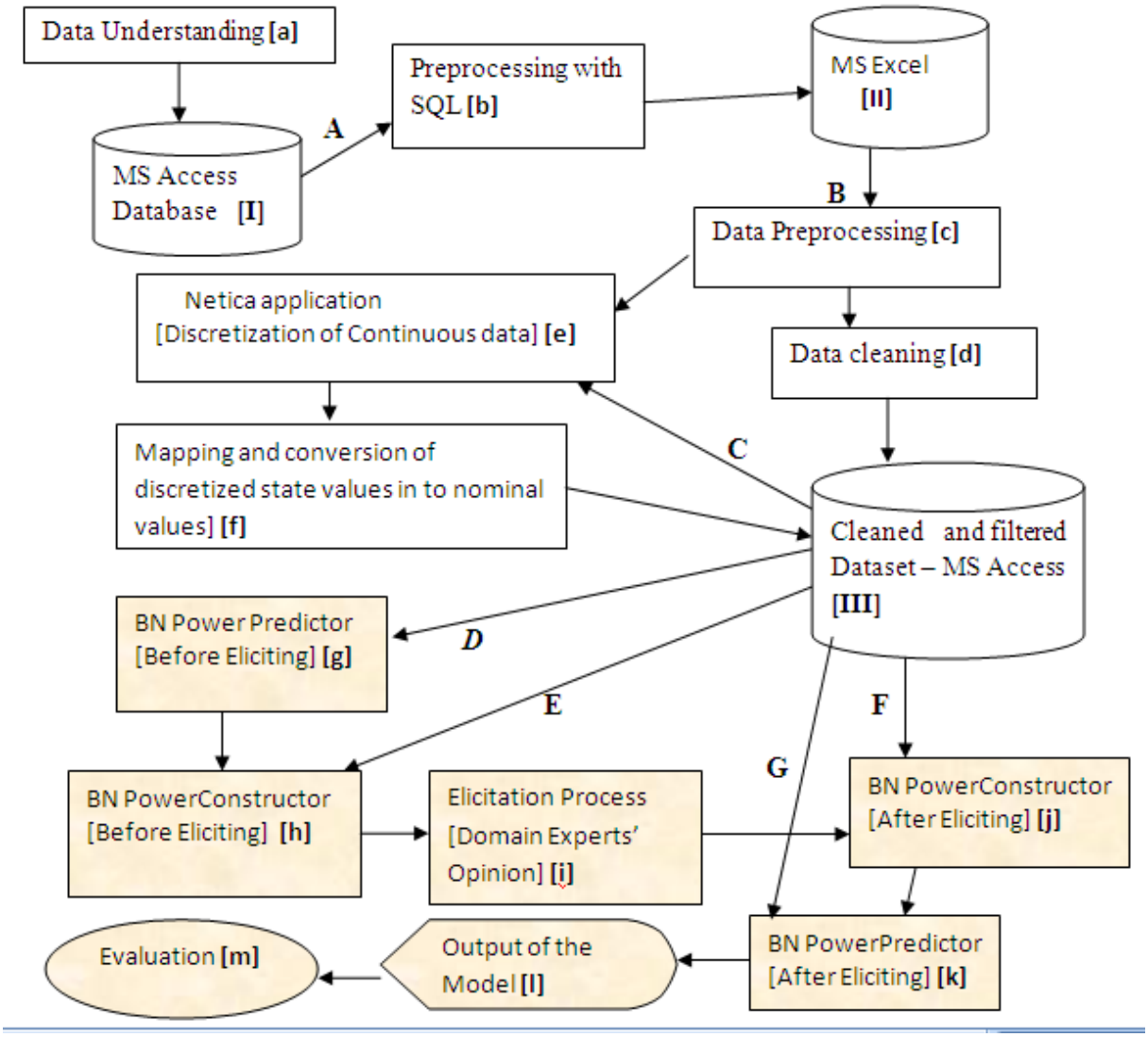


FIGURE 4.1: BN MODELLING PROCESS (It's not a standard)

In the figure 4.1, the unshaded parts depict tasks in preparing data for experiments, where as the shaded parts depict tasks in building the BN model or experimental parts. Others like**, [a], [b]… [m]** are sequence of tasks in BN modeling process, **[I], [III], & [III]** are the stored information at rest and **A, B, …, G** are the flow of information.

4.1.2 Data Collection and Understanding

The most important task that requires greater attention is collecting, analyzing and understanding the content and structure of data available. For the purpose of this study, the source of data about biological attributes has been obtained from the Institute of Biodiversity Conservation's database.

The data has usually been used for research purpose and it has also been updated frequently in MS Access database. The database has 25 tables, and accessing the data from a single table is impossible since they are organized in relational database form. Therefore, understanding about each of the tables is so difficult. So as to make good background knowledge of the database, the preliminary informal discussions and interviews have been conducted with the domain experts.

After proofing that, the actual data were found in the database that would be used for the experiment, selection of appropriate tables was done. Thus, five target tables were identified for preparing data for experiments. These are: IVI, Disturbance, VEG-Sap_Seed, Forest, and Species. The tables are presented in figure 4.2:

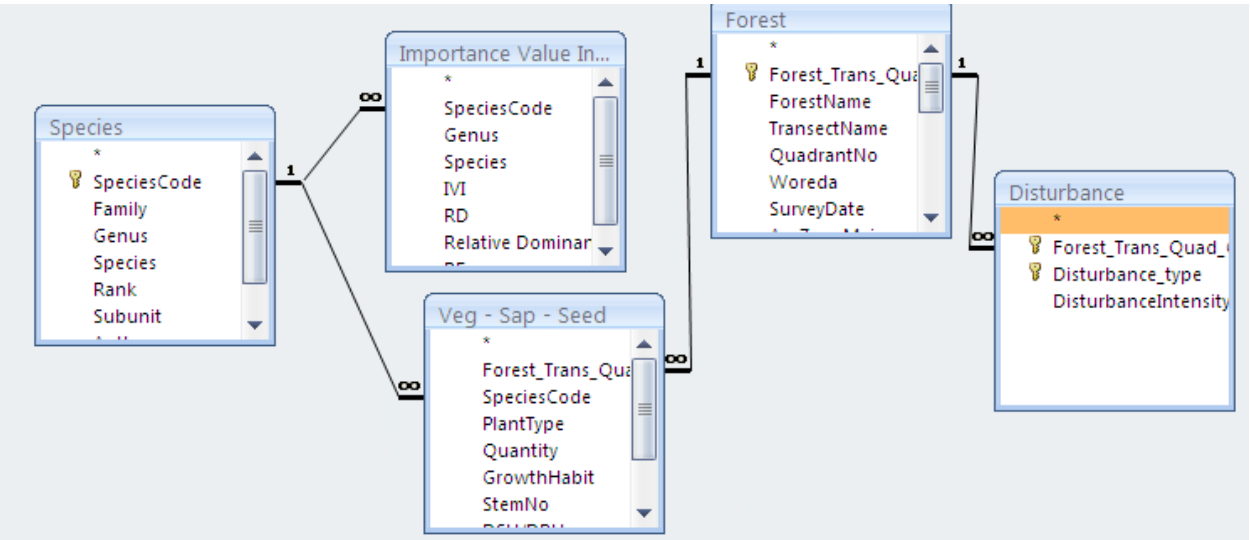


FIGURE 4.2: THE FIVE RELATIONAL TABLES

4.1.3 Transforming Data with Query by Example(QBE)

QBE is important to filter out and transform the major attributes with its corresponding valid values by formulating the query in the MS Access grid form.

To get the expected major biological attributes, the following simple query was formulated from the five tables as:

Query: ***Species*** **(SpeciesCode, Genus, Species),** ***Disturbance*** **(LoggingIntensity),** ***Forest*** **(ForestName),** ***ImportantValueIndex*** **(IVI, RF, RD, RDOM)** and ***Veg_Sap_Seed*** **(RS, SeedlingNo, SaplingNo).**

After formulating the query as stated above on the QBE, we generated the output as shown below in the table 4.1:

Masterdataset

Genus	Species	LoggingInter	IVI	RD	RDOM	RF	RS	SeedlingNo
Acacia	abyssinica	high	0.28	0.02	0.20	0.06	0	0
Acacia	abyssinica	high	7.15	0.44	5.02	5.19	0	0
Acacia	abyssinica	high	3.54	0.46	2.16	0.68	0	0
Acacia	abyssinica	medium	0.30	0.03	0.02	0.24	0	0
Acacia	abyssinica	medium	27.78	5.00	17.40	5.39	0	0
Acacia	abyssinica	medium	1.04	0.08	0.47	0.49	0	0
Acalypha	acrogyna	medium	1.04	0.23	0.01	0.24	6	0
Acalypha	acrogyna	medium	0.00	0.00	0.00	0.00	6	0
Albizia	grandibracteat	low	0.57	0.13	0.01	0.43	0	0
Albizia	grandibracteat	medium	0.14	0.01	0.00	0.13	0	0
Albizia	grandibracteat	high	3.49	0.68	1.82	0.99	0	0
Albizia	grandibracteat	medium	1.69	0.09	0.54	0.73	0	0
Albizia	grandibracteat	medium	0.31	0.06	0.01	5.93	0	0
Albizia	grandibracteat	high	2.04	0.80	0.45	0.48	0	0
Albizia	grandibracteat	high	10.50	2.26	4.66	6.16	0	0
Albizia	grandibracteat	high	1.42	0.29	0.15	0.99	37	0
Albizia	grandibracteat	high	1.46	0.67	0.13	0.66	50	50
Albizia	gummifera	medium	1.72	0.75	0.23	0.74	0	0
Albizia	gummifera	high	2.64	0.20	1.02	0.24	0	0
Albizia	gummifera	high	15.91	2.25	9.30	0.74	0	0
Albizia	gummifera	high	0.63	0.07	0.16	0.48	0	0
Albizia	gummifera	low	13.52	1.56	8.65	0.68	0	0
Albizia	gummifera	medium	2.46	0.16	1.15	1.15	0	0

TABLE 4.1: SAMPLE GENERATED OUTPUT FROM THE GIVEN QUERY.

4.1.4 Data Preprocessing

Next to data collection, data pre-processing is a step where the data to be used in finding

useful data is cleaned, reduced and organized, to a suitable format for Bayesian network tool. In the real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically large size or other unforeseen errors. Thus, while using the experiments, the data should be preprocessed in order to improve the quality of the data and to make it free from errors. Being cognizant of this fact, data cleaning and discretization of continuous data were the techniques to be used for data preprocessing for this study. Therefore, each of the techniques was illustrated as follows.

4.1.4.1 Data Cleaning and Organizing

As explained in (Ham and Kamber (2001)), a large number of errors are to be expected with original data set. This implies that usually, the raw data to be used for Bayesian network are not cleaned and organized, they have some errors, and irrelevant attributes which are not necessary for the goal of a biological modeling at hand. Data may be missed due to equipment problems, deletion of related records, transcription error while keying the manual record in an electronic format. Thus, data cleaning is just increasing the quality of the data to achieve the validity estimation of the model result.

According to Ham and Kamber (2001), data cleaning may involve selection of cleaned subset of the data, ignore the tuple, filling the missing value manually, and use the most probable value in the missing value. For instance, in real situations, as far as we have given values for the attributes such as IVI, RD, and RDOM, but values for RF has been missed during editing or entering the data in to the database, then we can obtain the values for the RF since RF is one of the variables to calculate the value for IVI. That is, IVI=RF + RD + RDOM. There fore, in such minor rescannable cases, it is possible to correct the values of the attributes. Thus, it is required to avoid mistakes, data incompleteness and noisy data, that may occur due to random errors, incorrect attributes, and other problems like duplicate records, incomplete data or data inconsistency.

