School of Spatial Sciences

Knowledge Representation
In
Geographic Information Systems

Robert Jonathan Corner

This thesis is presented as part of the requirements for
the award of the Degree of Doctor of Philosophy
of the
Curtin University of Technology

December 1999

ABSTRACT

In order to satisfy increasing demand for better, smarter, more flexible land resource information an alternative form of representation is proposed. That representation is to be achieved through the coupling of Expert System methods and Geographic Information Systems. Instead of representing resource information using entities such as soil types, defined by rigid boundaries on a map, a more fluid presentation is proposed. Individual resource attributes will be represented by surfaces that describe their probability of occurrence, at a number of levels, across a landscape. Such flexible representations, which are designed to better capture the mental models behind their creation, are capable of being combined and synthesised to answer a wide range of resource queries.

An investigation of methods of knowledge representation in a number of fields of research, led to the belief that a Bayesian Network provides a representational calculus that is appropriate to the "fuzzy" and imprecise conceptual models used in resource assessment. The fundamental mathematical principles of such networks have been tailored to provide a representation that is in tune with the intuitive processes of a surveyor's thinking.

Software has been written to demonstrate the method and tested on a variety of data sets from Australia and overseas. These tests and demonstrations have used a range of densities of knowledge and range of acuity in evidential data. In general the results accord with the mental models used as drivers. A number of operational facets of the method have been highlighted during these demonstrations and attention has been given to a discussion of them.

ACKNOWLEDGEMENTS

The execution of this research and the writing of this thesis would not have been possible without the assistance of a number of individuals and organisations. The work was carried out whilst the author was a full time employee of CSIRO Land and Water. Thanks are due to that organisation, and in particular the leader of the research group, Dr. Simon Cook, for allowing the research to be used in this way and allowing time for writing up. The development of the Expector software was funded in part through a grant from the Land and Water Resource Research and Development Corporation.

Soil scientists from the Western Australian Department of Agriculture made a valued contribution to the knowledge base used in the trial runs of the Expector software. Thanks go particularly to Gerard Grealish, Geoff Moore, Dr. Bill Verboom and Paul Galloway for acting as 'guinea pigs". In addition, land resource assessment staff from Queensland department of Natural Resources and the Tasmanian Department of Primary Industry and Fisheries assisted with demonstrations of the Expector Software. Thanks are also due to Dr. Paul Gessler, formerly with CSIRO Division of Soils, for permission to use the Sterling data.

Supervision of the work was ably carried out on behalf of the University by Dr. Robert Hickey, and on behalf of CSIRO by Dr. Simon Cook. Thanks are due to both of them for their efforts.

Lastly, thanks are due to three generations of my family. To my father, Dr. W. D. Corner for painstakingly checking my equations for errors, and to my wife and children for enduring the absences occasioned by combining full-time work and part-time study.

Robert J Corner
Perth WA
December 1999

A NOTE ON TERMINOLOGY

Throughout this thesis frequent reference is made to Geographic (or Geographical) Information Systems. The usual abbreviation of this is GIS. The author has followed the example of Bonham-Carter (1994) in using the single abbreviation GIS to describe both a single system and multiple, plural, systems.

In discussing Bayesian networks and their application to mapping, the statement is made (Chapter 7 page 69) that the states of a variable, which is represented as a map, are identical to the classes in a categorical map. Both terms are used in the text. In general if the discussion is concerning the nature and accuracy of map data, the term *class* will be used. If the discussion relates to probabilistic calculus, the term *state* will be used.

The majority of GIS analysis during this research was carried out using either ARC/INFO or ArcView. It is unavoidable in discussions of geographic data analysis that terminology specific to those systems has been used. Specifically, the term *coverage* is used to refer to a vector data set whilst the term *grid* refers to a raster data set. In addition the term *workspace* is used to refer to the computer directory in which the data are stored.

TABLE OF CONTENTS

Chapter **Page**

Chapter **Page**

Chapter **Page**

Chapter **Page**

Chapter **Page**

Chapter **Page**

Chapter **Page**

Chapter **Page**

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Table **Page**

Chapter 1

INTRODUCTION

1.1 The need for natural resource assessment

In the latter half of the twentieth century, there has been an increasing awareness of the fact that the resources of this planet are finite. Man's assault on those resources has increased with population growth, creating a greater demand for food and land on which to grow food (FAO,1999). Whilst the agrarian revolution, with the production by selective breeding of high yielding and disease resistant strains as well as the increasing use of pesticides, has increased grain production on a per hectare basis, the demand for agricultural land continues.

In some environments, such as Australia, there is also a growing awareness that conventional farming systems have an adverse impact on the soil resource on which they depend. In reaction to this, there are moves to ensure that future agricultural practices are truly sustainable. These two considerations, together with many others such as the increased taking of potentially productive land on urban fringes, have created a climate in which information about the basic soil resources of this and other lands is much in demand.

In recent years, there has also been an increase in the availability of computer based models to predict the growth of crops under a range of conditions. Whilst these research tools have been under development, the agricultural engineering industry has produced equipment which is capable of measuring crop yields and varying agronomic inputs at a very fine spatial resolution. This has lead to the development of an emerging discipline know as precision farming, the basic ideal of which is the matching of agricultural inputs to the productive capacity of the land.

The effective use precision farming techniques, crop growth models and the like in the planning and execution of a sustainable agricultural system requires good soil information. Moreover, that soil information must be capable of furnishing quantitative data about the state of such agronomically vital parameters as water holding capacity and availability of nutrients.

From an environmental standpoint, there is a greater understanding of the effects of over-application of agricultural chemicals. Tools are available to model the effects of these chemicals and other pollutants in groundwater. These tools can benefit from the input of spatial representations of variable soil properties.

1.2 The limitation of current mapping methods

Conventional choropleth soil maps do not furnish spatially distributed soil attribute information either readily or accurately. In the past, limitations of representation have meant that such maps delineate relatively broad soil types within which soil attributes can vary considerably. Estimates of soil properties can be derived from existing maps by assigning mean values to existing soil classes. The result is often less than satisfactory since large, generally un-specified, variance within the map units hinders interpretation.

Natural resource surveyors have always recognised the variability of fundamental properties. In mapping a new area, they must delineate a set of boundaries between map units. Before boundaries can be drawn in physical map space the map units must be defined. This requires that the units be defined, in a multidimensional soil attribute space. This problem is then compounded by the relative inflexibility of a paper map and its associated memoir as a representation what is, in fact, a very rich and diverse information stream.

In creating a paper map, surveyors build mental models which associate observable features, such as relief features, soil colour, etc. with variation in less readily observed soil properties. These 'landscape models', which represent the surveyor's knowledge, are recorded as sketches and field notes, as well as being stored as loose concepts in the surveyor's brain. If this knowledge can be in some way formalised, it can be used to generate quantitative maps of soil properties. A suitable medium for that may be through some form of knowledge representation operating within a geographic information system framework.

1.3 Geographic Information Systems

The development of Geographic Information Systems (GIS) enables soil mapping to break free from the constraints of paper maps and atlases. GIS provides a means

whereby raw data can be easily stored and synthesised into maps to suit particular queries or requirements. Early applications of the philosophy, even before the computer technology was mature enough to support such concepts, were to be found in the area of land use planning - supporting critical decisions about the allocation of natural resources (eg. McHarg, 1969).

However, whilst providing excellent repositories for data, GIS can, at worst, merely become the digital equivalent of a grey and uninviting map cabinet. In order to truly capitalise on their abilities to combine, synthesis and manipulate data, it is necessary to imbue them with some degree of intelligence, or at least some means of knowledge representation. If that is effected, such systems can become extremely powerful tools in the service of those charged with mapping our increasingly strained natural resources.

1.4 Representing knowledge

The sort of knowledge that a soil surveyor uses in map making involves a series of inferential processes. These are generally of the kind that associate an outward physical expression, such as a terrain attribute, with an understanding of the physical parameters likely to pertain in that area. This knowledge combines an understanding of the physical processes at work in the landscape with the fruits of experience and detailed observation. It will often be reinforced and refined by additional observations made in trial pits and shallow cores. For example, a suspicion held by a surveyor that mottling is present in the B horizon at a particular location may be based on their knowledge that this occurs in areas which have been subjected to cycles of inundation and drying. A core at that location will confirm or refute the presence of mottles.

The mental process involved is somewhat similar to that employed by a medical practitioner in diagnosing a patient's illness from presented symptoms. A doctor draws upon knowledge of human anatomy, the ills that beset it, and experience gained from the observation of similar cases in the past. Diagnoses are frequently assisted and confirmed by pathological tests.

Medical science was one of the first fields of endeavour to make use of computer 'expert systems' to assist in the diagnostic process. A number of methods were developed to exploit the ability of computers to store data and diagnosis rules. Systems were also devised which possess the ability to learn incrementally. All these systems embody a form of knowledge representation that is a close parallel to that required by land resource assessment.

1.5 Expert and rule based systems

Expert systems are not unique to medical diagnosis and have been developed for use in other fields such as financial management, production scheduling and geological prospecting. A common theme of such systems is their use to provide decision making assistance. "Does this child have a particular medical condition?" or "Should this client buy BHP shares?" are questions which might be asked. Such systems are generally non-spatial in their application. Even geological questions such as "Is this location prospective for gold?" are frequently treated by such systems as pertaining to isolated points rather than taken in their full spatial context.

Some of these systems assist in the "diagnosis" by means of a series of questions and answers. The choice of question and the choice of line of reasoning are often decided by the answer to preceding questions. Other systems are constructed to take all available evidence on board at once.

Different situations require different system architectures. A system designed to assist less well trained operatives by allowing them access to an 'expert oracle' will clearly differ from one designed to allow a highly skilled professional to codify their knowledge and decision making processes.

Two principle schemes of inference are generally found in such systems: Boolean logic and probabilistic inference. Within the realm of probabilistic inference, there are a number of representational calculi. These seek to represent knowledge and uncertainty and to provide means by which these concepts can be combined and reasoned judgements made. Interestingly, many of them have their ultimate origins at the gaming table (Bellhouse, 1993).

1.6 Aims of this thesis

The research reported in this thesis is essentially an investigation of the means whereby a GIS based method of knowledge representation and data combination may be applied to natural resource mapping in general, and soil mapping in particular. The mapping method and associated software that have been developed are known as Expector, the choice of name reflecting the fact that it is used to map, in quantitative form, a resource surveyor's expectations.

The thesis commences with a detailed analysis of current methods of natural resource mapping and the limitations imposed on the resulting maps by conventional means of representation. An alternative form of representation is suggested.

There then follows a review of the mathematical representation of knowledge, paying particular attention to the calculus of probability and belief. This leads to a review of expert systems, particularly those that use a probabilistic representation of knowledge, and an introduction to Causal Probabilistic Networks. Since the implementation of any such method for natural resource mapping is to be through the medium of a GIS, Chapter 5 is devoted to a discussion of that technology. Chapter 5 also includes a review of the efforts of other workers in establishing linkages between expert systems and GIS.

The thesis then turns to a study of the potential for quantifying the soil mapping process. This includes an examination of the form taken by the knowledge used in that process. Current trends in the automation of that process are examined and the influences of currently installed technology on system design are considered.

A description and definition of the functional stages of a quantitative soil mapping process sets a 'blueprint' for a new method to be known as the Expector method. The method is envisaged as comprising software for knowledge manipulation and editing which can operate synergistically with GIS. Synergistic links can be enabled by the development, within the framework of individual proprietary GIS, of data preparation and combination tools. An outline definition of the structure of the method discusses the reasoning behind the choice of programming language and hardware components of the implementation.

The development of expert system algorithms for the Expector method required the adaptation of the general principles of Causal Probabilistic Networks to suit the particular need of soil mapping, especially to cope with uncertain evidence. The discussion of this adaptation leads to a more detailed explanation of the tasks to be performed within the co-operating GIS. There then follows a description of the components and operation of the software. A copy of the Expector software and its user documentation comprise Appendix B which is attached to this thesis as a CD-ROM.

During the development of Expector, several experimental calculations were made with datasets covering areas ranging from a few hectares to whole catchments. The work on two sites, one at each end of this spatial scale, is reported in some detail and includes a discussion of the accuracy of the resulting maps. This is put in context by a discussion and examination of the accuracy of existing mapping products. Further examples of the application of Expector in natural resource mapping are then provided and are followed by an example of the use of the method to predict agricultural yield as an aid to precision agriculture.