In preparation of the data for experiment, data cleaning was conducted on MS Excel 2007. After data cleaning, we got 1200 (20%) records out of 6000 records (see table 4.2):

LoggingIntensity	IVI	RD	RDOM	RF	RS	SaplingNo	SeedlingNo	ConserThre
HIGH	0	0	0.2	0	0	0	0	HIGH
HIGH	3	0.44	1	1	0	0	0	Normal
HIGH	3	0.46	1	0.68	0	0	0	Normal
HIGH	0	0	0.02	0	0	0	0	HIGH
HIGH	3	1	1	1	0	0	0	Normal
HIGH	1.04	0.08	0.47	0.49	0	0	0	HIGH
MEDIUM	1.04	0.23	0	0	50	0	50	LOW
MEDIUM	0	0	0	0	50	0	50	LOW
HIGH	0	0.13	0	0.43	0	0	0	HIGH
HIGH	0	0	0	0	0	0	0	HIGH
HIGH	3	0.68	1	0.99	0	0	0	Normal
HIGH	1.69	0.09	0.54	0.73	0	0	0	Normal
HIGH	0	0	0	1	0	0	0	Normal
MEDIUM	2.04	0.8	0.45	0.48	0	0	0	Normal
MEDIUM	3	1	1	1	0	0	0	Normal
MEDIUM	1.42	0.29	0.15	0.99	50	0	50	Normal
MEDIUM	1.46	0.67	0.13	0.66	50	50	0	Normal
MEDIUM	1.72	0.75	0.23	0.74	0	0	0	Normal
MEDIUM	2.64	0.2	1	0	0	0	0	Normal
MEDIUM	3	1	1	0.74	0	0	0	Normal
MEDIUM	0.63	0	0.16	0.48	0	0	0	HIGH
MEDIUM	3	1	1	0.68	0	0	0	Normal
MEDIUM	2.46	0.16	1	1	0	0	0	Normal
LOW	1.97	0.57	0.32	1	0	0	0	Normal

TABLE 4.2: SAMPLE OF THE GENERATED OUTPUT AFTER DATA CLEANING

4.1.4.2 Discretization of Continuous Attribute Values

In order to use the Bayesian Belief Network technique, continuous variables need to be discretized (categorized). The discretization or categorization of a continuous variable is the process by which a continuous variable is converted into a discrete or categorical variable by grouping values into two or more categories as discussed in (Sando, 2005; Ham and Kamber, 2001).

Before we predicted or constructed the model with belief network software, detecting and discretizing data fields that contained continuous data were conducted. The training dataset contained seven continuous data and two discrete data for the application domain. For the purpose of discretization, we have used the Netica Application (hint: one of the Bayesian

belief network tool) since it is the recommended software for Bayesian application as explained in (Spiegelhalter et al., 1993; Lauritzen et al., 1988; Marcot et al., 2006)).

The state level (or threshold) of discretization may be decided depending on the problems situation. In some cases, we may use natural breaks for balancing of the state level by groupings (for instance, temperature could be divided into ($<0^0c$), (0^0c-100^0c) , and ($>100^0c$)) as explained in (Sando, 2005), and on the other hand, Lauritzen et al. (1988) and Marcot et al. (2006) argued that it has to be done using BN tools when attributes are found not possible for natural breaks and normally distributed (see APPENDIX V). They also justified that the BN tools like Netica to be used for this study is incorporated with plausible autodiscretizing method.

Netica software version 4.08 was used for discretization with three state levels and 0.02 (20%) rounding thresholds. For instance, if IVI value is 2.769, the software rounds it to 0.554.

However, in order to be benefited from both the software and the opinion of domain experts, the researcher has decided to combine the knowledge of two major tasks in discretizing the continuous data. These are presented as follows:

(i) First, Applying Netica Application Software to Discretize the Attribute Values (see figure 4.3, 4.4 and APPENDIX VIII).

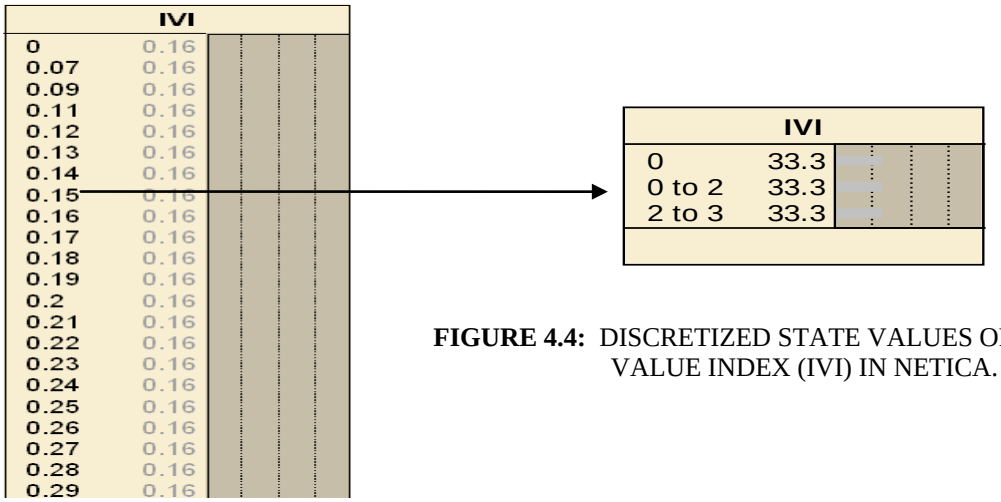


FIGURE 4.4: DISCRETIZED STATE VALUES OF IMPORTANT VALUE INDEX (IVI) IN NETICA.

FIGURE 4.3: SAMPLE VIEW OF IVI ATTRIBUTE BEFORE DISCRETIZATION IN NETICA

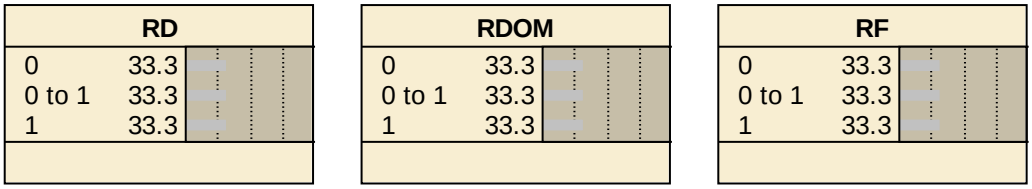


FIGURE 4.5: DISCRETIZED STATE VALUES OF RD, RDOM AND RF IN NETICA

(ii) Domain Experts' Opinion

The domain experts had almost relatively similar[8] experiences with the discretized state values that were done with Netica application (see figure 4.5). Because of this, they were given approval or acceptance for the state values that were discretized with Netica software.

After discretization, it was realized that further tasks should be done to make the discretized values more readable and responsive during interpretations. The three major subsequent tasks were further conducted and presented as follows:

(iii) Mapping of the Discretized State Values into Plausible Range of Real Values

[8] The implication of similar does not refer to the meaning of identical or the same (Source: Opinions of domain experts).

For each discretized attributes values, mapping into categorical values(nominal) were done for the sake of making states more readability and understanding of the experimental results in its interpretations, otherwise it is not obligatory to convert it in to nominal state values as explained in (Marcot et al., 2006). To do this, first, mapping into range of real values (see figure 4.6 and 4.7) and then assigning them into nominal values (see table 4.3) were done respectively.

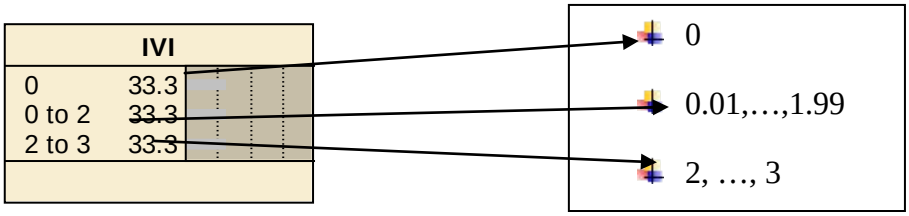


FIGURE 4.6: MAPPING OF IVI VALUES INTO RANGE OF REAL VALUES.

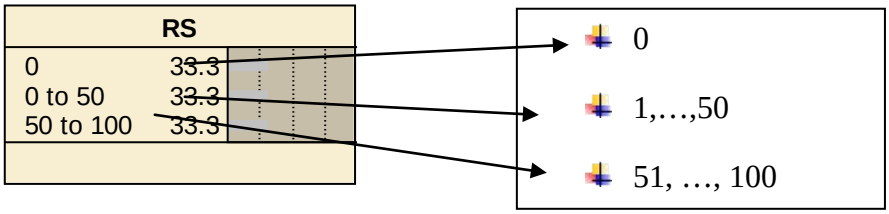


FIGURE 4.7: MAPPING OF RS STATE VALUES INTO RANGE OF REAL VALUES.

(iv) Assigning the Range of Real Values into Nominal (categorical values)

Experiences in the application domain have been indicated that each range of real values can usually be assigned or quantified with equivalent nominal values (Taye et al., 2002). Accordingly, we could able to depict nominal values for each attributes in the following manner (see table 4.3):

Attribute Name	Discretized Real Values	Nominal Values	Descriptions of Nominal Values at this particular instance
IVI	**0** **0.01,…,1.99** **2, ..,3**	**ZERO** **MEDIUM** **HIGH**	▪ ***Species require conservation, threat is high. More priority*** ▪ ***Species relatively threatened, but need attention, second priority*** ▪ ***Species relatively not threatened little or no conservation. Third Priority.***
RS	**0** **1,…., 50** **51, …,100**	**ZERO** **LOW** **HIGH**	▪ ***More threat expected. May be for rare species. First Priority.*** ▪ ***Relatively threatened species, Second Priority where they fall.*** ▪ ***Relatively no threats; third priority***
RDOM	**0** **0.01,…,0.99** **1**	**ZERO** **MEDIUM** **HIGH**	▪ ***Species require conservation, i.e., threat is high.*** ▪ ***Species relatively threatened, but need attention*** ▪ ***Species relatively not threatened, little or no conservation***
RD	**0** **0.01,…,0.99** **1**	**ZERO** **MEDIUM** **HIGH**	▪ ***Species require conservation, threat is high.*** ▪ ***Species relatively threatened, but need attention*** ▪ ***Species relatively not threatened, little or no conservation***
RF	**0** **0.01,…,0.99** **1**	**ZERO** **MEDIUM** **HIGH**	▪ ***Species require conservation, threat is high.*** ▪ ***Species relatively threatened, but need attention*** ▪ ***Species relatively not threatened, little or no conservation***
SaplinNO	**0** **1,…., 50** **51,…,100**	**ZERO** **LOW** **HIGH**	▪ ***More threat expected. May be for rare species..*** ▪ ***Relatively threatened species*** ▪ ***Relatively no threats***
SeedlingNo	**0** **1,…., 50** **51,…,100**	**ZERO** **LOW** **HIGH**	▪ ***More threat expected. May be for rare species*** ▪ ***Relatively threatened species*** ▪ ***Relatively no threats***

TABLE 4.3: MAPPING OF THE DISCRETIZED STATE VALUES INTO CATEGORICAL VALUES: Note: We tried to describe only based on the given (specific) nominal values for each attributes. Names of Nominal values are taken from the domain experts.

(v) Sample Records after Populating the Nominal Values into MS Access (see table below).

LoggingInter	IVI	RD	RDOM	RF	RS	SeedlingNo	SaplingNo	ConserThre
HIGH	ZERO	ZERO	MEDIUM	ZERO	ZERO	ZERO	ZERO	HIGH
HIGH	HIGH	MEDIUM	HIGH	HIGH	ZERO	ZERO	ZERO	Normal
HIGH	HIGH	MEDIUM	HIGH	MEDIUM	ZERO	ZERO	ZERO	Normal
HIGH	ZERO	ZERO	MEDIUM	ZERO	ZERO	ZERO	ZERO	HIGH
HIGH	HIGH	HIGH	HIGH	HIGH	ZERO	ZERO	ZERO	Normal
HIGH	MEDIUM	ZERO	MEDIUM	MEDIUM	ZERO	ZERO	ZERO	HIGH
MEDIUM	MEDIUM	MEDIUM	ZERO	ZERO	HIGH	ZERO	HIGH	LOW
MEDIUM	ZERO	ZERO	ZERO	ZERO	HIGH	ZERO	HIGH	LOW
HIGH	ZERO	MEDIUM	ZERO	MEDIUM	ZERO	ZERO	ZERO	HIGH
HIGH	ZERO	ZERO	ZERO	ZERO	ZERO	ZERO	ZERO	HIGH
HIGH	HIGH	MEDIUM	HIGH	MEDIUM	ZERO	ZERO	ZERO	Normal
HIGH	MEDIUM	MEDIUM	HIGH	MEDIUM	ZERO	ZERO	ZERO	Normal
HIGH	ZERO	ZERO	ZERO	HIGH	ZERO	ZERO	ZERO	Normal
MEDIUM	MEDIUM	HIGH	MEDIUM	MEDIUM	ZERO	ZERO	ZERO	Normal
MEDIUM	HIGH	HIGH	HIGH	HIGH	ZERO	ZERO	ZERO	Normal
MEDIUM	MEDIUM	MEDIUM	MEDIUM	MEDIUM	HIGH	ZERO	HIGH	Normal
MEDIUM	MEDIUM	MEDIUM	MEDIUM	MEDIUM	HIGH	HIGH	ZERO	Normal
MEDIUM	MEDIUM	HIGH	MEDIUM	MEDIUM	ZERO	ZERO	ZERO	Normal
MEDIUM	MEDIUM	MEDIUM	HIGH	ZERO	ZERO	ZERO	ZERO	Normal
MEDIUM	HIGH	HIGH	HIGH	MEDIUM	ZERO	ZERO	ZERO	Normal
MEDIUM	MEDIUM	ZERO	MEDIUM	MEDIUM	ZERO	ZERO	ZERO	HIGH
MEDIUM	HIGH	HIGH	HIGH	MEDIUM	ZERO	ZERO	ZERO	Normal
MEDIUM	MEDIUM	MEDIUM	MEDIUM	MEDIUM	ZERO	ZERO	ZERO	Normal

TABLE 4.4: SAMPLE RECORDS POPULATED WITH NOMINAL VALUES.

The records that were entered in MS Access format had the following distribution with respect to the target class labels:

ConservThreatStatus	# of Records	Records in (%)
HIGH	450	37.50
LOW	339	28.25
NORMAL	411	34.25
TOTAL	1200	100

TABLE 4.5: DISTRIBUTION OF RECORDS BASED ON THE TARGET CLASS LABELS

In table 4.5: 450 cases belong to high threat species, 339 cases belong to low threat species and 411 cases belong to normal threat species.

4.1.5 Descriptions of Major Biological Attributes

As explained in Taye et al. (2002), each major biological attributes and its relationship have been described as follows:

(i) **Important value index (IVI):** is the sum of relative density, relative frequency and relative dominance of a given species. IVI might be either to high (not rare or abundant), medium (relatively less threatened), or zero (highly threatened).

(ii) **Regeneration status:** refers to the rating of the sapling and seedling status of the species. For instance, the regeneration of a species might be either to high (highly reproducing itself, less priority species), low (moderately threatened) or zero (highly threatened; high priority species).

(iii) **Logging intensity**: refers to the degree of disturbance, for instance, logging of the matured vegetation. Logging intensity might be either to high (expected to be highly threatened, but not always if the species is highly regenerated), medium or low.

(iv) **Degree of threats (Conservation threat status**): refers to the target variable or biological outcome of interest. Conservation threat might be either to high (considered for conservation since the species is highly threatened with extinction), normal (average or need attention to minimize further threatened) or low (little or no conservation may be required).

(v) **Relative Frequency, Relative Density, Relative Dominance, Sapling, and Seedling:** Each of these biological attributes has causal effect relationship with the above major attributes. For instance, the rate of sapling of a species will affect the regeneration status of the species is either to be high, low or zero.

4.2 BUILDING THE NETWORK MODEL

Building the BN model for the biological modeling and conservation comprises two stages:

(i) Obtaining the network structure, which can also be termed as structural model building,

(ii) Eliciting and reinforcing the obtained structure.

The second step is attempting to combine the opinion of the domain experts with the results that we have got from prior information or objective data. This is, of course, one of the benefits of Bayesian network that makes it different from other machine learning (decision tree, neural network, etc).

In general, the next part of this study focuses on reporting the results of the experiment that would be generated from both prediction and construction of BN model respectively.

4.3 BN MODEL PREDICTION FOR BIOLOGICAL ATTRIBUTES

In order to predict the BN model for biological attributes, we used the 10-fold cross validation technique and the BN PowerPredictor software. The process was mounted on a 32-bit windows system on PC that had 2.0 GHz, 512 MB, and 40 GB hard disk; it was run on Windows XP. In APPENDIX VI, we could visualize how PowerPredictor was learning the CPT (BN)……..

In what follows, the experiment one that was implemented from the biological attributes is presented.

4.3.1 Experiment One

The experiments were done by dividing the dataset into test set and training set. We used 1200 cases or dataset having 120 test sets and 1080 training set. The procedure repeated ten

times. The nine folds were used for training and one fold was used for test set (validity checking). Accordingly, prediction accuracy of the BN model is presented as follows:

4.3.1.1 Prediction Accuracy Before Eliciting Opinions of Domain Experts

The training data sets were prepared as tables in Microsoft access database. Before eliciting the opinions of the domain experts, the estimated prediction accuracy of the 10-fold cross validation results were depicted in confusion matrix for each test datasets based on the target class labels (see APPENDIX III).

The following figure illustrates the best learned network out of the 10-fold experiments.

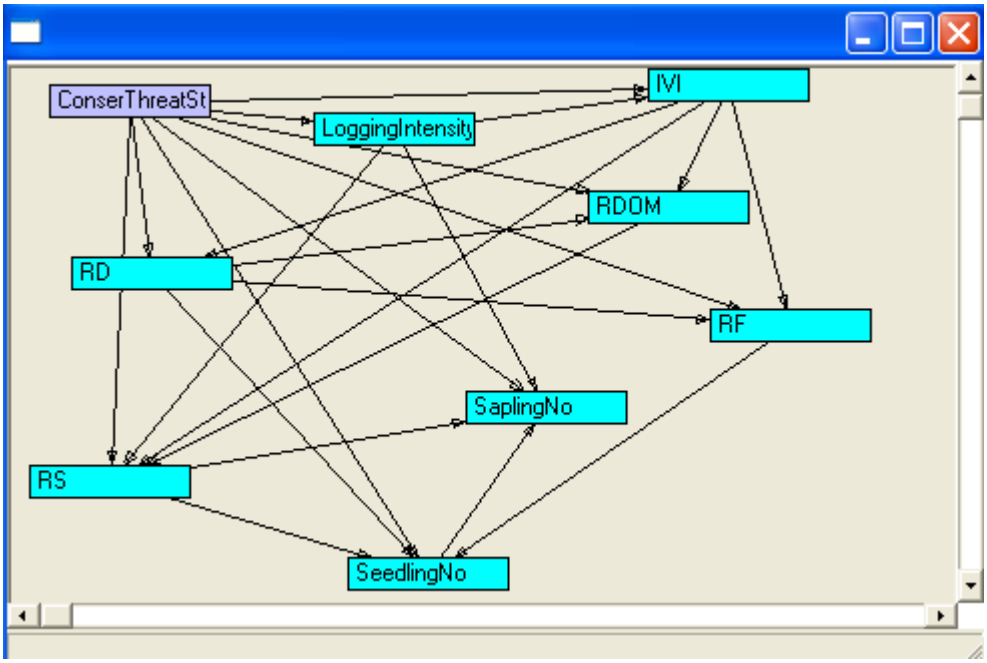


FIGURE 4.8: BEST PREDICTION LEARNED MODEL

The learned prediction model was appeared to be likely to explain the dependencies between attributes. For instance, in the figure 4.8, RS was observed to have four parents, and IVI had two parents. That is, RS was a direct causal effect of LoggingIntensity, RDOM, IVI and ConservThreatStatus. Similarly, IVI was the direct causal effect of LoggingIntensity and ConservThreatStatus. The results in the figure illustrated us how BBN tool captured the direct

influential attributes or dependency between them, however, this might not true from the views of the domain experts (see figure 4.10).

Finally, the average estimated prediction accuracy for ten run test set in the data set was summarized and calculated as follows:

TEST SET	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
PREDICTION ACCURACY (%)	77.69	72.29	73.52	73.98	70.93	72.31	71.67	70.46	76.20	72.22

TABLE 4.6: PREDICTION ACCURACY BEFORE ELICITING OPNIONS OF DOMAIN EXPERTS

In calculation, the average prediction accuracy was found to be 73.13 % with the maximum prediction accuracy set at 95% confidence interval, this means that only 21.87%(i.e., 95%-73.13%) of the target tree species were wrongly classified based on the observed or actual values of the attributes(table 4.4).

This implied that 21.87% of the instances were predicted as inaccuracy of estimating the target class labels. Hence, it was realized that there was risky in prediction of instances. To this end, the best prediction before eliciting opinions of domain experts was the one whose prediction accuracy was 77.69%, and the most risky prediction was the one with 70.46% predictive accuracy at 95% confidence interval. In order to reduce such a risky prediction, further enhancement was conducted with domain experts.

Next, constructing the BN model before and after eliciting the domain experts' opinions were presented.

4.4 CONSTRUCTING BN MODEL FOR BIOLOGICAL ATTRIBUTES

The BN PowerConstructor system was used to construct the BN model. It used the database as input and constructed the belief network structure as output. During construction of the BN model with biological attributes, learning parameters or CPTs for each attributes were accomplished (see APPENDIX VII).

In order to decide whether the BN tool was good in prediction of the network to be constructed, first constructing the network model before eliciting the experts' opinions and then, reinforcing the network model through eliciting the domain experts were conducted.