Both the developmental data processing and subsequent experience with the method highlighted a number of potential operational problems. The majority of these can be controlled by the exercise of caution by an informed user. The thesis concludes with a discussion of some of these problems. Solutions to some are offered and areas are indicated where further work may fruitfully be conducted.

Chapter 2

THE NATURAL RESOURCE MAPPING PROCESS

Traditional natural resource and soil mapping methods have reached a high level of sophistication. They use both conceptual and statistical models to represent landscapes. However, choropleth map representations impose some constraints on the flexibility of the outputs of those models. Alternative forms of representation may be better able to convey the rich information which the models embody.

2.1 The soil survey process

A soil surveyor producing a map using traditional methods follows well-established procedures, developed and documented over the past fifty or more years (eg. Soil Survey Staff, 1993; FAO, 1979; Dent and Young, 1981.) A typical survey begins with an inspection of the area, although this may be preceded by reference to existing sources of information. Unless the area is in a remote location that has never been subject to human activity, maps and reports of topography, geology etc. will be consulted. A satellite image will be consulted, at least as a guide to land use, as will any available aerial photographs.

Viewing aerial photographs in stereo is a valuable aid to constructing a mental model of the landscape. If the site is in a particularly remote area or in another country, this mental model may quite possibly be "roughed out" from photographs before the surveyor reaches the site. The development of that mental model draws not only on the available information, but also on the surveyor's prior experience and training. There then follows a period of fieldwork in which the surveyor is constantly testing and refining the mental model by field observation, adding additional information, exceptions and qualifications as they appear.

Finally, a map begins to take shape, often using the aerial photographs as a base. These are used partly to geo-reference points and lines, but also for the information they contain about subtle landform changes and land textures. The surveyor's mental model is now in a mature state and includes many exceptions and special cases. In traditional soil mapping, this complex multilevel model then has to be condensed

into a single-layer map, generally composed of polygons of supposedly homogenous soil types. This traditional cartographic representation has a considerable constraining effect on the map. Often the greatest insights in the map are to be found in the small text of the legend and in the accompanying memoir, rather than in the cartographic product itself. It is only in this textual form that the surveyor's knowledge of the variability of the soil resource can be expressed.

2.2 Problems and models in natural resources mapping

Soil surveying is just one part of the broader field of natural resource mapping and land evaluation. The process of mapping objects and phenomena in the natural world requires a considerable amount of abstraction and generalisation. Natural systems are extraordinarily complex at a number of levels. For example, they are unstable over time.

The rate of change of attributes in natural systems varies from long term geological and geomorphic processes such as isostasy - through cyclical effects due to seasons - to short term chaotic effects caused by weather. There are many processes occurring at any one time at any particular location in a landscape, and any one of them may have some bearing on the physical attributes present. A map of any of these attributes, whether represented as a traditional paper product or as a data layer in a GIS can only be a "snapshot" of the condition of the attribute at a particular time. As such, it is an abstraction, not only of reality, but of a particular reality from a constantly changing spectrum.

Some attributes are more stable than others and some are more easily mapped than others. There are interesting trade-offs between ease of mapping and level of abstraction and utility. For example, a map of the distribution of some large fauna such as Western Red kangaroos in a pastoral area may be easily made with the aid of aerial photography. Although only truly accurate for the time of acquisition of the photography, it nevertheless gives a good indication of the likely distribution of kangaroo under similar seasonal and climatic conditions, as well as an indication, by scaling up, of population density on a more regional basis. A similarly easily acquired map of tree cover and distribution would have greater temporal stability but may be less accurate in terms of species identification.

The characteristics that are pertinent to evaluation of land for agricultural use vary from relatively easily mapped attributes such as terrain features to more difficult ones such as sub-surface pH. A natural resource surveyor uses models to link the readily observed attributes to those that are more obscure. These models are either conceptual in nature (Hewitt, 1993) or statistical (Gessler et al., 1995). Both types of model represent an abstraction and simplification of the true relationship between the perceptible and obscure attributes.

With both types of model it is desirable that the attributes to be mapped be semi-permanent. That is to say, they should be attributes that have resulted from the action of longer-term soil and landscape formation processes, rather than the products of shorter-term land management effects. The natural processes involved will have interactions. Whether it is possible to ever have complete knowledge of the interactions between natural processes is debatable, but it may be possible to know the state of a process at any one time, especially if it is in equilibrium. Models may, therefore, be regarded as being only abstractions of knowledge.

In addition to conceptual and statistical models a third class of model is found in the soil survey - those offered by geo-statistics. These methods, such as kriging, are in the main designed as spatial predictors of values of attributes at un-known locations, based upon the spatial distribution of values of that attribute at known locations. They are generally used as interpolation methods and do not, therefore, fit the general soil survey model of using readily observed features to infer less visible characteristics. Recent developments of basic geo-statistics, such as co-kriging and regression kriging, do exploit the relationship between spatially distributed variables using techniques which have parallels in statistical models (Odeh et al, 1995).

2.2.1 Statistical models

By comparison with conceptual models, statistical models are both simpler to derive and easier to represent mathematically. In the context of soil mapping, they will generally have been developed by analysis of sample data. Whilst these data are not necessarily from the area being mapped, they will at least be from a landscape where similar geomorphic and pedo-genetic processes are believed to take place. Their ability to describe relationships is, theoretically, limited by a number of factors.

These may be explored by considering the statistical model as a polynomial equation relating one attribute to another. Figure 2.1 shows a hypothetical linear fit through an imaginary data set for two attributes.

In this figure, the relationship between the attributes is expressed by the solid line. If we consider a case in which the points represent the entire universe of both the predictor and predicted attribute, then the relationship between them could be described exactly by whatever complex polynomial follows the dotted line. This would be a true numerical representation of that inter-relationship. In reality, the points shown will represent only a very small part of the entire universe of possible points. In that case, even a line which fitted those points exactly will only be an estimate, since the location in attribute space of all other points is unknown.

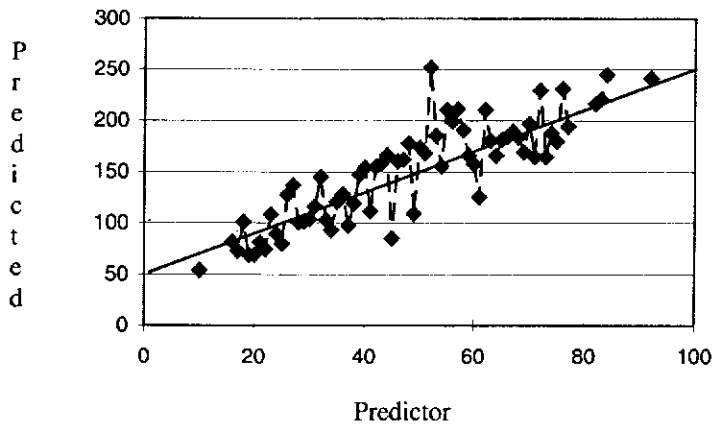

Figure 2.1 Regression mapping of a hypothetical attribute

It is, therefore, axiomatic that when predicting the value of one attribute from another there will be unknown values in at least one of those attributes. In addition, there will be error in the measurement of both the predictor and predicted attribute. Pragmatically, the polynomial fitted through any data set is unlikely to be of high order, although it may include trigonometric or logarithmic functions. These may be generated by optimisation routines which search data for interrelationships, but in practice are more likely to be determined by an analyst. The analyst may superimpose some pre-conception – based on an expectation of the process that links the predictor with the predicted attribute.

The strength or otherwise of the prediction is then a function of:-

a) The degree to which the sample used truly represents all possible combinations of predictor and predicted.
b) The error inherent in the measurement of the sample.
c) The degree to which the chosen polynomial fits the sample.

The description above is a considerable simplification of real situations. In reality, it is very rare for two attributes to be linked in such a way as to exclude interactions with other spatially varying attributes. A variety of methods exist to handle such situations, and descriptions of a number of these may be found in Burrough (1986). In essence, however, they all suffer to some degree from the limitations described above.

2.2.2 Conceptual models

Conceptual models of landscape development are 'fuzzy' and imprecise. They are frequently informal and even when apparently formal often contain subtle modifiers to accommodate local knowledge or special conditions. They are, however, immensely useful to a skilled surveyor in mapping natural resource attributes. This is in part due to their informal nature which allows exceptions and modifiers, and in part due to the fact that they are processed and synthesised using natural language and the human brain, rather than mathematics and a computer.

The processes used in traditional soil mapping are well represented by conceptual models. Although they do not embody all possible knowledge, they do encapsulate the surveyor's belief in the way the landscape has developed and in the likelihood of encountering certain attributes in certain combinations. Such models are rarely crisply defined. In a review of the development of soil survey techniques since the 1960s, Burrough et al. (1997) discuss a continuum of models going from the double crisp (crisp classes in attribute space linked to crisp classes in geographical space), through crisp-continuous to double continuous.

In general, the model used by the surveyor will be more or less fuzzy, which suggests intuitively that probability may be used to represent it mathematically.

Unfortunately, the conventional output of this fuzzy conceptual model is generally a choropleth map. This imposes some limitations on its subsequent interpretation.

2.3 The limitations of choropleth maps

Since soil variation is known to be effectively continuous, the hard boundaries in a traditional soil map are more an imposition of the techniques used than of reality. The nature of the theme mapped, usually soil type, is such that membership of a particular soil type or class encompasses a range of physical and chemical properties. Whilst occasional hard distinctions do occur, it is more usual for properties to grade between points in the landscape. This fuzzy-ness in class membership is nicely paralleled by the fuzzy-ness of most conceptual models of landscape development, but is not represented by the hard boundaries between units on the map

This is not necessarily a limitation unless the map is used for purposes such as land management at a detailed level. It then manifests itself by a failure to provide the level of resolution required. There has been a loss of information, which can obstruct unambiguous interpretation of the soil map.

We may take as an example of this ambiguity the selection of suitable areas to establish a new crop type. The particular soil factors appreciated by the new crop will be known from elsewhere and could be promulgated as a rather nebulous piece of information such as a preference for *"free draining slightly acid sands"*.

It would be possible to determine areas that fit this description by reference to a traditional soil map. However, they would include areas which may be outside the comfortable range of conditions for the crop, simply because the description of requirements is broad, as are the descriptions of the soil classes in the traditional map.

The new crop's requirements could be defined with more precision by reference to detailed threshold values for a range of soil attributes. The description *"free draining slightly acid sands"* could be redefined in terms of at least three attributes. These are moisture retention capability, pH and particle size. There will no doubt be others.

Access to maps of these soil properties would, therefore, allow us to select not only those areas that exactly matched the specifications, but also those where one or more of the attributes was sub-optimal. In cases where that attribute was capable of manipulation by management techniques, the range over which the new crop was viable could be considerably increased.

2.4 Representing soil attributes

The methods of data storage and presentation offered to us by GIS enable a soil attribute or other similar natural parameter to be represented as a continuous surface. Such a map would be of considerable use, but would still suffer from the limitation that it contains no statement about its precision. Indeed, it would invite an assumption that it was a definitive statement. If this map results from the application of a fuzzy conceptual model or of a statistical model with in-built and estimable error, the use of a second surface as a map of precision is a possible solution.

Bouma (1989) suggests that a statement that a particular soil property has a particular value, or lies within a particular range, may be of less use (from an environmental and legislative perspective) than a statement of the probability that the value lies within a particular range. However, a single statement of precision is not without its drawbacks.

If we assign an absolute value to an attribute for a particular surface segment or raster cell, together with a single value as an expectation of accuracy, we are still not providing information about the distribution of the implied uncertainty. However, a statement that there is a 75 percent chance that an attribute exceeds a value of four, but is less than five, can be accompanied by probabilities that it exceeds five and that it is less than four. This is of more use to a decision-maker than the simple statement that the attribute value is four, with a probability of 75 percent. The latter method tells us nothing about the distribution of the variable in attribute space.