The network model that was constructed before eliciting the opinions of domain experts was visualized as shown in figure 4.9:

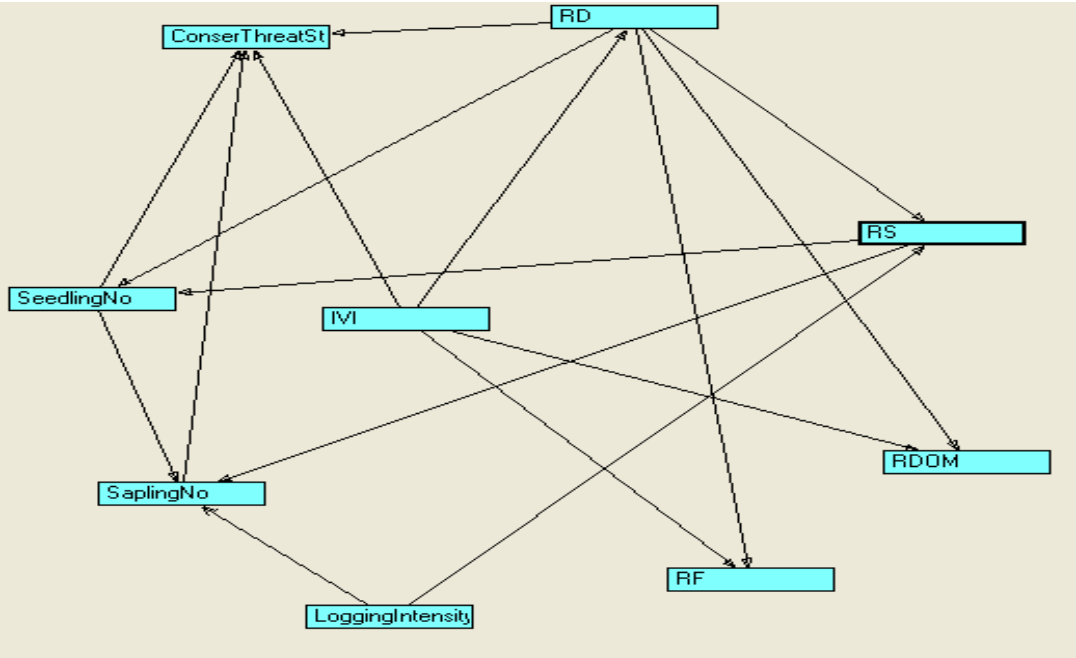


FIGURE 4.9: INITIAL BN MODEL NETWORK USING POWERCONSTRUCTOR

During initial BN model construction, cause-effect relationships or dependency between the attributes were displayed in the following manner:

LoggingIntensity -> RS LoggingIntensity -> SaplingNo IVI -> RD IVI -> RDOM

IVI -> RF IVI -> ConserThreatStatus RD -> RDOM RD -> RF RD -> RS

RD -> SeedlingNo RD -> ConserThreatStatus RS -> SeedlingNo

RS -> SaplingNo SeedlingNo -> SaplingNo SeedlingNo -> ConserThreatStatus

SaplingNo -> ConserThreatStatus

As indicated inside the box above, for instance, LoggingIntensity -> RS means RS was a direct causal effect of the LoggingIntensity. At this particular point no opinions of experts. The BN software itself captured this information. In other words, we couldn't modify or interfere it without getting the opinion of the experts. Hence, getting the opinion of the domain experts thereby reinforcing or modifying the network would be done. For this purpose, we made a discussion with three concerned experts (see partial profile in APPENDIX II) on BN model for biological network that was done before (see figure 4.9).

During modification we visualized the problems such as wrongly assigned arcs, miss directionality or not correctly depicted based on cause-effect relationships, but no node found without getting link as we could see in figure 4.9. For instance, wrongly assigned arcs were depicted in the box below based on the opinion of the domain experts:

LoggingIntensity -> RS RD -> RDOM RD -> RF RD -> RS RD -> SeedlingNo

RD -> ConserThreatStatus SeedlingNo -> SaplingNo

SeedlingNo -> ConserThreatStatus SaplingNo -> ConserThreatStatus

Similarly, wrongly miss directed arcs were depicted as follows in the box:

IVI -> RD IVI -> RDOM IVI -> RF RS -> SeedlingNo RS -> SaplingNo

Thus, both wrongly assigned (orientation) and miss directed (missing orientation) arcs were removed and reoriented respectively. However, no nodes were made out of the network instead correction (adjustment) was done to make the wrongly orientation to properly link based on the knowledge of cause-effect relationship.

Based on the conditional independence of the causal network, the presence of the attribute LoggingIntensity was found to be the cause or relevant for all attributes below it. That is, it was independent of all other nodes in the entire network. This implies that no node has come before it. The final network after reinforcing the BN model was visualized in the following manner:

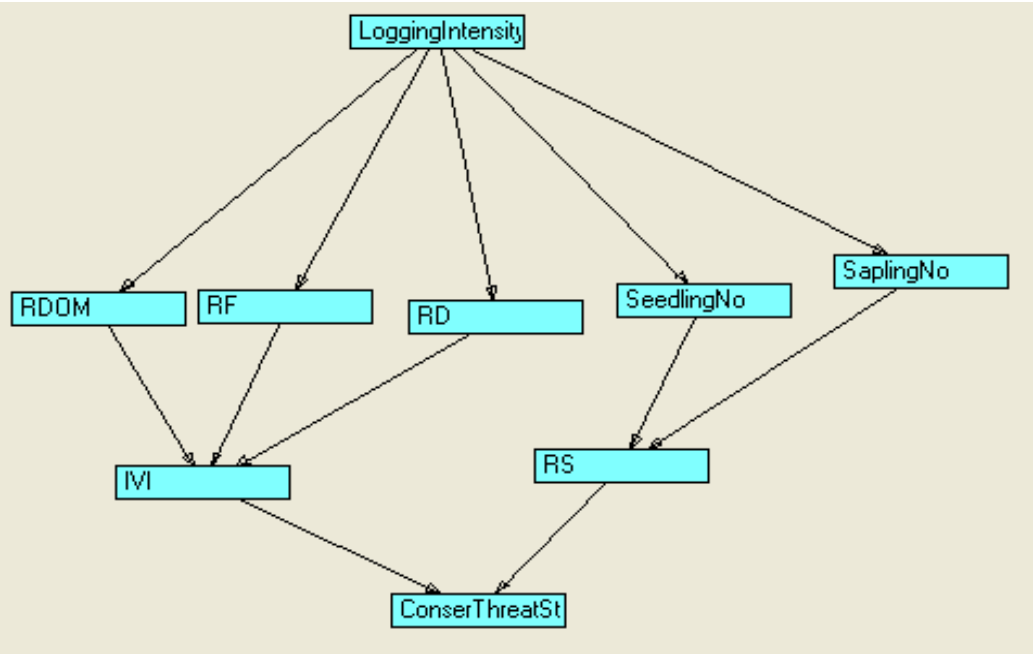


FIGURE 4:10: FINAL BN MODEL FOR BIOLOGICAL NETWORK USING POWERCONSTRUCTOR

In figure 4.10: IVI and Rs are directly influence the target variable, and decision making as to the conservation threat depends on the two major influential attributes (see section 4.4.2).

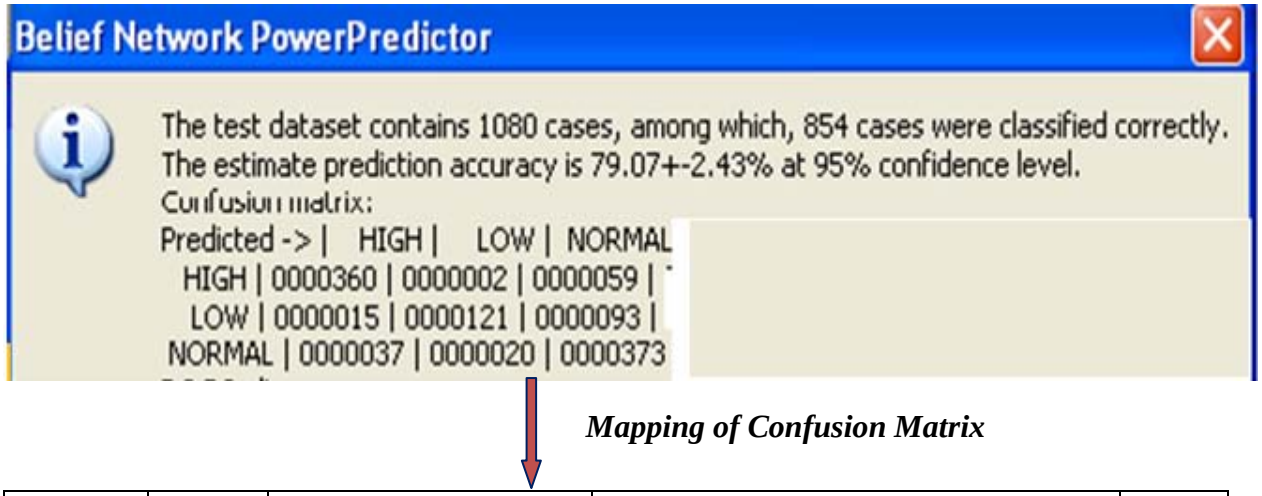
4.4.1 Experiment Two

Experiment two was carried out after eliciting the opinions of domain experts. In similar fashion, prediction of accuracy on the model was done as experiment one.

4.4.1.1Prediction Accuracy after Eliciting Opinions of Domain Experts

In the process of elicitation there was modification on the initial BN model as explained in figure 4.10. As a result, we repeated the same process as we did before eliciting the opnions of domain experts with 10-fold cross validation. The estimated prediction accuracy of the 10-fold run test results were depicted in confusion matrix for each partition of the datasets based on the target class labels (see APPENDIX III). Sample print screen output for the first test dataset was depicted in figure 4.11.

FIGURE 4.11: SAMPLE OUTPUT OF CONFUSION MATRIX AND PREDICTION ACCURACY



		ConservThreatStatus	Predicted			Total
			HIGH	LOW	NORMAL	
Test set 1	**Actual**	**HIGH**	**360(85.51%)**	**2(0.47%)**	**59(14%)**	**421**
		LOW	15	121	93	229
		NORMAL	37	20	373	430

TABLE 4.7: SAMPLE OUTPUT OF CONFUSION MATRIX TEST SET ONE

From the table 4.7, it is explained that a confusion matrix is a simpler and more useful outcomes, and it compares predicted with actual (or real data) outcomes. In the example given here,

- The confusion matrix depicts that out of the total test records provided to the BN tool; about **79.07%** of the records were classified correctly.

- Confusion matrix, the chance of classification of a "**HIGH**" threat category species into a "**NORMAL**", which may be risky in the learning process was only **0.14** (**14 %** of actually **HIGH** threat species were predicted as **Normal** species).
- 360 cases (i.e., **85.51%**) in which the species was actually high in threat the model correctly predicted.

Finally, the average estimated prediction accuracy for ten run test set in the data set was summarized and calculated as follows:

TEST SET	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
PREDICTION ACCURACY (%)	79.07	76.30	78.06	74.07	74.54	75.00	73.98	78.43	72.64	76.09

TABLE 4.8: PREDICTION ACCURACY AFTER ELICITING OPNIONS OF THE DOMAIN EXPERTS

In calculation, the average prediction accuracy that was computed from the confusion matrix was found to be 75.76 %, with the maximum prediction accuracy set at 95%, this means that only 19.24% of the target tree species were wrongly classified based on the actual values of the attributes (table 4.4).

The result implied that 19.24% of the instances were predicted as inaccuracy of estimating the target class labels. Hence, it was again visualized that there was risky or variation in prediction of instances, though the result in experiment two was found to be promised for further indications and research work in the problem domain.

Comparing the two experiments (i.e., before and after eliciting process) in prediction accuracy; there was a slight discrepancy or variation i.e., 75.76-73.13=2.63 %.