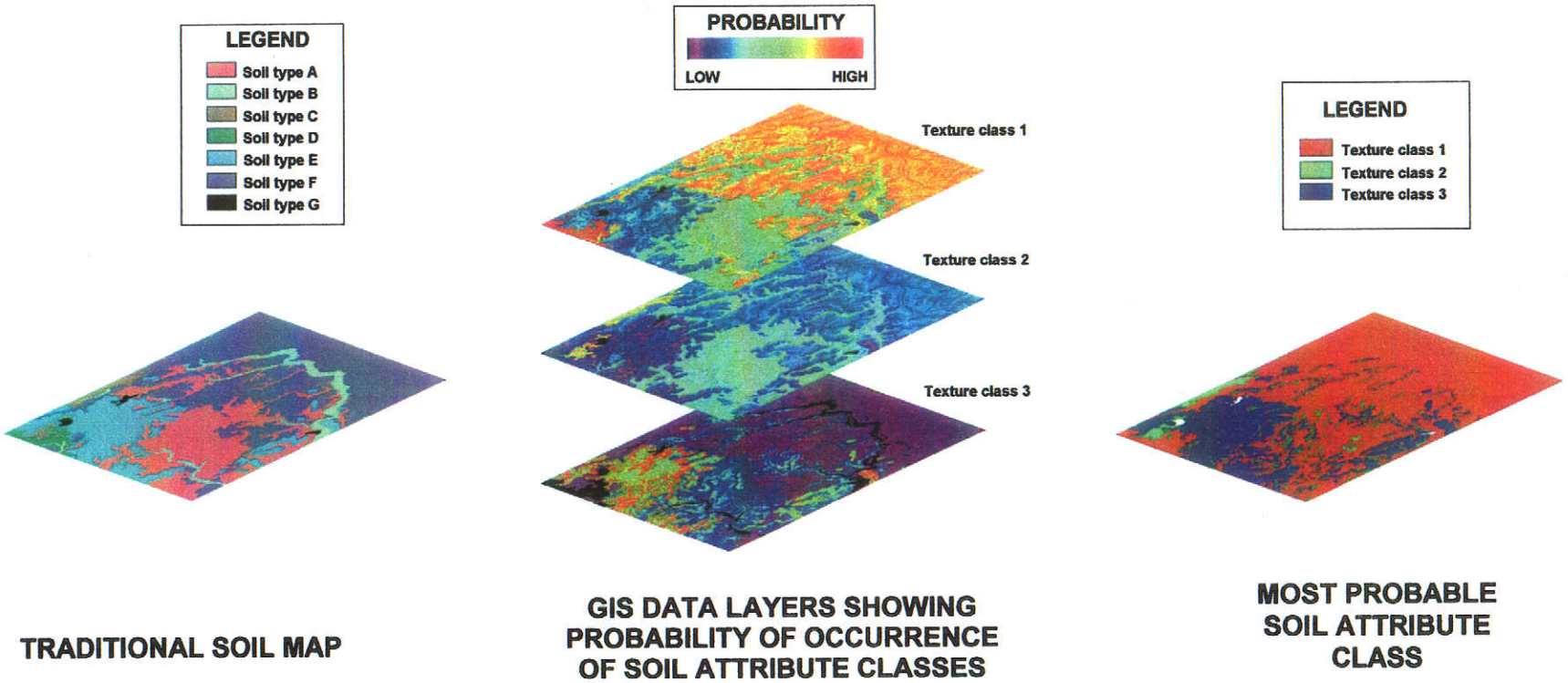

Figure 2.2 Different ways of presenting soil information

Paradoxically, this suggests that even when working with soil properties rather than with soil classes, it is maybe convenient to divide them into classes. This allows us to assign a probability of membership to those classes. Our ideal map has now migrated from one restricted by technology to polygons describing quasi-homogenous soil types, to a multi-layer representation. Each layer will represent our expectation of the occurrence of a particular class of one of several soil attributes. Figure 2.2 illustrates the traditional and proposed concepts.

2.5 Summary

Soil mapping is a subset of the wider domain of natural resource mapping, both of which can make effective use of statistical and conceptual models. Statistical models have a number of limitations. Although conceptual models are also limited by virtue of being abstractions, they have considerable flexibility and are intuitively appealing to the surveyor.

The output from models has traditionally been limited by their representation as choropleth paper maps. The use of GIS offers the opportunity to represent the underlying soil information in novel ways. In addition to representing spatially variable soil attributes, it is also possible to store information about the spatially variable accuracy of that representation. It is suggested that the traditional (choropleth) soil map be replaced by a multi-layer representation of the likelihood of occurrence of particular soil attributes.

Conceptual models combine a number of competing ideas or threads of evidence to produce a 'most likely' scenario. In order to capitalise on the power of computers and GIS, some means is required of representing the knowledge embodied in this process. The next chapter looks at some ways of mathematically representing knowledge.

Chapter 3

THE MATHEMATICAL REPRESENTATION OF KNOWLEDGE

Knowledge is a philosophical concept that is an inherent part of human decision making processes. The internal representation of knowledge in the human brain is somewhat informal, loosely defined and incorporates 'fuzzy' concepts such as belief, expectation and an understanding of fairness and balance. In order to use digital computers as decision-making aids in the natural resource mapping process, some scheme is required by which such knowledge may be represented and manipulated mathematically.

The field of research known as Artificial Intelligence (AI) aims to produce systems which emulate the human brain. Expert or knowledge-based systems form a subset of AI which endeavours to capture the reasoning behind decisions made by experts in particular domains of knowledge. Related research on topics such as data mining seeks to develop knowledge-like relationships from the analysis of large data sets. These endeavours are all relevant to the use of knowledge in geographic data analysis.

Knowledge representation schemes typically use either logic or probability (and the related concept of belief) as representational frameworks. Either method is a contender as a means of representing the natural resource mapping process.

This chapter provides first a review of AI in general, and then a discussion of some of the different types of expert systems. It then briefly examines the use of logic in knowledge based schemes. Since many knowledge based systems use probability for knowledge or uncertainty representation, a discussion is provided on the origins of probability theory and Bayes' rule. A simple example, drawn from the literature, illustrates the use of Bayes' rule for updating. The chapter ends with a brief discussion of Dempster-Shafer theory.

3.1 Artificial Intelligence

3.1.1 The goals of AI

Artificial Intelligence (AI) is a branch of research that seeks to develop machines which imitate as much as is possible of human mental activity. It has been argued (Penrose, 1989) that this goal is impossible, since there are aspects of human consciousness, particularly those to do with feelings, that cannot be replicated by a machine. In other words, the machine can never be said to have true understanding. However, from the point of view of applying artificial intelligence to the solution of geographical problems, this may not be a limiting factor.

One particular branch of AI that is of relevance to geographic data analysis is that of expert or knowledge-based systems. Bender (1996, p. 18) discusses the breadth of scope of AI and suggests that it includes (as goals) systems which are capable of reasoning, planning and learning. He then defines an expert system, for a particular field, as exhibiting abilities in that field, accepting input about a particular problem, and delivering advice and actions using domain specific knowledge.

In reviewing the first 25 years of AI research, Duda and Shortliffe (1983) cite two of the goals of AI research as being the development of cognitive models of intelligent behaviour and the development of computer programs to solve problems normally thought to require human intelligence. They then define expert systems as "a class of AI computer programs intended to act as consultants for decision making". This and Bender's (1996) definition of expert systems indicate that their capabilities fall short of full artificial intelligence.

The field of AI and knowledge based systems has a rich and developing literature, much of which is beyond the scope of this thesis. Indeed, some developments have occurred so recently as to have been contemporary with the research reported here (e.g. Heckerman, 1997). This review is confined to methods and systems that have an application in natural resources mapping. An extensive review of mathematical methods in AI is provided by Bender (1996) and numerous compilations of papers such as those edited by Shafer and Pearl (1990) and Garcia and Chien (1992) cover the field in great detail.

3.1.2 Expert systems as consultants

Although Duda and Shortliffe (1983) suggested that none of the goals of AI had been fully achieved, they expressed the opinion that some expert systems designed as consultation programs may have reached performance levels similar to those of human experts. These successes seem to have been achieved by extensive effort to formalise and organise large amounts of knowledge.

In most cases, the knowledge base is described as being a substantial collection of semi-organised, often subjective, information which may even be incomplete. Duda and Shortliffe (1983) suggest that encoding this in a program renders that knowledge explicit and, therefore, more uniformly applicable for decision making. They make the interesting observation that it is easier to emulate 'expert' problem solving than to devise programs to make common sense deductions or to learn speech and language in the same way that a child does. This lack of "common sense" is suggested by the authors as being one of the reasons why computer expert systems fall short of human experts. They simply lack the necessary background and context.

Consultant expert systems have been developed in a number of fields. Writing nearly two decades ago, Duda and Gaschnig (1981) listed 26 systems, all in a mature state of development. These primarily cover the fields of medicine, engineering, geology, chemistry and the design and analysis of electronics. More recent works (e.g. Garcia and Chien, 1992) offer additional examples from fields as diverse as computer configuration and satellite failure diagnosis. In many cases, problem-solving strategies transcend discipline boundaries. This is also true of human problem solving.

3.1.3 Types of expert systems

Expert systems themselves have many subdivisions. Bender (1996, p. 19), discussing expert system shells (tools for the construction of expert systems), suggests the hierarchy of systems shown in Table 3.1. Those that have found their way into natural resource evaluation typically fall into the "Reasoning" class of this hierarchy.

Class of system	Sub-class	Examples of methods
Reasoning	Qualitative	Logic
		Production rules
		Semantic nets
	Quantitative	Bayesian
		Fuzzy logic
Hybrid	Uses more than one method	
Pattern classification	Rule extraction	
	Decision trees	
	Neural networks	Hopfield-like
		Feed forward

Table 3.1 A hierarchy for expert systems (after Bender (1996))

3.1.4 Knowledge bases and inference engines

Buchanan and Shortliffe (1987) describe expert systems as having two essential components. Those are a Knowledge Base and an Inference Engine. The Knowledge Base is a collection of facts and associations about the relevant subject area, often represented as a set of rules. The Inference Engine is an interpreter which uses the knowledge base to solve the problem (Davis, 1987). It is within the inference engine that such tasks as dealing with uncertainty and combination of evidence are performed.

3.2 Logic as a knowledge representation mechanism

The use of logic to represent knowledge has an intuitive appeal and, in the case of complete and all encompassing knowledge, may even be appropriate. Mathematical logic is founded on the two concepts TRUE and FALSE and makes use of a number of operators familiar to users of procedural programming languages. These include constructs such as IF.....THEN.....ELSE, AND, OR, and NOT. Bender (1996, Ch. 3) provides a good tutorial on the use of logic in AI.

There are numerous examples of logic based expert systems. They cover a range of fields as diverse as computer configuration (Barker and O'Connor, 1989) and land evaluation (Rossiter, 1990). Such systems often involve the user in a question and answer dialogue using the answers to navigate a decision tree. In this regard, they

have been likened to the taxonomic keys used manually for tasks such as plant identification (Bender, 1996).

A simple logical set of rules can both represent knowledge and provide an inference engine for navigating those rules. Navigation may either take the form of a rigid progress through the rules, as illustrated by a plant identification key, or use some less exact construct such as certainty factors (Shortliffe, 1974). Such constructs generally use probabilistic methods as a means of combining and manipulating knowledge.

3.3 Probability as a knowledge representation mechanism

3.3.1 Origins of probability theory

There is a popular belief that probability theory originated from gambling and games of chance. In a discussion of the role played by roguery in the history of probability, Bellhouse (1993) notes that the proof of that suggestion is often cited as the fact that many of the early probability theorists (from the mid 17th century onwards) used analysis of games of chance to formulate and describe their ideas. There is little to suggest that a knowledge of probability was inherent in the *design* of games of chance, many of which have been played since antiquity. A useful discussion on the origins of probability in philosophical thinking is provided by Hacking (1975). He traces the beginnings of modern probability from Pascal's responses in the 1650s to questions from a member of the French nobility concerning games of chance. Pascal then continued to elaborate on the ideas of chance in his famous philosophical "wager" on the existence (or otherwise) of God.

The development of probability can then be traced from these beginnings through to modern concepts of epistemic probability which have to do with knowledge and evidence. According to Hacking (1975), a numerical scale of probability was first enunciated by Liebnitz in a 1665 paper. Liebnitz's interest in such matters derived from civil law which, according to Hacking (1975), shares with epistemic probability the fact that it must distinguish between testimony and circumstance.