To this end, the best prediction after eliciting process was the one whose prediction accuracy was 79.07%, and the most risky prediction was the one with 72.64% predictive accuracy at 95% confidence interval.

This completes the discussion of the experiments conducted in relation to the prediction and constructing of the model.

Following this, samples of demonstration were taken for the purpose of checking its accuracy[9] and how far the results were able to classify the target class label based on the arbitrary input cases (see section 4.4.1.2).

4.4.1.2 Demonstration of Instance Classification

This was the step where we attempted to see by taking arbitrary input cases as the class label instantly. See below the sample instance classification:

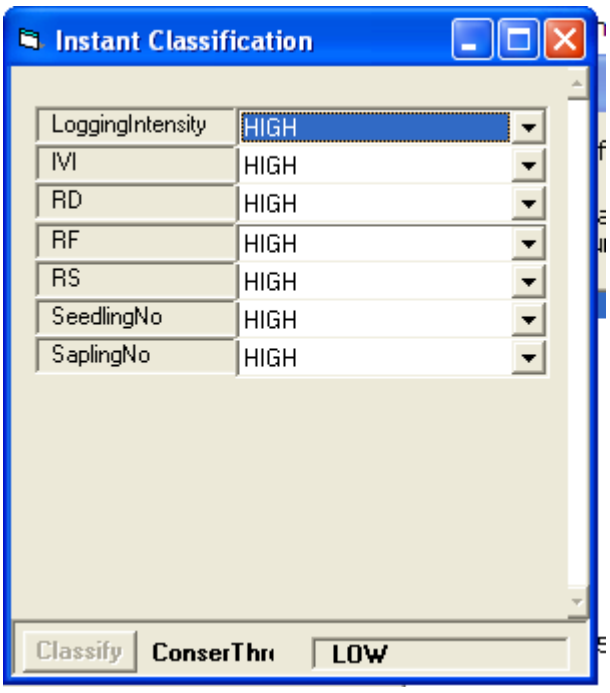


FIGURE 4.12: INSTANCE CLASSIFICATION USING PowerPredictor

[9] Accuracy refers to the existing real situations that the domain experts expect to obtain through instance input cases (Taye et al., 2002).

From the figure 4.12: Input cases, for instance, IVI=HIGH, RS=HIGH, the correct class label value was found to be LOW. This implied that the values that were automatically generated for each attributes predicted the class label as LOW. Hence, when we compared with the real situations (i.e., knowledge of the domain experts), the predicted class label was found valid.

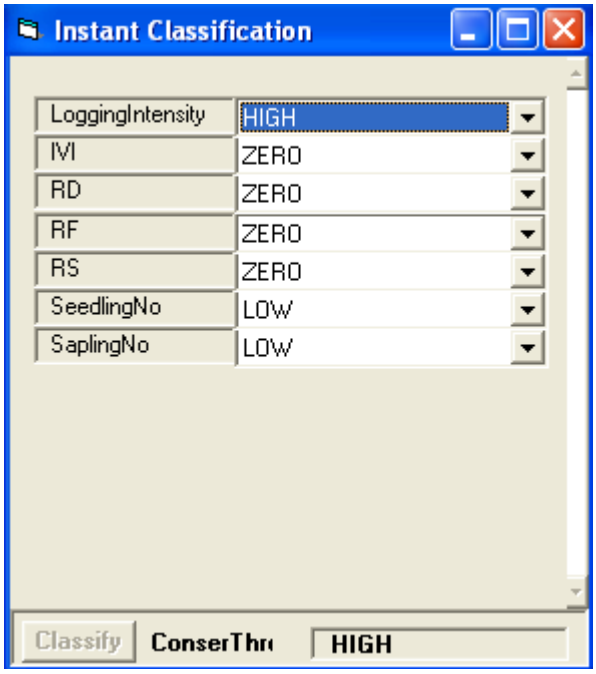


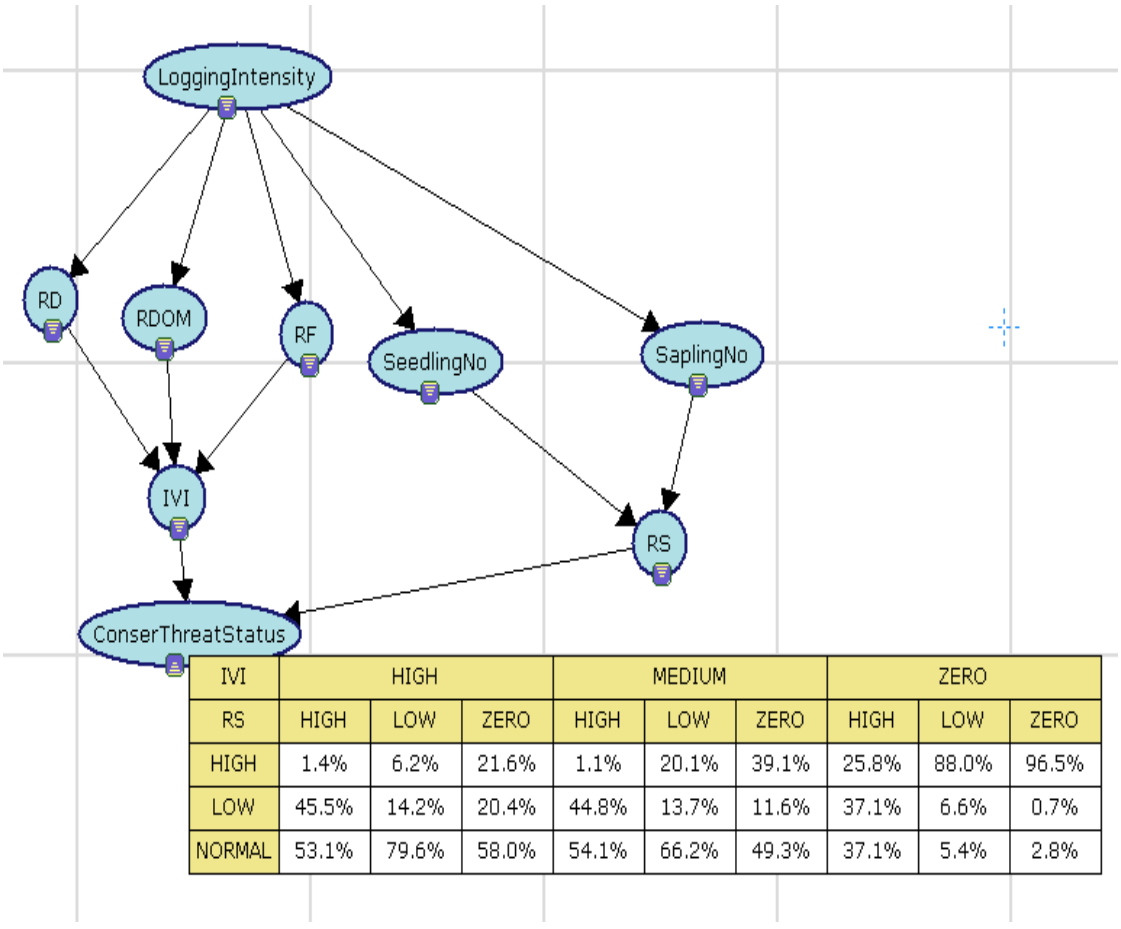
FIGURE 4.13: INSTANCE CLASSFICATION USING PowerPredictor

From figure 4.13: Input cases, for instance, IVI=ZERO, RS=ZERO, the correct class label value was found to be HIGH. Again, the predicted class label was found valid. The results were illustrated based on the direct causal effect of the target variable (see figure 4.10), i.e, **RS → CoservThreatStatus ← IVI.** That is, other variables were not important or irrelevant to make decision of such type. However, this doesn't mean they are not essential or relevant for BN model construction since they are either independence or conditional independence given its parents and thereby its cumulative effect observed in the conditional probability tables for the target variable (see figure 4.14).

Therefore, these two instance examples (i.e., figure 4.12 and 4.13) implied that it is likely to develop the full fledged system and/or machine learning classifiers for the domain experts (users) and thus, making a good decision in the future.

4.4.2 Visualization of Conditional Probability Tables(CPTs)

As indicated below we visualized the amount of conditional probability tables of the target variable "ConservThreatStatus" using the BNJ software that was obtained from the BN network (see figure 4.10).



IVI	HIGH			MEDIUM			ZERO		
RS	HIGH	LOW	ZERO	HIGH	LOW	ZERO	HIGH	LOW	ZERO
HIGH	1.4%	6.2%	21.6%	1.1%	20.1%	39.1%	25.8%	88.0%	96.5%
LOW	45.5%	14.2%	20.4%	44.8%	13.7%	11.6%	37.1%	6.6%	0.7%
NORMAL	53.1%	79.6%	58.0%	54.1%	66.2%	49.3%	37.1%	5.4%	2.8%

FIGURE 4.14: SAMPLE VISUALIZATION OF CPT FOR TARGET VARIABLE.

From the visualization in figure 4.14, we could see that the values of CPT when RS=ZERO, IVI=ZERO, and CoservThreatStatus=HIGH was 96.5%, which was very much likely to mirror the image of the real world situation (i.e., knowledge of the domain experts).

In other words, the conditional probability was computed as:

P (CoservThreatStatus=HIGH| RS=ZERO, IVI=ZERO) =0.965. i.e., the result of this particular instance from the CPT was showed us there was highly probably threatened tree species that should be considered for conservation.

Sample queries from the visualization,

Query 1: What is the chance of being LOW threat species given RS=HIGH and IVI=HIGH?

i.e., P (ConservThreatStatus=LOW|RS=HIGH, IVI=HIGH) =0.455. i.e., it was likely to consider species with little or no conservation based on the observed cases. Since species potential to regenerate itself is high, and the availability of species on the field is also abundant.

Query 2: What is the chance of being NORMAL threat species given RS=HIGH and IVI=HIGH?

i.e., P(ConservThreatStatus=NORMAL|RS=HIGH, IVI=HIGH) =0.531. i.e., it was highly likely to monitor or pay attention the species based on observed cases.

Query 3: What is the chance of being HIGH threat species given RS=HIGH and IVI=HIGH?

i.e., P (ConservThreatStatus=HIGH|RS=HIGH, IVI=HIGH) =0.014. This result also implied there was unlikely to consider the target tree species (species) for conservation when both the RS and IVI were high. Since there was high regeneration and important value index in the area where sampling was taking during inventory work, and thus, the belief of the domain experts were also low under the given evidences.

4.5 DISCUSSION OF THE MODEL RESULTS

In order to determine the accuracy of the BN Model for biological network we have used the dependency analysis method with ten-fold cross validations technique. The validation has

been implemented by first randomly breaking the full dataset (1200 dataset) into ten partitions. Randomization is done internally (for training and test set) as required by BN PowerPredictor withoout the researcher intervention. This procedure is repeated ten times. The experimental results show that on average, the estimated prediction accuracy before and after eliciting opinions of domain experts are 73.13% and 75.76 % respectively.

Findings indicated that there is no great discrepancy between experiment one and two, and hence, the Bayesian belief network with PowerConstructor system is found to be a good predictor of such type of work even in the absence of the domain experts. Moreover, both examples of instance classification and conditional probabilities have demonstrated the validity (i.e., strength) estimation of the model results.

However, with active participation or involvement of the domain experts, there is indeed an opportunity of enhancing the BN model prediction for the application domain and thus, making a good prediction inference unquestionable.