Further interesting developments chronicled by Hacking (1975) include the use of the term "expectation" by Huygens, first published in a work of 1657. We are

clearly advancing now towards concepts which we will need when considering the combination of multiple threads of evidence. Pascal and Huygens also worked on data combination. In his extensive work, Hacking (1975) goes on to discuss the development, principally in Holland and the UK, of mortality statistics and annuity tables. These introduce the concept of probability distributions that are central to the use of probability as a means of representing knowledge.

3.3.2 Bernoulli and the first limit theorem

Although it is the work of Thomas Bayes (1763) (reprinted with commentary 1958) on which the majority of reasoning used in modern probabilistic expert systems is based, Bayes' work drew on that of Bernoulli. In "Ars conjectandi", published posthumously in 1713, Bernoulli expounded what has been described as the first limit theorem of probability. In addition, according to Hacking (1975), Bernoulli was the last person before modern times to seriously discuss the concept that probability may be non-additive. He put forward ideas such as the suggestion that the probability of an event and of its converse may both exceed 0.5. These have re-emerged more recently in the thinking of Dempster (1967) and Shafer (1976). In addition, Bernoulli expanded on ideas of subjective probability which are at the heart of some of today's evidence combination methods.

Bernoulli's limit theorem states that in a set of n trials, on a chance set-up, the probability of achieving a 'success' (which has a prior probability of P), will tend to P as the number of trials increases. This is sometimes described as the First Law of Large Numbers (von Mises, 1964). From this basis, Bernoulli went on to introduce the concept of conditional probability.

It was then left to Bayes to determine mathematical methods of working with conditional probabilities. The contemporary reader can turn to more recent and accessible texts such as Montgomery and Runger (1994, p. 78 et seq) for a definition of the concept of conditional probability. They illustrate that the probability of an event P(A) and the conditional probability P(A|B) (the probability of A given that B has happened) are really the probabilities of the same event, computed under two different states of knowledge.

3.3.3 Additive probabilities

Since Bernoulli's time, there has been a convention that probabilities are additive, and that the probability of an event and its converse sum to unity. That is:-

$$\mathbf{P(A) + P(\overline{A}) = 1} \tag{3.1}$$

By extension of this it can be assumed that, if there exists a set $\boldsymbol{A}$ of mutually exclusive, and exhaustive events A_1, A_2, to A_N, then :-

$$\mathbf{P(\boldsymbol{A}) = \sum_1^N A} \tag{3.2}$$

This is sometimes referred to as the Law of Total Probability (Montgomery and Runger, 1994).

3.3.4 Bayesian methods and conditional probability

The work of Bayes is fundamental to problems of inference. The problem that he set out to solve is stated in his 1763 paper (Bayes, 1763; Bayes, 1958) as follows:-

> "*Given,* the number of times in which an unknown event has happened and failed. *Required,* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability"

Following this statement of problem are a number of definitions. According to Hacking (1965), these are not necessarily the first such definitions but they do describe the computational context of Bayes' work. The most important of these definitions is the following statement:-

> "Events are independent when the happening of any one of them does neither increase nor abate the probability of the rest."

Other important points are that Bayes considers the word *chance* to equate with *probability*, and defines the probability of any event as:-

> "the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening."

In his analysis of Bayes' work, Hacking (1965, p193) refers to this definition as being that of a "fair bet".

In his solution of the problem Bayes makes a number of propositions. One of those is highly relevant to the combination of probabilities and is stated succinctly as a corollary that:-

> "if of two subsequent events the probability of the first be **a/n** and the probability of the both together be **p/n**, then the probability of the second on supposition that the first happens is **p/a**"

The proof of this is based on frequency counts and, if restated in modern notation, gives us the now familiar definition of conditional probability. It can be seen that if we write **P(A) = a/n** , **P(B,A) = p/n** and **P(B|A)** as **p/a** then :-

$$\mathbf{P(B \mid A) = \frac{P(A,B)}{P(A)}} \tag{3.3}$$

Where **P(A,B)** is the joint probability of **A** and **B** together. This equation is now generally referred to as Bayes' rule (von Mises, 1964).

3.3.5 Twentieth century readings of Bayes' work

To modern eyes, Bayes' work makes hard reading and it is useful to turn to modern interpretations for the complete story. Both von Mises (1964) and Hacking (1965) provide discussions. Hacking (1965) is more accessible and uses Bayes' own example of balls thrown upon a perfectly level table. It is noted that, although Bayes was the first to use this proof as a basis for statistical inference, he may not have been the first to see it. According to Hacking, what Bayes demonstrated was that the probability of event A happening, given events B and C, is proportional to the product of the probability of B happening given A and C and the probability of A happening given C. That is :-

$$\mathbf{P(A \mid BC) \propto P(B \mid AC) . P(A \mid C)} \tag{3.4}$$

This can be shown to hold in a number of cases, perhaps the most relevant being that in which **A**, **B** and **C** are regarded as propositions and conditional probability is regarded as the degree to which one proposition supports another.

Hacking (1965) puts this into context by suggesting that this be viewed as a case in which **A** is a hypothesis, **C** is initial knowledge for assessing that hypothesis and **B** is new data - for example an experimental result. We now read **P(A|BC)** as being the posterior support for **A** (in light of the new experiment). Similarly **P(A|C)** is the prior support for **A**, that is in the light of **C**, but before learning of **B**. **P(B|AC)** is the support for the result **B** based on the hypothesis and our initial data **C**. In other words, this is the likelihood of getting result **B** if hypothesis **A** is true. The importance of this to problems involving the combination of geographical data is that it offers a method of updating probabilities in the light of new evidence.

A more readily comprehensible exposition of Bayes' theorem, together with an example drawn from the geo-sciences, can be found in Davis (1986). Starting from Equation 3.3 above we invert the definition of conditional probability to give:-

$$\mathbf{P(A,B) = P(B \mid A) \, . \, P(A)} \tag{3.5}$$

Similarly we construct the relationship: -

$$\mathbf{P(B,A) = P(A \mid B) \, . \, P(B).} \tag{3.6}$$

Since by definition:-

$$\mathbf{P(A,B) = P(B,A)} \tag{3.7}$$

We can now write:-

$$\mathbf{P(B \mid A)P(A) = P(A \mid B)P(B)} \tag{3.8}$$

which can in turn be rewritten as:-

$$P(B \mid A) = P(A \mid B)\frac{P(B)}{P(A)} \tag{3.9}$$

This gives us a useful tool for inverting conditional probabilities. Under the assumption that a mutually inclusive and exhaustive set of events B_i exists, which are conditionally related to A we can rewrite the total probability rule (Equation 3.2) as:-

$$P(A) = P(A,B_1) + P(A,B_2) + ... + P(A,B_i) \tag{3.10}$$

Using this and the definition of conditional probability from Equation 3.4 we can generalise Equation 3.9 as the most generally quoted version of Bayes' theorem:-

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)} \tag{3.11}$$

3.3.6 A geological example of updating

Davis (1986) goes on to illustrate the use of Bayes' theorem by considering a case in which a previously unknown marine fossil species has been found in a stream bed at a point below the confluence of two tributaries. The problem is to determine which of the two basins will be more fruitful in the search for further remains. The area of each stream basin is known and from these areas and their sum, the probability that the fossil came from each can be calculated. Naturally the larger basin will have the higher probability. Davis' example calculated these as being 0.64 for the larger basin B_1 and 0.36 for the smaller basin B_2.

Further evidence is available in the form of a geological map, which indicates that 35 percent of the sediments in the larger basin are marine, whereas 80 percent of the rocks in the smaller basin are marine. Assuming these two basins as being the only sources of that fossil, this may be taken as the conditional probability that a fossil from basin B_1 is marine $\{P(A|B_1) = 0.35\}$ and the conditional probability that a fossil from basin B_2 is marine $\{P(A|B_2) = 0.8\}$. Equation 3.11 therefore reduces to a two case situation, in which the conditional probability of the fossil originating from basin B_1, $P(B_1|A)$ is given by:-

$$P(B_1 \mid A) = \frac{P(A \mid B_1)P(B_1)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2)} \tag{3.12}$$

Solving this equation gives us a conditional probability of 0.44 that the fossil came from the larger basin. An analogous computation gives a figure of 0.56 for the probability of it having come from the smaller basin. This is in contrast to the original hypothesis, which was based solely on the area of the basins and shows the effect of updating that hypothesis in the light of new or additional evidence.

3.3.7 Bayesian networks

Most problems in real life are more complex than the simple examples given here. Multiple pieces of evidence support or refute, intermediate pieces of evidence which in their turn support or refute the final hypothesis. Bayesian methods enable us to propagate probabilities though such complex networks. A further discussion of Bayesian networks is provided in Chapter 4.

3.4 Probability and belief

Discussions on the origins of modern probability theory such as Hacking (1975) and Bellhouse (1993) trace the terminology of probability through terms such as *chance, belief and expectation.* We have seen that Bayes regarded probability as being the same as chance and that Huygens introduced the term expectation. The term belief is linguistically linked to expectation. All these terms have mathematical as well as linguistic meanings and are less interchangeable in mathematics than they are in language. An important distinction is that between probability and belief. This is of particular relevance in systems that use subjective or user supplied probabilities. The chief difference between probability and belief is that whilst probability is generally regarded as operating in exclusive and exhaustive conditions, belief is not exclusive. This, by removing the additive constraints imposed on probability, leads to a better representation of ignorance and is exploited in a belief calculus know as Dempster-Shafer theory.

3.4.1 Dempster-Shafer theory – belief and ignorance

Shafer (1976) gives an example of the representation of ignorance. If we consider the possibility that life exists in the vicinity of a particular star, we can construct two

possibilities, Θ_1 that life exists and Θ_2 that it does not. Under the assumption that they are the only two possibilities, we can equate their sum to unity. It is therefore possible to mathematically describe the state of complete ignorance as to whether life exists by assigning setting both Θ_1 and Θ_2 to 0.5. For a simple situation this is adequate.

However, Shafer (1976) argues that, if we complicate the situation by introducing a consideration of whether planets necessary to support life exist near that star, we now have a set of possibilities with three members. They are ζ_1, that life exists near the star, ζ_2, that there are planets but no life, and ζ_3, that there are no planets and hence no life. Considering the relationship to the earlier situation it can be seen that Θ_1 equates to ζ_1 and that Θ_2 equates to $\zeta_2 + \zeta_3$. Using Bayesian calculus will cause an inconsistency.

Specifically, continuing to represent ignorance about the first situation as $\Theta_1 = \Theta_2 = 0.5$, is at odds with a statement of ignorance about the new situation which states that $\zeta_1 + \zeta_2 + \zeta_3 = 1$, since if the three possibilities ζ are equal, $\zeta_1 = \zeta_2 = \zeta_3 = 0.3$. It could be argued that this is a case more of the problem being incorrectly specified than a serious computational problem, but it does raise the question of whether the sum of 'beliefs' requires to be unity.

Dempster-Shafer theory, which incorporates Dempster's Rule of Combinations, has been developed as a calculus for combining belief functions under the assumption that belief need not sum to unity. It is expounded in Shafer (1976) and many subsequent texts including Bender (1996).

3.5 Summary

The subset of Artificial Intelligence research known as knowledge-based or expert systems uses mathematical representations of knowledge. These two principal knowledge representation schemes use logic and probability. Most probabilistic schemes use Bayes' rule or modifications thereof to manipulate and combine representations of knowledge. Whilst other combination calculi exist, that of Bayes'

is currently prevalent and its derivation has been covered extensively here. Examples of expert systems using both logic and Bayesian networks are provided in Chapter 4.

Chapter 4

SOME EXAMPLES OF EXPERT SYSTEMS

The knowledge representation schemes based on logic and probability reviewed in the previous chapter have been embodied in numerous expert systems. Some of these are from the field of medical diagnosis; some are from geological prospecting and land evaluation. Many more examples exist from a diverse range of activities where attempts have been made to formalise and automate decision making. Since this research is directed towards knowledge representations for natural resource mapping, examples from the geosciences will be examined in some detail. However, there are close parallels between decision strategies in land resource assessment and in medical diagnosis. Agterberg (1989), in discussing this similarity, notes that both emerged at about the same time; the one with Agricola's 'De Re Metalica" of 1556 and the other with Paracelsus' sixteenth century theory of similarity (reprinted in Paracelsus (1967)). Both disciplines require the reading of signs which point with a varying degree of certainty to, respectively, 'pay-dirt' or a successful and correct diagnosis.