The final estimated average error rate is 19.24%. The error rates indicated that further research and appropriate database (i.e., consistence and reliable data) are required for BN model and hence, not only the expected BN model prediction will be highly enhanced in its performance, but also it will be highly matched with the domain experts' experience. Hence, making an effective and efficient decision (or inference) will be easy.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1. CONCLUSION

In this study, the findings indicated that the results in the experiment has an indication to model a framework for learning input from the database and for estimating the prediction accuracy of the BN model for biological modeling of the target tree species.. It is also found that the BN model can predict or estimate the conservation status of species based on the prior information for pragmatic analysis.

Findings indicated that there is no great discrepancy between experiment one and two and hence, the Bayesian belief network with PowerConstructor system is found to be a good predictor of such type of work even in the absence of the domain experts.

The current study is also made in an effort to model the BN in biological modeling for conservation action, which assists the information provision to conservation biologist, foresters and agricultural researchers.

In the experimental results as we have attempted to depict the confusion matrix for each training sets with 10-fold cross validation technique, there is on average 19.24% error rate. It implied that the BN model requires further study. The Model has not yet perfect to do a good inference in all cases. Moreover, the model concentrates on the major biological attributes and hence, it should also consider the socioeconomic factors that are believed to be important in biological modeling and conservation of target tree species.

At last, loss of information during discretization may reduce the prediction accuracy of the model and hence, use of other strategy is also pertinent to improve it.

5.2 RECOMMENDATIONS

In this study, the following recommendations are made:

1. As BN model is a machine learning tool, further testing and validation/evaluation of the system is a necessity. For that matter, users, outside of the panel experts who helped in the early modeling of BN should be actively involved for reinforcement and enhancement of the prediction of the model result.
2. In the current work, only a few target tree species' records were considered for BN model. In order to complete the model result, it is also necessary to include all the available researched target tree species.
3. The study used dependency analysis approach to see the dependency between the biological attributes. Other approaches have to be studied in order to make comparison.
4. The study has only tried to see where case has with unknown/known structure and complete data, however, further study is also important where cases have with unknown structure and partial data, and known structure and partial data.
5. In the current study we used 10-fold cross validation; other standard quality network measures or validation techniques should be used in order to have good comparisons of the different evaluation techniques in prediction of the BN model for conservation of target tree species.
6. The model used in this research was simplistic and considered only a few variables. More variables need to be considered in order to formulate the physical model which is close to reality, for instance, socioeconomics factors.
7. The study could able to see only one of the uncertainty management techniques that are Bayesian belief networks. It is also important to study and implement other uncertainty techniques.

8. In this study, a combined method (i.e., Netica application and Opinions of domain experts) has been used for discretization of the continuous data. The problem of the discretization of continuous variables arises as an issue in Bayesian Belief Network technique because conditional probabilities are based on discrete variables. There is a possibility of loss of important information due to discretization. It is recommended that better methods of discretization be studied and implemented to reduce the loss of information that might be caused by using simple categorization methods.

REFERENCES

Ames, D.P.(2003). Using Bayesian networks to model watershed management decisions: an East Canyon creek case study. http://www.hydromap.com/papers/ames_ecc.pdf

Bayes, T. (1763). An essay toward solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society, pp. 370–418.

Breese, J.S. and Heckerman, D.(1999). Decision theoretical troubleshooting. IEEE Transactions on Systems, Man and Cybernetics, Part A (Systems and Human), 26(6):838-842.

Cain et al.(1999): Belief networks: a Frame work for the participatory development of natural resources management strategies.

Cheng, J. (1998). ” Learning Bayesian Networks from Data: An Efficient Approach Based on Information Theory”, Technical Report, Department of Computer Science, University of Alberta.

Cooper, G.F. and Herskovits, E. (1992). A Bayesian Method for the induction of probabilistic networks from data. Machine Learning, 9 (pp. 309-347).

Finkel,(1996): An introduction to Bayesian inference for ecological research and environmental decision- making. Ecol. Appl. 6: 1036-1046.

Girma Balcha (2002). Conservation and sustainable use of Forest Genetic Resources. Addis Ababa, Ethiopia. PP. 150-157. In Proceedings of A National Conference on Forest Resources of Ethiopia.

GBS (Global Biodiversity Strategy), (1992). Harvard University, USA: 19-20.

Green et al. (2005): Complexity in ecology and conservation: mathematical, statistical, and computational challenges. Bioscience, 55:501:510.

Ham and Kamber (2001): Data mining: concepts and techniques.

Hansson, O. and Mayer, A. (1989). Heuristic search as evidential reasoning. In Proceedings of the 5th Workshop on Uncertainty in Artificial Intelligence.

Heckerman, D. (1995). A tutorial on learning Bayesian networks. Technical Report MSR-TR-95-06. Microsoft Research.

Heckerman, et al. (2000). Dependency Networks for Inference, Collaborative Filtering, and Data Visualization. Journal of Machine Learning Research, 1: p. 49-75.

IUCN (The World Conservation Union), (1994). IUCN Red List Categories. Gland: 21 PP.

Jeffrey, R. C. (1992). Probability and the art of judgments. Cambridge University Press, Cambridge, UK.

Jones et al.(2002): A strategy for habitat supply modeling for British Columbia.

Kanga and Kangas (2004): Probability, possibility and evidence: approaches to consider risk and uncertainty in forest decisions analysis.

Kuikka et al. (1999): modeling environmentally driven uncertainties in Baltic cod Management by Bayesian influence diagrams.

Long., W. (1989). Medical diagnosis using a probabilistic causal network. Applied Artificial Intelligence, 3:367-383.

Lauritzen, S. (1988): "Local computations with probabilities on graphical structures and their application to expert systems" in J. Royal Statistics Society B, 50(2), 157-194.

Marcot, et.al.(2006). Guidelines for Developing and Updating Bayesian belief network applied to ecological modeling and conservation, NRC, Canada.

McCann K., Marcot G., Ellis R. (2006): Bayesian belief networks: applications in ecology and natural resource management.

McCarthy et al.(2001). Assessing Spatial PVA models of arboreal marsupials using significance tests and Bayesian statistics: Bio. Conserv. 98:191-200

McNay et al.(2006): a Bayesian Approach to evaluating habitat for woodland caribou in north-central British Columbia.

McNeely, J.A. (1988). Economics and Biological Diversity. IUCN, Gland, Switzerland

Mead et al.(2006). Applications of Bayesian networks in ecological modeling, Montana State University – Bozeman, USA.

Namkoong, gene and Koshy, Mathew P. (2000), Decision Making in gene Conservation. Forset genetic Resources 28.

Neapolitan, R. E.(2004). "Learning Bayesian Networks", Prentice Hall Series in Artificial Intelligence,

Newton, A., Oldfield, S., Fragoso, G., Mathew, P., Miles, L., & Edwards, M.(2003). Towards a Global Tree Conservation Atlas. UNEP-WCMC/FFI. http://www.unep-wcmc.org/resources/ publications/treeatlas

Nikovski, D.(2000). Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics. IEEE Transactions on Knowledge and Data Engineering, 12(4):50-516.

Oldfield, S.F., Lusty, C. and MacKinven, A.(1998) .The World List of Threatened Trees. World Conservation Press, Cambridge,

Olson et al.(1990a): A framework for modeling uncertain reasoning in ecosystem management II: Bayesian belief network.

Pearl, J.(1988). Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann.

Possingham, H. P.(1997). State-dependent decision analysis for conservation biology. Pages 298–304 in S. T. A. Pickett, R. S. Ostfield, M. Shachak, and G. E. Likens, editors. The ecological basis of conservation: heterogeneity, ecosystems and biodiversity. Chapman and Hall, New York, New York, USA.

Rahel Bekele (2005). Computer-Assisted Learner Group Formation Based on Personality Traits, Hamburg, Germany.

Regan, H. M., M. Colyvan, and M. A. Burgman. (2002). A taxonomy and treatment of uncertainty for ecology and conservation biology. Ecological Applications 12:618–628.

Samir Abduselam (2001). An Application of Expert Systems on Species Selection: The case of Forestry Research Center.

Sando, T. (2005). Modeling Highway Crashes Using Bayesian Networks: The Florida State University, College Of Engineering.

Schroth, G., et.al. (1996). Forest Ecology and Management, Volume 84, Issues 1-3, Pages 199-208.

Spiegelhalter, J.(1993). “Bayesian analysis in expert systems” in Statistical Science, 8(3), 219-283.

Spirtes, P.(2000). ”Causation, Prediction and Search”, 2nd Edition, MIT Press,

Staarfield, A.M., and Bleloch, A.L.(1986). Building models for conservation and wildlife management. Macmillan Publishing Co., New York.

Taye Bekele, Kumelachew Yeshitila, Shiferaw Dessie and Günther Haase (2002). Priority Woody Species of the Moist Montane Forests of Southwest Ethiopia: Consideration for Conservation.

Taylor et al. (2000). Incorporating uncertainty into management models for marine mammals. Conserve. Biol. 14: 1243-1252

Whitten, I.H., (2005). Data mining: Practical Machine Learning tools and techniques, Second Edition.

APPENDICES

APPENDIX I: LIST OF TARGET TREE PECIES

(Source: Institute of Biodiversity Conservation(IBC), Addis Ababa)

Genus	Species
Acacia	abyssinica
Acalypha	acrogyna
Albizia	grandibracteata
Albizia	gummifera
Albizia	schimperiana
Alchornea	laxiflora
Allophylus	abyssinicus
Allophylus	macrobotrys
Antiaris	toxicaria
Apodytes	dimidiata
Arundinaria	alpina
Bersama	abyssinica
Blighia	unijugata
Bridelia	micrantha
Brucea	antidysenterica
Buddleja	polystachya
Calpurnia	aurea
Canthium	oligocarpum
Cassipourea	malosana
Celtis	africana
Celtis	philippensis
Celtis	zenkeri
Clerodendrum	myricoides
Coffea	arabica
Cordia	africana
Croton	macrostachyus
Cyathea	manniana
Diospyros	abyssinica
Discopodium	penninervium
Dombeya	torrida
Dracaena	fragrans
Ekebergia	capensis
Elaeodendron	buchananii
Elaeodendron	buchananii
Embelia	schimperi
Entada	abyssinica

Erythrococca	trichogyne
Euphorbia	abyssinica
Euphorbia	abyssinica
Fagaropsis	angolensis
Ficus	exasperata
Ficus	mucuso
Ficus	ovata
Ficus	sur
Ficus	sycomorus
Ficus	thonningii
Ficus	vallis-choudae
Ficus	vasta
Flacourtia	indica
Gardenia	ternifolia
Grewia	ferruginea
Hagenia	abyssinica
Hallea	rubrostipulata
Hypericum	revolutum
Ilex	mitis
Lecaniodiscus	fraxinifolius
Lepidotrichilia	volkensii
Lobelia	giberroa
Maesa	lanceolata
Malacantha	alnifolia
Manilkara	butugi
Maytenus	addat
Maytenus	arbutifolia
Millettia	ferruginea
Mimusops	kummel
Morus	mesozygia
Myrsine	africana
Nuxia	congesta
Olea	capensis
Oncoba	spinosa
Oxyanthus	speciosus
Pavetta	abyssinica
Phoenix	reclinata
Phyllanthus	reticulatus
Polyscias	fulva
Premna	schimperi
Prunus	africana
Rhus	natalensis
Ricinus	communis
Ritchiea	albersii
Sapium	ellipticum
Schefflera	abyssinica