This chapter commences by looking at examples of expert systems that use logic, with a specific example of some expert system software for land evaluation. It continues with a look at some early probabilistic expert systems and reviews in detail the PROSPECTOR mineral exploration system.

The development of general schemes of Bayesian networks is then considered and the chapter concludes by discussing a medical diagnosis program which uses Bayesian networks.

4.1 Knowledge based systems using logic

The use of fuzzy and probabilistic systems of knowledge representation has somewhat overshadowed logical inference as an automated decision making strategy. This is partly due to the fact that complex real world problems are not often definable in 'crisp' logical decision trees.

There are examples in the literature of several 'toy' systems. Duda and Gaschnig (1981), in an article discussing the theory of expert systems, published code for a system to identify a number of mammals from observed characteristics. In the field of land evaluation, the Automated Land Evaluation System (ALES) (Rossiter, 1990; 1998) is a complex system for evaluating land characteristics by logical reasoning through a decision tree constructed by an expert.

4.1.1 The Automated Land Evaluation System (ALES)

The Automated Land Evaluation System (Rossiter, 1990; 1998) is not itself an expert system, but rather a system to enable users to build expert systems for the complex task of land evaluation. Its author suggests that three classes of people will interact with the system. Model builders will use it to construct land evaluation models, model users will enter land attribute data and request evaluations, and finally end users such as land use planners will take and use the printed results.

At the heart of a model built using ALES is a knowledge base schema. Each of these contains a set of proposed land use types, a set of outputs and a set of land characteristics. Within each land use type are land use requirements, each of which has user supplied severities in terms of the corresponding land qualities. Inferential knowledge is represented by decision trees.

This inferential knowledge includes the following:-

a) Determination of land quality severity levels from land characteristics
b) Determination of physical suitability and crop yields from land quality severity levels
c) Inference of land characteristics from other land characteristics

An exhaustive discussion of land characteristics, qualities, and their relationships is provided by Rossiter (1996).

ALES takes input from the user in the form of a series of land characteristics for a given site or map unit to populate its database. The user then requests an evaluation of that site or map unit for one or several land use types. The result is provided in the form of a report. Explanations are available of the path taken through the

decision tree to arrive at the final evaluation. Recent developments include the linking of ALES to the IDRISI Geographic Information System (Eastman, 1997).

4.2 Expert systems using probabilistic reasoning

An advantage of probability as a knowledge representation system is that it can readily represent uncertainty. Conflicting and competing pieces of evidence are effectively weighted so that the most probable are propagated towards the final decision. Much of the early work in this field concentrated on the production of systems to act as expert consultants - that is to embody the collected knowledge of experts in a form which would make that knowledge readily available to those with less skill and experience. Medical diagnosis is one of the knowledge domains in which this approach has been explored.

4.2.1 MYCIN

Much of the early research on expert systems was carried out at Stanford University in California. One of the best known products of that work is the medical diagnosis system MYCIN (Shortliffe, 1974; Buchanan and Shortliffe, 1987). It has been credited (Bender, 1996, p. 22) with being able to diagnose illness in its area of competence as well as, or better, than most physicians. MYCIN is a rule based expert system which allows for uncertainty. This means that the rules are not hard and fast but are defined as being only true part of the time. The degrees to which rules pertain are called certainty factors.

As a diagnostic system, MYCIN starts from evidence in the form of symptoms, test results, etc. and proceeds towards the identification of causes. It is described by its authors (Shortliffe and Buchanan, 1987) as using an approximation of Bayes' theorem. The approximation is a device to handle the subjective nature of many of the probabilities used and the inexactness with which some of the interactions can be specified. However, Adams (1987) shows that the model used is largely equivalent to probability theory under assumptions of conditional independence.

4.2.2 EMYCIN

The MYCIN team generalised their system to produce a tool for use by 'knowledge engineers' in producing expert systems. This tool is known as Essential MYCIN or

EMYCIN (van Melle et al., 1987). It has been applied to a number of problems such as structural analysis and further medical diagnostic tasks (Bennet and Engelmore, 1987). GEOMYCIN (Davis and Nanninga, 1985) is an example of an expert system constructed using EMYCIN. It was designed for environmental management and, together with its linkages to a GIS, is discussed in Chapter 5.

4.3 The PROSPECTOR mineral exploration consultant

Like MYCIN, PROSPECTOR was a product of Stanford Research International (SRI) which developed the software with the intention providing an expert consultant on mineral exploration (Hart et al., 1978). The choice of probabilistic reasoning as an "inference engine" is attributed to the fact that mineral prospecting is as much an art as it is a science, and its current state does not permit the construction of rigorous models. The decision rules in PROSPECTOR have been coded into a GIS framework by Katz (1991); some of the following description of the system is drawn from his work.

4.3.1 Interacting with PROSPECTOR

Transcripts of PROSPECTOR sessions reveal its use of a question and answer expert dialogue in which the user answers questions about the existence of various kinds of evidence. The user also needs to choose a mineralisation model and supply certainty factors for each piece of evidence. Those certainty factors may range between a value of 5 (indicating definite positive evidence) through 0 (indicating a lack of opinion) to a value of -5 (indicating definite negative evidence).

PROSPECTOR incorporated a simple parser to decode the language of 'volunteered' information and assign an initial certainty to such information. For example, the volunteered information 'there *are* carbonates' is credited an initial certainty of 4, whereas the certainty of information that 'there *might* be Sphalerite' is initially scored as only 2. The user has the option to update these estimates during the consultation. The inference rules used and, therefore, the pattern of the dialogue, are explicit to the mineralisation model.

The inference rules in PROSPECTOR have the following general form :-

IF

E_1 and E_2 and .... and E_N

THEN (to degree **LS** and **LN**)

H

Where $E_1 ... E_N$ are pieces of evidence and H is a hypothesis. In plain language, such a rule means that the n pieces of evidence E suggest to some degree the hypothesis H (Hart et al., 1978).

The PROSPECTOR user is required to supply, for each piece of evidence, two additional parameters known as the sufficiency ratio (LS) and necessity ratio (LN). These quantify the degree to which it is encouraging to find that piece of evidence present and the degree to which it is discouraging to find it absent. The quantity LS is analogous to that formally defined in statistics as the likelihood ratio (Hart et al., 1978).

4.3.2 Updating in PROSPECTOR

For ease of computation, PROSPECTOR uses an odds formulation of Bayes' rule. If the probability of some event **A** occurring is **P(A)** then the odds **O(A)** of it occurring are given by:-

$$\mathbf{O(A)} = \frac{\mathbf{P(A)}}{\mathbf{(1 - P(A))}} \tag{4.1}$$

Other definitions from Hart et al. (1978) are of the sufficiency ratio LS and necessity ratio LN as:-

$$\mathbf{LS} = \frac{\mathbf{P(E|H)}}{\mathbf{P(\overline{E}|H)}} \tag{4.2}$$

$$\mathbf{LS} = \frac{\mathbf{P(E|\overline{H})}}{\mathbf{P(\overline{E}|\overline{H})}} \tag{4.3}$$

PROSPECTOR uses a chaining system which builds up through intermediate 'evidence spaces' until the combined evidence supports a hypothesis. The graphical

representation of such a structure is referred to as semantic net. Figure 4.1, reproduced from the original PROSPECTOR documentation, illustrates part of one such semantic net. PROSPECTOR uses Bayesian updating as one its means of propagation up through this net, the other being Boolean logic. The prior probabilities required for Bayesian updating have been supplied by the expert who defined the rules.

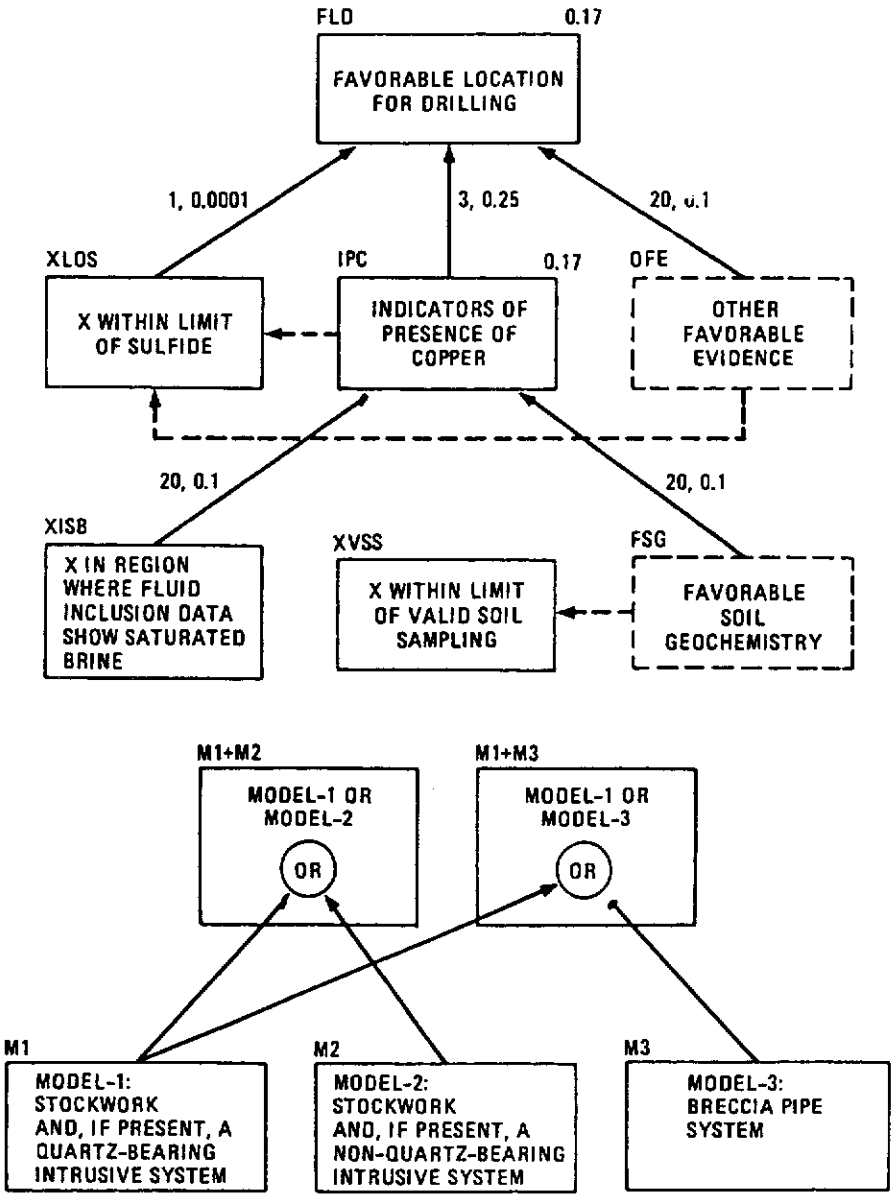

Figure 4.1 Part of a Prospector inference network (from Duda et al., 1978)

In the case of an evidence space at the bottom of a chain, the user-supplied certainty (on the scale of 5 to -5) is compared to that prior probability. The calculus of a Bayesian update also requires the conditional probability of that evidence space being true, given the user supplied certainty. This conditional probability is indicated by P(E|E') and is calculated from the certainty C and the Prior probability P(E) in the following manner (Katz,1991):-

$$\mathbf{P(E|E')} = \mathbf{P(E)} + \frac{\mathbf{C}}{\mathbf{5}}[\mathbf{1} - \mathbf{P(E)}] \qquad \textbf{for } \mathbf{C > 0} \tag{4.4}$$

$$\mathbf{P(E|E')} = \mathbf{P(E)} + \frac{\mathbf{C}}{\mathbf{5}}\mathbf{P(E)} \qquad \textbf{for } \mathbf{C \leq 0} \tag{4.5}$$

For any one update within the network, we also have values for **LS** and **LN** supplied by the user. The expert who built the network supplied the prior probability for the next evidence space in the chain. The Bayesian update is intended to provide an updated value for the probability **P(H)** of both the next evidence space being true, and the associated probability $\mathbf{P(\overline{H})}$ of it not being true.

Prospector proceeds by using the odds formulation of Bayes' Rule which Hart (1978) defines as:-

$$\mathbf{O(H|E)} = \mathbf{LS} \times \mathbf{O(H)} \qquad \text{where evidence is present} \tag{4.6}$$

and

$$\mathbf{O(H|\overline{E})} = \mathbf{LN} \times \mathbf{O(H)} \qquad \text{where evidence is absent} \tag{4.7}$$

The choice of Equation 4.6 or 4.7 is dependent on the value of the user-supplied certainty of the evidence. This odds formulation is then converted back to a probability which is regarded by Prospector as being **P(H|E)**. What is required by the user is **P(H|E')**. This is calculated from this value **P(H|E)** and the values of **P(E|E')** derived using Equation 4.4 or 4.5. The expression used, which is derived from Bayes' rule, depends again on whether the evidence is seen as being positive or negative. In the case of positive evidence, the expression is:-

$$P(H|E') = P(H) + \frac{P(H|E) - P(H)}{P(\overline{E})}[P(E|E') - P(E)] \quad (4.8)$$

The expression for negative evidence is:-

$$P(H|E') = P(H|\overline{E}') + \frac{P(H) - P(H|\overline{E})}{P(\overline{E})}P(E|E') \quad (4.9)$$

The propagation of probabilities then proceeds to the next level in the chain.

4.3.3 Combining evidence in PROSPECTOR

When two or more evidence spaces converge into one, Prospector uses, depending on the rules laid down by the expert, either a Boolean construct or likelihood ratios. In the case of a Boolean construct, the rules in Table 4.1 are used to determine which of several competing probabilities is propagated (Katz, 1991).

Boolean operator	**Rule**
AND	Use minimum probability
OR	Use maximum probability
NOT	Negate evidence 1 - probability

Table 4.1 Rules for updating using Boolean operators (after Katz, 1991).

In the case of a Bayesian update using likelihoods, the following procedure is used (Katz, 1991). For each contributing evidence space **i,** a value of $P(H|E_i')$ is calculated using either Equations 4.4 and 4.8 in the case of positive evidence, or Equations 4.5 and 4.9 for negative evidence. These are then converted to odds using Equation 4.1. The likelihood ratio LE_i is then calculated as the ratio between the odds value $O(H|E_i')$ and the original odds on the hypothesis $O(H)$. That is :-

$$LE_i = \frac{O(H \mid E')}{O(H)} \quad (4.10)$$

These various values of LE_i are then combined multiplicatively and used to update the odds on the hypothesis using:-

$$O(H|E_1', E_2', .., E_n') = O(H) \times LE_1 \times LE_2 \times ....\times LE_n \quad (4.11)$$

This odds value is then converted back to a probability using the inverse of Equation 4.1.

4.3.4 Some inconsistencies in PROSPECTOR

In its original form, PROSPECTOR was coded with three principal mineralisation models. Katz' (1991) emulation in a GIS used just one of the sets of rules in the original version. Although PROSPECTOR was credited with the discovery of a major molybdenum deposit in Washington State in the United States of America, it does contain inconsistencies that render its use questionable.

Both Rhodes and Garside (1991) and Bonham-Carter (1994) examine the use of conditional probability in such situations and point out that any one evidence space in a PROSPECTOR network is overspecified. In particular, there is an inherent inconsistency between user supplied values for LS and LN. Both can be calculated from conditional probabilities and thence in turn from joint probabilities. It can be shown that whilst PROSPECTOR constrains the values of LS and LN, they are in fact related. PROSPECTOR ignores this relationship. A particular example of this is that, in a correctly specified system, if LS = 1 then LN = 1. PROSPECTOR insists that the two values never be equal.

4.4 Causal probabilistic (Bayesian) networks

Causal networks provide a graphical means of representing causal relationships within a knowledge domain (Jensen, 1996). When Bayesian probability is used as a means of propagating evidence through such a network, it is referred to as either a Bayesian network or a Causal Probability Network (CPN). The use of Bayesian networks in expert systems was pioneered by Pearl (1986).

4.4.1 Causation in Bayesian networks

The concept of causation is important. In a general causal network, the hypothesis causes the evidence. We can use this as basis for reasoning from hypothesis to evidence. Such reasoning can be either certain or uncertain. However, with problem solving, particularly in natural resource mapping, we are trying to reason from presented evidence to hypothesis.

In order to do this we must invert the causation. Bayes' rule (Equation 3.3) provides a tool to do just that. Texts on Bayesian networks such as Jensen (1996) illustrate this with examples that range from the trivial to the complex. To return to the marine fossil problem of Davis (1986), discussed in Section 3.3.6, the marine sediments in the two stream basins $\mathbf{B_1}$ and $\mathbf{B_2}$ are causal in the production of marine fossils **A** found in the stream bed below the confluence. That causality can be expressed as a conditional probability of finding a marine fossil given a particular percentage of marine sediments. These conditional probabilities are expressed in the general form $\mathbf{P(A|B)}$. The question Davis (1986) wishes to answer is that of which basin a fossil came from. That is expressed by the conditional probability $\mathbf{P(B|A)}$, to determine which requires the use of inversion.

When using such a system for natural resource mapping, it is important to understand the relative directions of causation and reasoning. For example, if (when mapping soils) we have evidence in the form of a gravel pit, we can reason that we are in a gravely soil. The causation in this case is that gravel pits are caused by the presence of gravely soil. The evidence is caused by the hypothesis and the reasoning is the inverse of the causation, A converse case is presented by the reasoning that a low position in the landscape, as expressed by a high wetness index, has led to moist conditions conducive to the accumulation of organic matter. Here the hypothesis that we have organic matter is caused by the evidence of high wetness index and the causation and reasoning are aligned.

4.4.2 Bayesian networks defined

Jensen (1996, p18) describes a Bayesian network as containing the following characteristics:-

a) A set of variables and a set of directed edges between them;
b) A finite set of states exists for each variable;
c) The variables and the edges form a directed acyclical graph. (Acyclical means that there is no feedback);
d) For each variable **A** with 'parents' $\mathbf{B_1}$ .....$\mathbf{B_n}$ there is a conditional probability table $\mathbf{P(A|B_1,B_2.....B_n)}$.

4.4.3 Variables in Bayesian networks

In the context of Bayesian or Causal Probabilistic Networks, there are three types of variable. These are hypothesis variables, evidence variables and mediating variables. Each has a finite number of states. If evidence is available for any one of these variables, it is said to be instantiated (Jensen, 1996). Examples of the three types of variable can readily be found in the context of land evaluation or natural resource mapping.

Hypothesis variables may be exemplified by soil properties which are typically difficult or expensive to map such as, clay content, acidity or stoniness; each of which can have a range of states. Typical states for clay content might be 0-5 percent, 5-10 percent, 10-20 percent and 20-100 percent.

Evidence variables may, in a land evaluation context, be properties which are relatively cheap or easy to map. Examples would be topographic attributes such as slope and aspect or remotely sensed spectral signatures. Again, examples of states can be quoted. Aspect could have the four states North, South, East and West.

Mediating variables are introduced either to reflect the independence properties of various pieces of evidence or to facilitate the allocation of conditional probabilities. Jensen (1996) cautions against their use simply to produce a more refined model. An example of a mediating variable in resource mapping might be 'favourable topographic position'. Whilst such a variable may be synthesised probabilistically from, say, slope and aspect, it may be more relevant to the causal logic of the model to use a non-probabilistic construct such as a topographic index.

4.4.4 Learning in Bayesian networks

A recent development in Bayesian networks research is that of model seeking and data mining. This falls into the general category of *learning*, which Jensen (1996, p. 53) defines as having two components in this context. One is the specification of the model from available data; the other is the specification of conditional probabilities within a model. The field of medical diagnosis provides examples of model learning such as CHILD (Spiegelhalter et al., 1993). Heckerman (1997) discusses data mining in general and the learning of probabilities from a Bayesian network.

4.4.5 The CHILD program

CHILD is a an example of what its authors (Spiegelhalter et al,. 1993) describe as an 'idiot Bayes" model using directed acyclical graphs and Bayesian conditional probability. The construction of the model is described as comprising three stages: qualitative, probabilistic and quantitative. The qualitative stage considers general relationships within the model, the probabilistic stage defines joint distributions and the quantitative stage involves the specification of conditional probability distributions.

The CHILD model was intended to demonstrate the possibilities of expert systems for medical diagnosis. The knowledge base component is for the identification of congenital heart disease in babies. The propagation strategy used is Bayesian and the model basically has two forms. One is as a subjective Bayesian network specified by experts, the other is as a batch learned network specified by analysis of a large database. It is reported (Jensen 1996, p. 60) that in tests both models performed at a similar level to diagnosis by a physician, with the subjective network performing slightly better than the batch learned network.

4.5 Summary

A number of expert systems have been constructed which use either logic or probability as a means of representing knowledge. Many of these have been designed for use in the field of medical diagnosis. There are interesting similarities between medial diagnosis and the process of natural resource mapping, chiefly the

need to elucidate information about hidden conditions from the examination of signs, symptoms or surface expression.

The Automated Land Evaluation System (ALES) provides an example of a logic based system, whilst probabilistic systems such as MYCIN and its derivatives have their origins in medical diagnosis.

PROSPECTOR, a system designed as an expert mineral prospecting consultant, is a example of a further development of probabilistic systems into the geo-sciences. A detailed account has been given of the calculus behind PROSPECTOR as well as comments on some of its shortcomings.

Bayesian networks, or Causal probabilistic networks (CPN), constitute a more general class of probabilistic systems. The general principles of such networks have been discussed and some comment provided on the direction of causation with reference to resource mapping. Examples have also been drawn from the field of natural resource assessment to illustrate the concept of variables in Bayesian networks. The concept of CPNs has been further illustrated by a short discussion of CHILD, a CPN designed for medical diagnosis purposes.

If use is to be made of expert system techniques in a quantified natural resource assessment method, tools are required to handle the spatial component of resource assessment. Spatial tools are provided by GIS. The next chapter discusses the fundamentals of GIS and cartographic modelling and looks at ways in which they can be linked to knowledge processing systems.

Chapter 5

CARTOGRAPHIC MODELLING, GEOGRAPHIC INFORMATION SYSTEMS AND KNOWLEDGE

Cartographic modelling and data integration are concepts which are inherent, although not always explicit, in the processes used by human interpreters when addressing natural resource problems. Geographic Information Systems (GIS) provide powerful tools to assist in that task. In order to capitalise fully on those tools, a consideration of the role played by expert knowledge in those interpretations is required. Also required is some means of representing that knowledge.

This chapter looks at the development of GIS and its various forms and data processing paradigms. The use of GIS as a cartographic modelling tool in natural resource assessment is then considered, along with the concept of data integration. This is followed by a discussion of knowledge in the context of GIS and natural resource mapping which suggests that probability provides a convenient vehicle for the representation of such knowledge.

After a brief consideration of the probabilistic tools offered by one of the proprietary GIS programs, the chapter concludes with some examples, drawn from the literature, of interfaces between knowledge based or 'expert' systems and GIS.

5.1 Data representation in GIS

5.1.1 Data storage models

There are two principal geographic data storage models, raster and vector. These are described in many standard GIS works, for example, see Bonham-Carter (1994). Briefly, the raster data model stores each data layer, or theme, as an ordered collection of cells. The co-ordinates of each cell are inferred from a knowledge of the origin and dimensions of the raster and of the cell dimension. A vector model considers the world to be made up of points, which may in turn form the vertices of lines (arcs). These arcs may in their turn define polygons. Each of the point, line and polygon entities may exist only as spatial data or may have multiple attributes attached. A raster cell must have a value assigned to it, and may have additional

attributes attached by means of look-up tables. The selection of a particular data storage type is dependant on the nature of the problem and of the data sets to be used (Goodchild, 1991). The raster data model is well suited to the representation of surfaces. This is not confined to topographic surfaces, but also applies to surfaces of spatially varying attributes and of probability. That ability, together with the ease of integration of remotely sensed data products, render it highly appealing for natural resource mapping.