Schrebera	alata
Senna	petersiana
Stereospermum	kunthianum
Strychnos	mitis
Syzygium	guineense
Tamarix	aphylla
Teclea	nobilis
Trema	orientalis
Trichilia	dregeana
Trichilia	prieuriana
Trilepisium	madagascariense
Turraea	holstii
Vangueria	apiculata
Vernonia	amygdalina

APPENDIX II: PARTIAL PROFILES OF DOMAIN EXPERTS

Qualification	Experience	Remark
PHD, Ecology	15	
PHD, Siliviculture	20	
M.Sc. GIS and Earth Observation	10	

APPENDIX III: RESULTS OF ESTIMATED PREDICTION ACCURACY

BEFORE ELICITING OPNIONS OF DOMAIN EXPERTS [Experiment One]

CONFUSION MATRIX FOR TRAINIG TEST SET 1

			Predicted		
			HIGH	LOW	NORMAL
Test set 1	**Actual**	HIGH	348	6	67
		LOW	12	129	88
		NORMAL	28	40	362

In training set 1: 839 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **77.69%** at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 2

			Predicted		
			HIGH	LOW	NORMAL
Test set 2	**Actual**	HIGH	295	11	104
		LOW	2	105	104
		NORMAL	28	40	362

In training set 2: 780 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **72.29%** at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 3

			Predicted		
			HIGH	LOW	NORMAL
Test set 3	**Actual**	HIGH	320	10	83
		LOW	16	92	106
		NORMAL	33	38	382

In training set 3: 794 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **73.52%** at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 4

			Predicted		
			HIGH	LOW	NORMAL
Test set 4	**Actual**	HIGH	311	1	89
		LOW	10	93	113
		NORMAL	24	44	395

In training set 4:799 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **73.98%** at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 5

			Predicted		
			HIGH	LOW	NORMAL
Test set 5	**Actual**	HIGH	281	16	105
		LOW	4	105	106
		NORMAL	8	75	380

In training set 5: 766 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **70.93%** at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 6

			Predicted		
			HIGH	LOW	NORMAL
Test set 6	**Actual**	HIGH	279	9	115
		LOW	11	85	107
		NORMAL	23	34	417

In training set 6: 781 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **72.31%** at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 7

			Predicted		
			HIGH	LOW	NORMAL
Test set 7	**Actual**	HIGH	311	1	89
		LOW	10	93	113
		NORMAL	24	44	395

In training set 7: 774 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **71.67%** at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 8

			Predicted		
			HIGH	LOW	NORMAL
Test set 8	**Actual**	HIGH	288	10	115
		LOW	11	119	116
		NORMAL	20	43	354

In training set 8: 761 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **70.46** % at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 9

			Predicted		
			HIGH	LOW	NORMAL
Test set 9	**Actual**	HIGH	287	4	93
		LOW	6	119	99
		NORMAL	18	37	417

In training set 9: 823 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **76.20%** at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 10

			Predicted		
			HIGH	LOW	NORMAL
Test set 10	**Actual**	HIGH	266	12	92
		LOW	6	125	109
		NORMAL	21	60	389

In training set 10: 780 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **72.22%** at 95% confidence interval.

AFTER ELICITING OPNIONS OF DOMAIN EXPERTS [Experiment Two]

CONFUSION MATRIX FOR TRAINIG TEST SET 1

			Predicted		
			HIGH	LOW	NORMAL
Test set 1	**Actual**	HIGH	360	2	59
		LOW	15	121	93
		NORMAL	37	20	373

In training set 1: 854 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **79.07%** at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 2

			Predicted		
			HIGH	LOW	NORMAL
Test set 2	**Actual**	HIGH	327	2	84
		LOW	9	112	90
		NORMAL	30	41	385

In training set 2: 824 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is 76.30 % at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 3

			Predicted		
			HIGH	LOW	NORMAL
Test set 3	**Actual**	HIGH	332	2	79
		LOW	11	96	107
		NORMAL	24	14	415

In training set 3: 843 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is 78.06 % at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 4

			Predicted		
			HIGH	LOW	NORMAL
Test set 4	**Actual**	HIGH	317	1	83
		LOW	9	94	113
		NORMAL	28	46	389

In training set 4: 800 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **74.04** % at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 5

			Predicted		
			HIGH	LOW	NORMAL
Test set 5	**Actual**	HIGH	319	6	77
		LOW	7	98	110
		NORMAL	27	48	388

In training set 5: 805 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is 74.54 % at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 6

			Predicted		
			HIGH	LOW	NORMAL
Test set 6	**Actual**	HIGH	320	6	77
		LOW	10	95	100
		NORMAL	39	38	397

In training set 6: 810 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is 75.00% at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 7

			Predicted		
			HIGH	LOW	NORMAL
Test set 7	**Actual**	HIGH	319	7	68
		LOW	12	116	120
		NORMAL	27	47	364

In training set 7: 799 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **73.98** % at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 8

			Predicted		
			HIGH	LOW	NORMAL
Test set 8	**Actual**	HIGH	356	1	56
		LOW	14	140	96
		NORMAL	45	21	351

In training set 8: 847 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is 78.43% at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 9

			Predicted		
			HIGH	LOW	NORMAL
Test set 9	**Actual**	HIGH	291	4	89
		LOW	4	109	111
		NORMAL	21	73	378

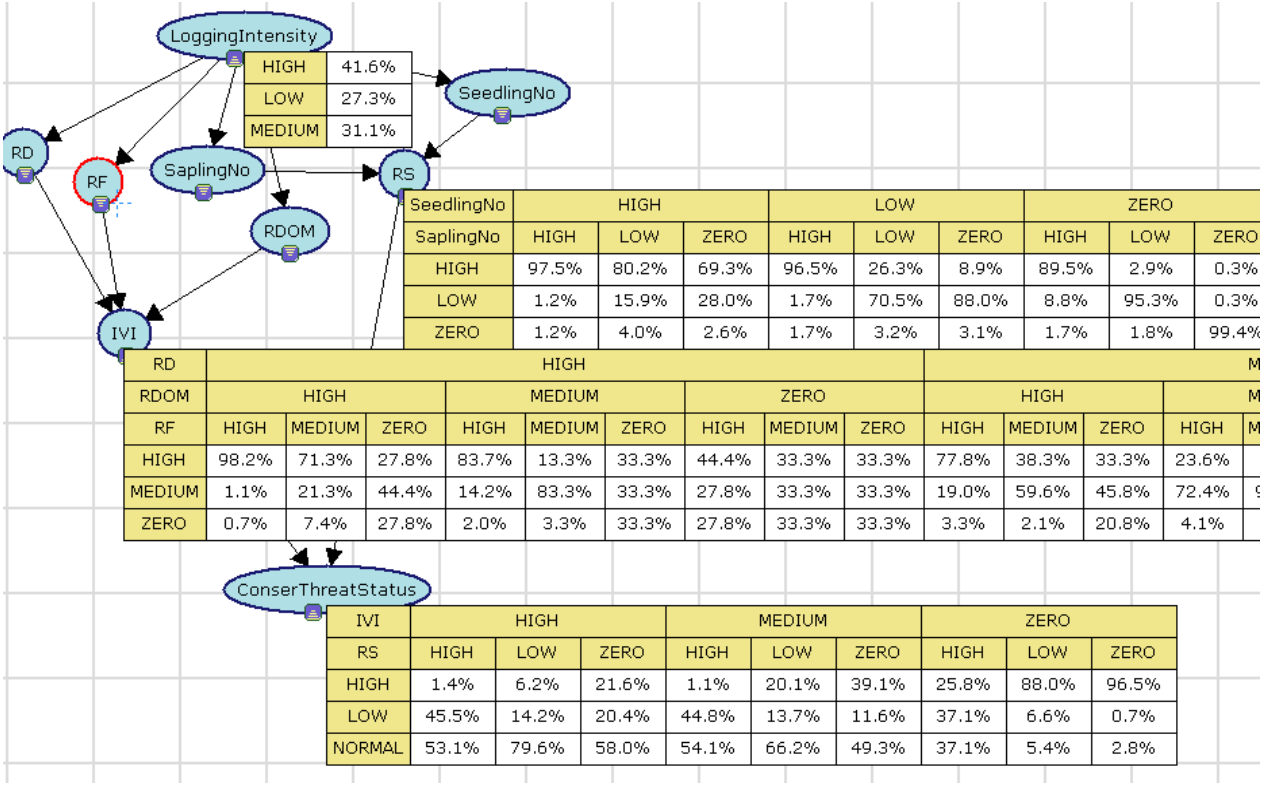
In training set 9: 778 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **72.64** % at 95% confidence interval.

CONFUSION MATRIX FOR TRAINIG TEST SET 10

			Predicted		
			HIGH	LOW	NORMAL
Test set 10	**Actual**	HIGH	288	8	73
		LOW	8	139	93
		NORMAL	26	50	394

In training set 10: 821 cases out of 1080 test data set were classified correctly and the estimated prediction accuracy is **76.09%** at 95% confidence interval.

APPENDIX IV: VISUALIZATION OF CPT USING BNJ



SeedlingNo	HIGH			LOW			ZERO		
SaplingNo	HIGH	LOW	ZERO	HIGH	LOW	ZERO	HIGH	LOW	ZERO
HIGH	97.5%	80.2%	69.3%	96.5%	26.3%	8.9%	89.5%	2.9%	0.3%
LOW	1.2%	15.9%	28.0%	1.7%	70.5%	88.0%	8.8%	95.3%	0.3%
ZERO	1.2%	4.0%	2.6%	1.7%	3.2%	3.1%	1.7%	1.8%	99.4%

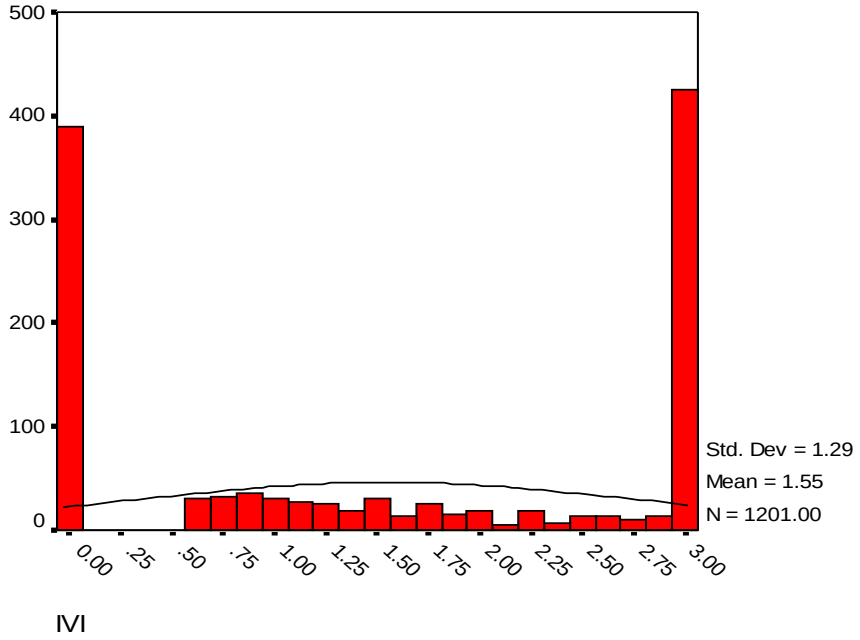
RD	HIGH													M
RDOM	HIGH			MEDIUM			ZERO			HIGH				M
RF	HIGH	MEDIUM	ZERO	HIGH	MEDIUM	ZERO	HIGH	MEDIUM	ZERO	HIGH	MEDIUM	ZERO	HIGH	M
HIGH	98.2%	71.3%	27.8%	83.7%	13.3%	33.3%	44.4%	33.3%	33.3%	77.8%	38.3%	33.3%	23.6%	
MEDIUM	1.1%	21.3%	44.4%	14.2%	83.3%	33.3%	27.8%	33.3%	33.3%	19.0%	59.6%	45.8%	72.4%	9
ZERO	0.7%	7.4%	27.8%	2.0%	3.3%	33.3%	27.8%	33.3%	33.3%	3.3%	2.1%	20.8%	4.1%	