5.1.2 Map Algebra and cartographic modelling

Central to the concept of GIS is the notion of being able to perform geographical calculations or cartographic modelling. The 'Map Algebra' devised by Tomlin in the 1980's (Tomlin, 1990) provides a framework for cartographic modelling which has been embedded into the raster data handling modules of systems such as ARC/INFO GRID (ESRI, 1997). Implicit in this algebra are local, zonal, incremental and focal operators which act as functions on one or more data layers to create at least one new data layer. Some of the more complex operators produce multiple layers as output.

5.2 The nature of GIS

The history of Geographic Information Systems as we know them today is inextricably bound up with the history of computing systems. A typical definition of a Geographic Information System is provided by Burrough (1986, p. 6), who defines them as "a powerful set of tools for collecting, storing, retrieving at will, transforming and displaying spatial data from the real world." Today, such a system is certain to be computer based.

The basic concepts of storing and analysing spatial data, particularly when considering multiple data layers, was in use in such fields as geological exploration and land use planning well before computers were commonplace. McHarg (1969) provides a lavishly illustrated example of a land use planning study for the Potomac river basin in the United States. This study considered multiple themes and was carried out using mylar overlay. Matrix tables defined the inter-compatabilties of various land uses, their natural determinants and their consequences. Such work would now routinely be carried out using a GIS.

Geographic information systems can be viewed in a number of ways. They can be considered as a computer system, or as a series of analytical functions and processing paradigms. A third view concentrates on the data stored and analysed by the system. Others have drawn comparisons between GIS and decision support systems

5.2.1 The computer systems view of GIS

Antenucci et al. (1991) comment that GIS relies on the integration of three aspects of computer technology, namely database management systems (DBMS), spatial analysis tools and graphic display capabilities. Similarly, Maguire (1991) describes GIS as a fusion of database management, computer-aided design, remote sensing and computer cartography. His inclusion of remote sensing acknowledges the impetus given to the concept and development of GIS by the ready availability of digital raster scan satellite data which started with the US Landsat program in the 1970s.

5.2.2 Functional concepts in GIS

Bonham-Carter (1994), in discussing the application of GIS to problems in the geo-sciences, suggests that GIS achieves its goal by performing one, or more, of a number of basic activities on spatial data. He defines these activities as: organisation, visualisation, query, combination, analysis and prediction.

In further analysing the nature of GIS, Maguire (1991) describes three views of GIS. These are the map view (focusing on cartographic aspects), the database view and a view which focuses on spatial analysis.

5.2.3 GIS processing paradigms

Maguire (1991) goes on to define three 'designs' for GIS, namely the file processing, hybrid and extended designs. The latter type stores both geographical and attribute data in a DBMS which also provides spatial analysis functions. This type of system is poorly represented in today's GIS marketplace.

Intergraph Corporation, a major company in the fields of Computer Aided Drafting and Design, introduced their Topologically Integrated Geographic and Resource Information System (TIGRIS) in the late 1980s as a venture into the extended type of system (Herring, 1987). Unfortunately, it was not adopted by users and was

discontinued by the vendor in favour of development of the Microstation GIS Environment (MGE).

MGE belongs to the hybrid category of systems which also include the ESRI ARC/INFO software (ESRI, 1997). Hybrid systems typically store geographical data within the GIS, whilst holding attribute data in an external DBMS. Spatial data processing is handled by the GIS, with the DBMS being responsible for attribute query operations.

The file processing paradigm finds an embodiment in systems such as IDRISI (Eastman, 1997) and the raster processing modules of hybrid systems such as ARC/INFO and MGE. This is the simplest of the processing paradigms and holds both geographic and attribute data in simple raster files.

5.2.4 GIS as a decision support system

In a review which examines the differences between Database management systems, Computer aided drafting systems and GIS, (Cowen, 1988) discusses the concept of GIS as a decision support system. He suggests that a GIS is "best defined as a decision support system, involving the integration of spatially referenced data in a problem solving environment". This view is amplified by Jankowski (1995), who describes the process of integrating GIS and multi-critera decision making methods.

5.3 GIS, the environment and natural resource assessment

Cartographic modelling is ideally suited as a tool for use in environmental and land-use planning applications. Indeed, McHarg's work in the late 1960s (McHarg, 1969), although carried out manually, must be considered to be an exercise in cartographic modelling. Fedra (1993) claims that whilst both GIS and environmental modelling are well established as methods and as fields of research, their integration is in its infancy. This contrasts with the fact that one of the early texts used in the teaching of Computer Cartography and GIS was that by Burrough (1986) which approached the topic from a natural resource assessment point of view.

Natural resource assessment has always relied on cartography to assemble and present its data. Those functions are today taken over by GIS. When the analytical

power of GIS is brought to bear on the problem, the task of the resource assessor is rendered more productive and access is granted to data streams which in the past have been poorly represented.

In a description of the analytical capabilities of GIS, Berry (1993) identifies four classes of primitive operations in map analysis. These are reclassification, overlay of two or more maps, measurement of distance and connectivity and characterisation of cartographic neighbourhoods. All of these operations lie at the heart of natural resource assessment work, especially in an environmental management context.

5.4 Data integration in GIS

The ability to integrate data may be regarded as one of the chief benefits of the adoption of GIS. A simple view of data integration is as a process of making different data sets compatible with each other (Rhind et al., 1984). Flowerdew (1991) considers four basic questions which need to be answered when data are being integrated. These are: what *type* of data, to *where* do the data refer, to *when* do the data refer and how *accurate* are the data. The effect of data accuracy in the integration process is a theme of discussion in a number of papers collected by Goodchild and Gopal (1989). Geographical data sets, which are an abstraction of the real world, will contain both thematic and positional error. These errors need to be handled effectively if the analytical power of GIS is to be harnessed in a meaningful fashion.

Alternative views of the process exist. Some authors (eg. Moon, 1990; Bonham-Carter, 1991) represent cartographic modelling itself as being the data integration process. Bonham-Carter (1994) uses the term "data integration modelling" to describe a process which predicts the occurrence of an attribute based on the spatial coincidence of a number of pieces of evidence.

Common to these descriptions of the data integration process is the concept of synthesising new information from existing data. The new information is intended to be a 'value added' product which is more useful than its constituent data. This is an analogue of the data combination process carried out by a human interpreter when

considering multiple data themes. In the case of the human interpreter, the process draws on expert knowledge.

5.5 Knowledge in the context of GIS

We can characterise the knowledge that needs to be represented in a GIS as belonging to three categories. Firstly, there is knowledge about the spatial relationships between entities stored in the system. This is often referred to as topology. Secondly, there is knowledge about the composition of those entities - this is thematic knowledge. Thirdly, there is knowledge about the relationships between the entities that comprise our data and those entities that we wish to derive as the result of data combination and analysis procedures. This thesis concentrates principally on the third of these categories.

5.5.1 Spatial relationships

Topology is the branch of mathematics that deals with set theory. By analogy and extension, it has been taken into the argot of GIS to describe the spatial relationship between entities. Its particular characteristic, and that which ties it to the truly mathematical branch of the science, is the fact that relationships between spatial entities are preserved irrespective of any geometric transformations to which those entities may be subjected.

Classical examples of topology include the 'knowledge' embodied in a line as to the identity of the polygons on either side of it or of the identity of the two nodes at its terminations.

5.5.2 Thematic knowledge

This is perhaps the simplest and easiest of forms of knowledge to represent in a GIS. It refers to the identity and characteristics of an entity. Examples of this are attributes related to a linear feature such as a road. Its width, route class, surface type, etc. may all be stored. Similarly, for area features, information about land-cover type or the value of real estate lots may be stored as attributes of polygons or grid cells. However, it must be remembered that not all such knowledge is absolute and that a degree of uncertainty may exist.

An example of uncertain knowledge is offered by a grid cell whose land cover type has been determined as the result of classification of multi-spectral remotely sensed data. That grid cell will have some degree of class membership attached to it by the classifying algorithm. Some image processing systems store class membership data in separate data layers, one for each class. Terms such as 'typicality' are used to describe the degree of class membership. The handling of uncertain thematic data is one of the topics explored in this thesis.

5.5.3 Knowledge of relationship

Spatial data is an abstraction and representation of the real world. It is also, of necessity, a simplification. We may choose to map land-cover type, perhaps because it is relatively easily done, rather than some more abstract attribute such as bird habitat. We hope to be able to synthesise not only the bird habitat, but possibly other attributes from our base land-cover data. The degree of success with which this can be carried out depends two factors. The first is how well the necessary relationships can be described and represented. The second is the degree to which the descriptions of the base data actually relate to the habitat. Such relationships will rarely be direct one to one links and will frequently require reference to other 'base' attributes. In this example, those attributes may be topography or prevailing wind direction.

Continuing the habitat example, we may find that bird population fluctuations, perhaps brought on by year to year climatic variation, will cause uncertainty as to the extent of the preferred habitat. What is considered marginal under low population pressure may seem prime habitat under higher population pressure. In other words, an apparently well defined spatial relationship may vary through time. Similar examples can be found in the field of human endeavour. Even in the field of geology, where the situation may be regarded as being more static, there will be grey areas. It is this uncertainty which allows, and indeed invites, the use of probabilistic methods to represent knowledge about relationships. Some such methods have already been discussed in Chapters 3 and 4.

5.6 Probabilistic tools in proprietary GIS

Few proprietary Geographic Information Systems contain implementations of expert systems or probabilistic reasoning tools. However, most have a scripting language

that enables such tools to be built by the user. One exception to this is IDRISI (Eastman, 1997). IDRISI is a low cost raster GIS and contains implementations of 'fuzzy logic' and both Bayesian inference and Dempster-Shafer belief functions. Although the procedure for using these is a little cumbersome, it can be enhanced using IDRISI's scripting language.

5.7 Integrating GIS and expert systems

Most developments in the field of expert systems and GIS integration have concentrated on developing linkages between systems rather than in building new systems afresh. One possible reason for this is the complexity and maturity of such commercial GIS as ARC/INFO.

Burrough (1992) discusses the development of intelligent GIS, but makes no suggestion of anything other than a linkage between the two technologies. More recently, however, reporting as part of a US National Center for Geographic Information and Analysis (NCGIA) research initiative on spatio-temporal reasoning in GIS, Smyth (1998) envisages a modular system. Increases in interoperability of computer applications through mediums such as Open Database Connectivity and embeded controls, like those used by ESRI's Map Objects, have rendered this modular approach more practical. Developments such as Open GIS (Crisp, 1998) will further assist such ventures.

Numerous examples of linkages between GIS and expert systems are to be found in the literature. Some of these have used expert systems to parameterise models in GIS, whilst others have used spatial data layers as representations of knowledge. The difference in approach is essentially in the extent to which knowledge and expectations are spatially represented. In an indirect linkage, all knowledge is represented outside the GIS as a system of rules. In the direct linkage paradigm, knowledge can be represented spatially by thematic maps of uncertainty. There is a general class of systems known as spatial decision support systems (SDSS) to which this latter group more properly belong. Jankowski (1995) discusses SDSS as one of the ways in which multiple criteria decision analysis and GIS may be linked.

Ferrier and Wadge (1997) describe four basic approaches to the task of integrating GIS and expert systems. These are :-

a) Construction of a fully integrated expert GIS;
b) Enhancement of an expert system with GIS tools;
c) Development of an interface between an GIS and an expert system;
d) Enhancement of the GIS with expert reasoning facilities.

The development of an interface is the method commonly chosen by workers in this field. However there is a certain degree of hybridisation between the four basic methods.

Two main linkage methods are commonly found in the literature. These are a link between an expert system which has no spatial component and a GIS, and a more closely integrated system in which the rules of the expert system take on a spatial dimension.

5.8 Linkages with non-spatial expert systems

5.8.1 Geomycin

EMYCIN is a generic expert system shell derived from the medial diagnosis system MYCIN (see Section 4.2.1) by removing all its rules (van Melle et al., 1987). The name is derived from Empty MYCIN. Researchers in Australia used EMYCIN to build a geographic problem solving tool which they termed GEOMYCIN (Davis and Nanninga, 1985). GEOMYCIN is not a direct linkage between a GIS and an expert system but rather uses the expert system to supply parameters for the GIS.