IVI	HIGH			MEDIUM			ZERO		
RS	HIGH	LOW	ZERO	HIGH	LOW	ZERO	HIGH	LOW	ZERO
HIGH	1.4%	6.2%	21.6%	1.1%	20.1%	39.1%	25.8%	88.0%	96.5%
LOW	45.5%	14.2%	20.4%	44.8%	13.7%	11.6%	37.1%	6.6%	0.7%
NORMAL	53.1%	79.6%	58.0%	54.1%	66.2%	49.3%	37.1%	5.4%	2.8%

APPENDIX V: SAMPLE DISTRIBUTION OF VALUES USING SPSS

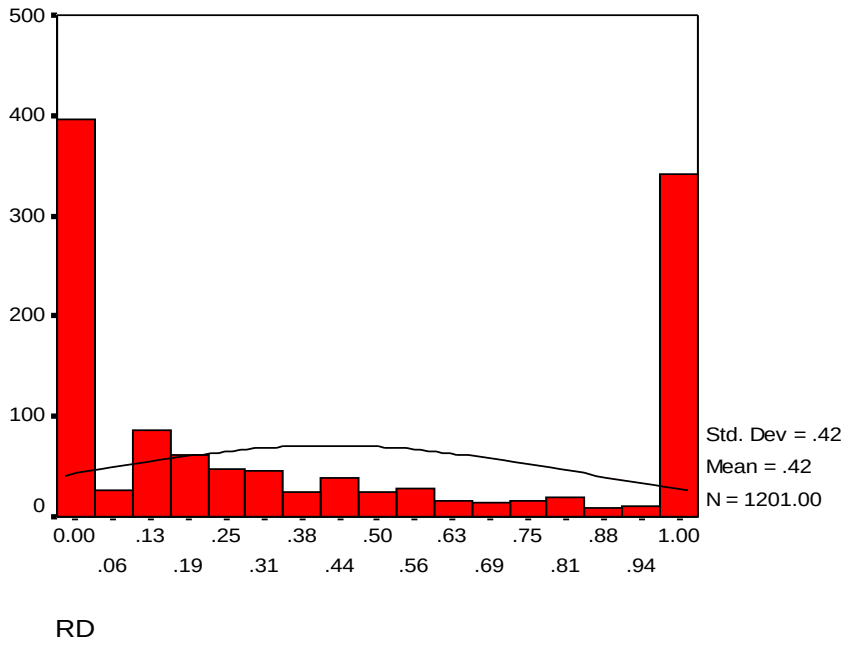
Distribution values of IVI using SPSS 11.0

[Not normally distributed with the given mean and standard deviation]

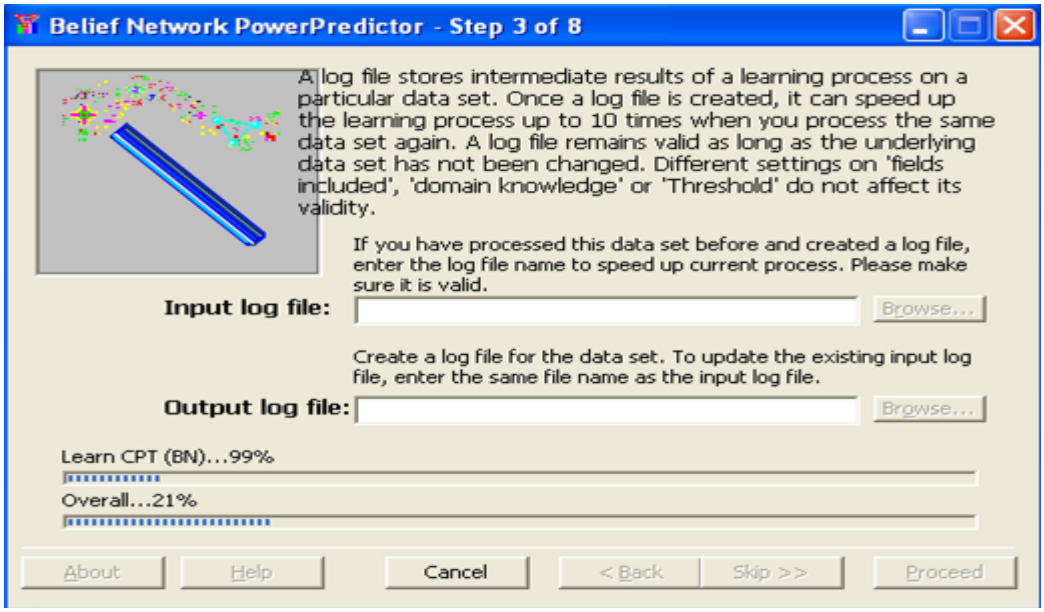


Distribution values of RD using SPSS 11.0

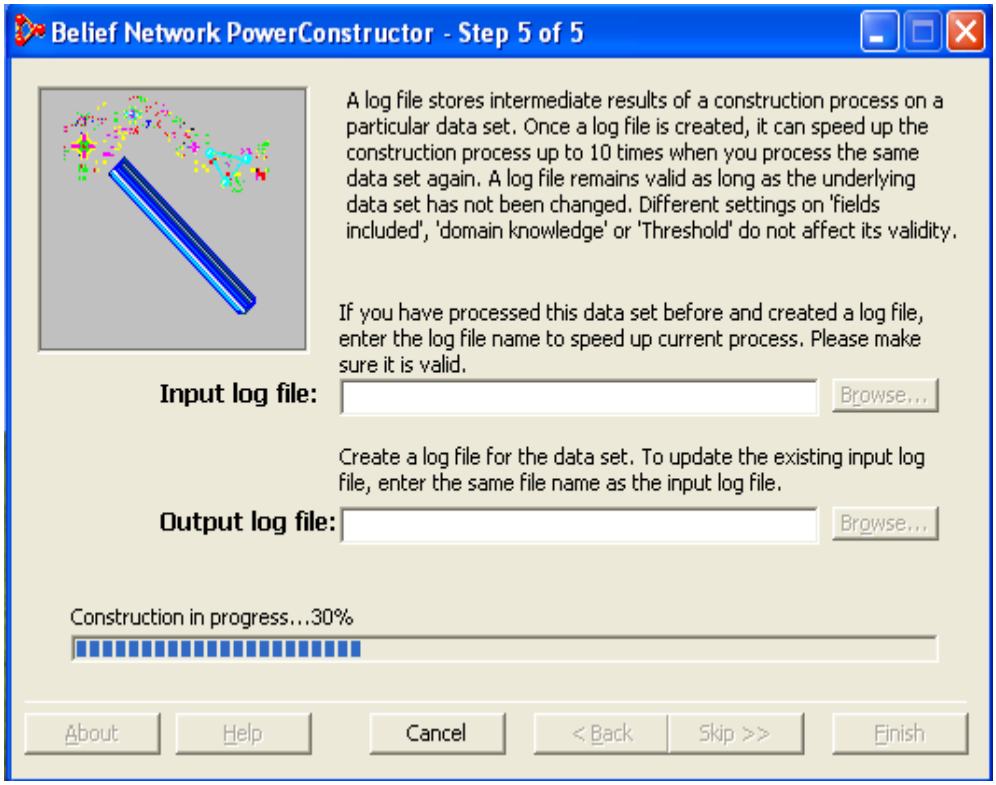
[Not normally distributed or bell shaped with the given mean and standard deviation]



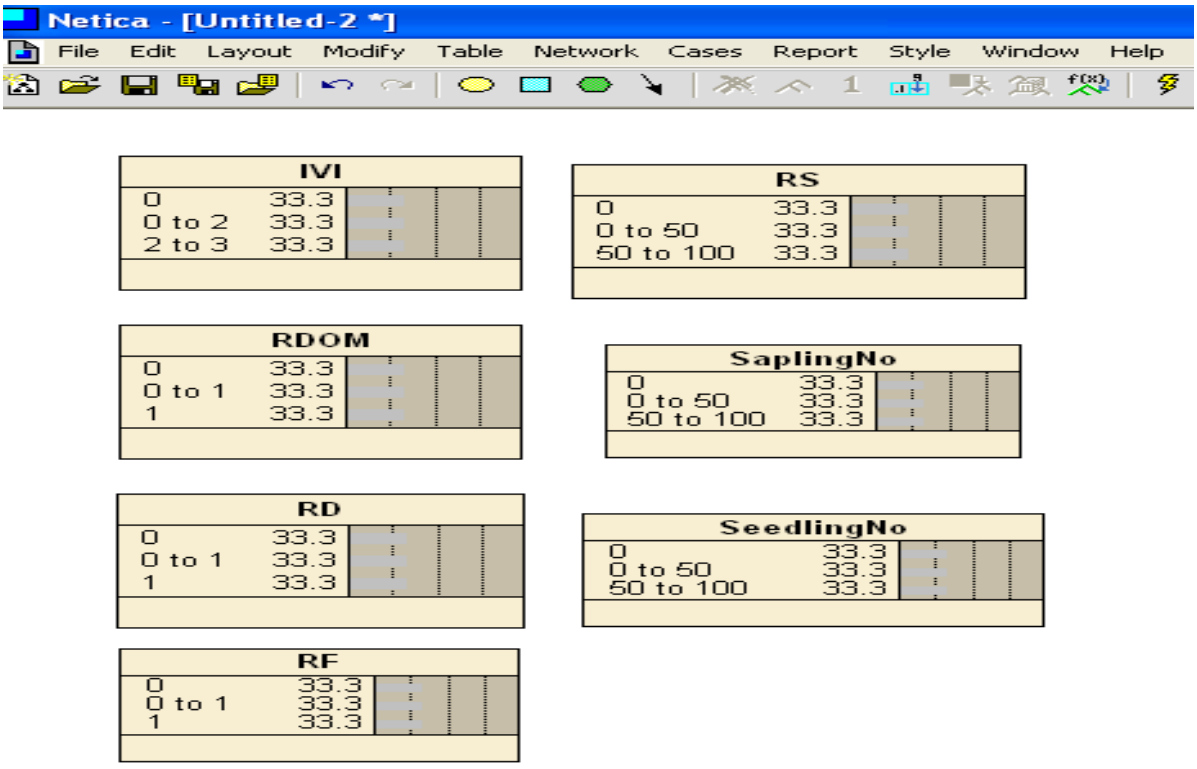
APPENDIX VI: LEARNING CPT (BN) PROCESS USING BN PowerPredictor



APPENDIX VII: BN PowerConstructor WHILE LEARNING PARAMETERS



APPENDIX VIII: APPLYING NETICA APPLICATION SOFTWARE FOR DISCRETIZATION OF CONTINIUOUS DATA



DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

_____

Behailu Getachew Wolde

March 2009

The thesis has been submitted for examination with my approval as university advisor

_____

Dr. Rahel Bekele

March 2009