Davis and Nanninga (1985) report on a study, using GEOMYCIN, the objective of which was the production of a system to predict fire damage in the Kakadu National Park. A rules base was constructed which enabled values for parameters such as 'fire danger' to be determined by consideration of several geographical and climatic input values. In order to simplify the spatial application of the rules base a scheme, referred to as geographical equivalence, was devised. This identified areas which, although different, behaved the same way for the operation of particular rules. This seems to have been necessitated by the relatively limited computer power available at the time. The MYCIN engine being used had been designed for use in a question and answer mode rather than drawing input from large geographical data sets.

5.8.2 Logic based systems and GIS

The Automated Land Evaluation System (ALES) is a logic based decision tree method (Rossiter, 1998). Although it does not itself have a spatial component, it has been linked successfully to GIS for a variety of applications.

Van Lanen and Woperis (1992) report the use of a linkage between ALES and an unspecified GIS to assess the suitability of land mapping units for the direct injection of animal manure. GIS data on land use (derived from government statistical data) and soil types in polygon format were synthesised into unique land mapping units. ALES was used to handle the land characteristics and carry out the evaluation using decision trees based on the user supplied knowledge base. The physical suitability ratings thus derived were then passed back to the GIS for production of maps and statistical tables.

5.9 Spatial expert system approaches

Spatial expert systems encapsulate more of the knowledge as layers within the GIS rather than in the 'rules base' of the previous examples. There are a number of approaches to this, which differ chiefly in the degree to which knowledge is held in the GIS or in accompanying tabular databases. A few examples from Australia and overseas serve to illustrate some of these differing approaches.

5.9.1 Weights of evidence

The use of a weights of evidence method of data combination has been pioneered in the geo-sciences by Bonham-Carter (1991, 1994). The method used is essentially a simplified causal probabilistic network designed to use binary input data and is designed for mineral potential mapping. The odds form of Bayes' rule is used to calculate posterior probabilities in a manner similar to that employed by PROSPECTOR (see Section 4.3). The spatial dimension is added by presentation of evidence as binary maps showing the favourability (or otherwise) of particular attributes.

Each evidence layer is assigned a weight that can be directly derived from a coincidence table. This coincidence table has been developed by analysis of known occurrences of the mineral whose potential is being mapped and the attribute whose

favourability for that mineralisation is being considered. Spatial data sets are represented as raster (quad-tree) elements and manipulated using the SPANS GIS.

5.9.2 Mapping forest systems in New South Wales

A combination of expert system and GIS was used to map forest types in New South Wales from classified Thematic Mapper data and digital elevation model parameters (Skidmore, 1989). In this case, the geographical data were held in the SPIRAL GIS and expert system calculations, using Bayes' rule, were performed using custom written software. Maps of a number of topographic variables such as slope, aspect and topographic position were prepared for the 7.5km square study area.

Expert knowledge was derived by a variety of methods including a questionnaire of foresters. This was used to develop conditional probability tables which relate classes in the input maps to eucalypt species. The output from a non-parametric classification of TM data was also used. The expert system based classification is reported as performing better than a simple classification of the remotely sensed data alone.

A related study (Skidmore et al., 1991) mapped forest soils directly from terrain attributes for a 3km square study area. Again, conditional probabilities were obtained from a panel of experts. Results were only evaluated quantitatively using a small number of soil pits. However, the expert system correctly classified soils at 14 out of 21 pits. Visual comparison of the expert system map with a conventional map revealed a number of points of agreement and of disagreement. The authors attribute some of the disagreement to the vagueness of soil landscape descriptions.

5.9.3 Wildlife habitat mapping in Scotland

Aspinall (1992) describes the use of a Bayesian method to combine geographical data sets when modelling the winter distribution of red deer in the Grampian region of Scotland. A Bayesian method was chosen because it emulates the way in which habitat suitability might be assessed by a wildlife expert.

This study noted that the conditional independence assumption of Bayes is often not met when using environmental data sets, and endeavoured to minimise the risk of

violation by using the smallest number of datasets possible in the development of the models. Presence/absence data from a red deer census and GIS data layers representing environmental variables were used as inputs. Probability calculations, using an inductive process, were based on coincidence tables developed from a geographical subset of the available census data. The procedure was implemented using in-house raster GIS software.

Data from another part of the same region were used to test model output. Red deer presence in that area coincided with a predicted probability of 0.8 or greater at the majority of test sites. However only about one third of a random selection of sites with no recorded deer presence had low (<0.2) probability of red deer presence, while more than half had probabilities in excess of 0.8. These sites had habitat that was suitable for deer but fell outside the census area. This points to the value of such methods for representing the state of knowledge about environmental variable such as habitat distribution. The fact that the species is not recorded in an area does not mean that the habitat is not suitable.

5.9.4 Desertification risk in burned Greek forests.

In work carried out concurrently with that described in this thesis, Stassopoulou et al. (1998) investigated the use of Bayesian (causal probabilistic) networks in GIS. They discuss the problems of using continuous and discrete data in the same model before adopting the discrete method in their study. They also provide a useful discussion of means of estimating the uncertainty in input data.

Their network contains a situation which, at first, seems to violate the need for conditional independence of evidence. In the example given, they have two nodes on different lines of reasoning which have a common parent. The authors claim that by instantiating that parent, that is assigning it a fixed value, the apparent loop through the data is effectively blocked and analysis can proceed. The values of the two child nodes cannot influence each other through the parent since the parent value, being part of the evidence data is not subject to change. This study uses a training data set to establish joint and conditional probabilities rather than expert knowledge. Since the study is essentially a proof of concept, no rigorous estimation of accuracy was made.

5.9.5 A spatial emulation of PROSPECTOR

The original PROSPECTOR expert system (Hart et al., 1978) was designed as a non-spatial query session system. Its original authors added some graphical capability through an interface to a simple image processing system capable of handling 128*128 pixel arrays. Katz (1991) emulated the system using the MAPS raster GIS. Katz's implementation required the user to provide individual maps for each of the evidence spaces in the inference net. These individual maps each represent the probability the proposition represented by any evidence space is 'true,' given the available evidence.

In order for this construct to work logically, evidence spaces must represent concepts and propositions such as 'proximity to intrusive contact' or 'having favourable Au concentration'. Maps of these concepts need to be constructed from base geographic data. Methods for such conversion include buffering, expert assignment and the use of favourability functions (essentially non-linear buffering). Individual calculations were carried out using map algebra from MAPS.

The PROSPECTOR model relies heavily on subjective probability as part of its knowledge base and, as described in Section 4.3, can become poorly conditioned under situations in which parameters are over specified. The MAPS implementation was not immune to this, as there was no checking that the values provided by the expert were logically coherent. Using data extracted from the original PROSPECTOR reports the MAPS, emulation was able to reproduce the final output for the Copper Island deposit study reported in Hart et al. (1978). An accuracy of 80 percent was claimed, although the MAPS emulation used only a subset of the original evidence spaces.

5.10 Summary

Geographic information systems, although a relatively recent technology, draw on the long established methods and concepts of cartographic modelling. There are a number of ways in which a geographic information system may be regarded and described. Central to all of them is a system for storing spatial data and a set of tools for the analysis and display of that data. Such systems may also be used for the integration of data sets to provide enhanced information products.

Natural resource mapping and data interpretation have always used such tools as part of the manual process of creating representations of natural phenomena. The concept of expert knowledge is intrinsic to this process. In replicating and enhancing natural resource assessment by the application of GIS, the nature of this knowledge and the means by which it can be represented need to be considered.

Several examples have been provided of linkages between GIS and knowledge representation and manipulation systems. Before applying a similar linkage to the natural resource mapping process, it is worth reanalysing that mapping process in the light of these examples. The next chapter discusses options that may be used to quantify the process of soil mapping.

Chapter 6

QUANTIFYING THE SOIL MAPPING PROCESS

Previous chapters have examined the natural resource mapping process, in particular soil mapping. The conventional process involves the application of knowledge. A number of methods for represented that kind of knowledge have been discussed. Examples from the literature have illustrated ways in which analogous processes may be automated and quantified by the use of knowledge representation techniques and GIS. We now return to soil mapping and consider how it, in particular, may be quantified and formalised.

It has previously been noted that the soil survey method makes effective use of inexact concepts and that the cartographic representation of its current map output could be improved to better represent the knowledge contained in the process. These two observations can act as a guide to the construction of a quantified system that retains both these characteristics.

6.1 General considerations

In general terms, the task of making a process quantitative needs to be considered from three standpoints. These are:-

a) The amenability of the process to quantification;
b) The impact on those individuals and organisations who carry out the process;
c) Software and hardware issues associated with implementation.

Naturally, there is some degree of overlap between these. For example, the degree of change required by soil survey practitioners has an effect on the amenability to change of the whole system. Similarly, the software and hardware resources already in place in an organisation will influence the skills base and amenability to change of those practitioners. This chapter considers first the organisational and theoretical aspects of quantifying the method, and then examines some of the hardware and software issues involved. It concludes by describing the overall concept for a quantitative soil mapping method.

6.2 Current trends in automation of soil mapping

In a major review of Agricultural Land Evaluation in Australia, Shields et al. (1996) highlighted the need to move to a quantitative land evaluation process. A long term goal of "..the formation of objective, integrated, land evaluation systems which essentially consider productivity and resource conservation" was expressed (Shields et al., 1996, p. 2). It was, however, recognised that such a situation will take time to achieve. Among the primary attributes used in the land evaluation process are those pertaining to the basic soil resource. A good first step in the overall quantification of the process can be made by improving the mapping of soil attributes.

In Chapter 2, the process of soil survey was described as being the construction and subsequent cartographic representation of a conceptual model. An alternative method of representation, using surfaces to show the probability of occurrence of individual soil attributes was also discussed. Irrespective of the means of cartographic representation, there are gains to be made from the development of a means of quantitatively representing the knowledge used in the construction of those conceptual models.

These gains come principally from the increased flexibility of output that GIS gives us the opportunity to exploit. Not only do we have the potential to produce maps of fundamental soil attributes that can be combined as required, but we can also readily ascribe certainties of occurrence to those attributes.

Within Australia, the task of mapping soil resources lies, typically, with State Government agencies such as Departments of Agriculture. Other surveys, often at higher spatial resolution, are carried out from time to time by private consultants. These agencies and consultants now routinely use GIS as an aid to the cartographic compilation of surveys.

A land resource assessment team will typically consist of at least one surveyor assisted by technical staff. Those technical staff will be responsible for the preparation and analysis of samples and for the cartographic compilation using GIS. In some organisations, the surveyors may undertake part of the cartographic compilation. In essence, the survey is carried out not by an individual, but by a

multi-disciplinary team. Some degree of automation is currently found in two areas of the process. These are, firstly, the development of databases holding site data descriptions and the results of soil tests and secondly, the previously mentioned use of GIS for cartographic compilation

6.3 Steps to a quantitative soil survey method

6.3.1 A description of the process

The process of soil survey, as described in Chapter 2, is represented schematically in Figure 6.1. Not all of this process is amenable to being quantified or is capable of being enhanced by the use of GIS. Some fundamental elements must remain unchanged; these include the visual reconnaissance and field work components. However, the process of building the conceptual model and map production can be considerably enhanced. This enhancement involves a greater formality in the recording of the knowledge component of the conceptual model.

The provision of a framework to formalise (and quantify) a conceptual model also makes that model more transferable between areas and even between individual surveyors. Whilst it is true that no two soil surveyors will ever have exactly the same interpretation of a landscape, there will be a degree of commonality. By describing the model in explicit numerical terms, it can be tested, verified and refined.

The handling of all data in a GIS from an early stage in the process, means that the previous constraint of delineating quasi-homogenous map units can be removed. A more flexible scheme can be devised which allows the representation of individual resource attributes and their inherent continuous variation. These can then be combined, as required, in answer to any particular query. In order to do this, a basic change is required in the way the conceptual model is constructed.

Instead of being directed towards the homogenous units in the traditional model, the surveyor's conceptual thinking must now be directed towards individual attributes. At first encounter, this may seem to the surveyor to be an extra imposition. In practice, it will soon be seen that the models have always been directed towards attributes. They have merely been constrained by the need to work within the polygons of a traditional natural resource map. An additional advantage is that there

should be greater consistency between surveyors in the mapping of attributes than there is in the mapping of classes. This is because the grouping of those attributes into classes, takes place in multi-attribute space and offers opportunities for different, and often conflicting, boundary definitions in that space.

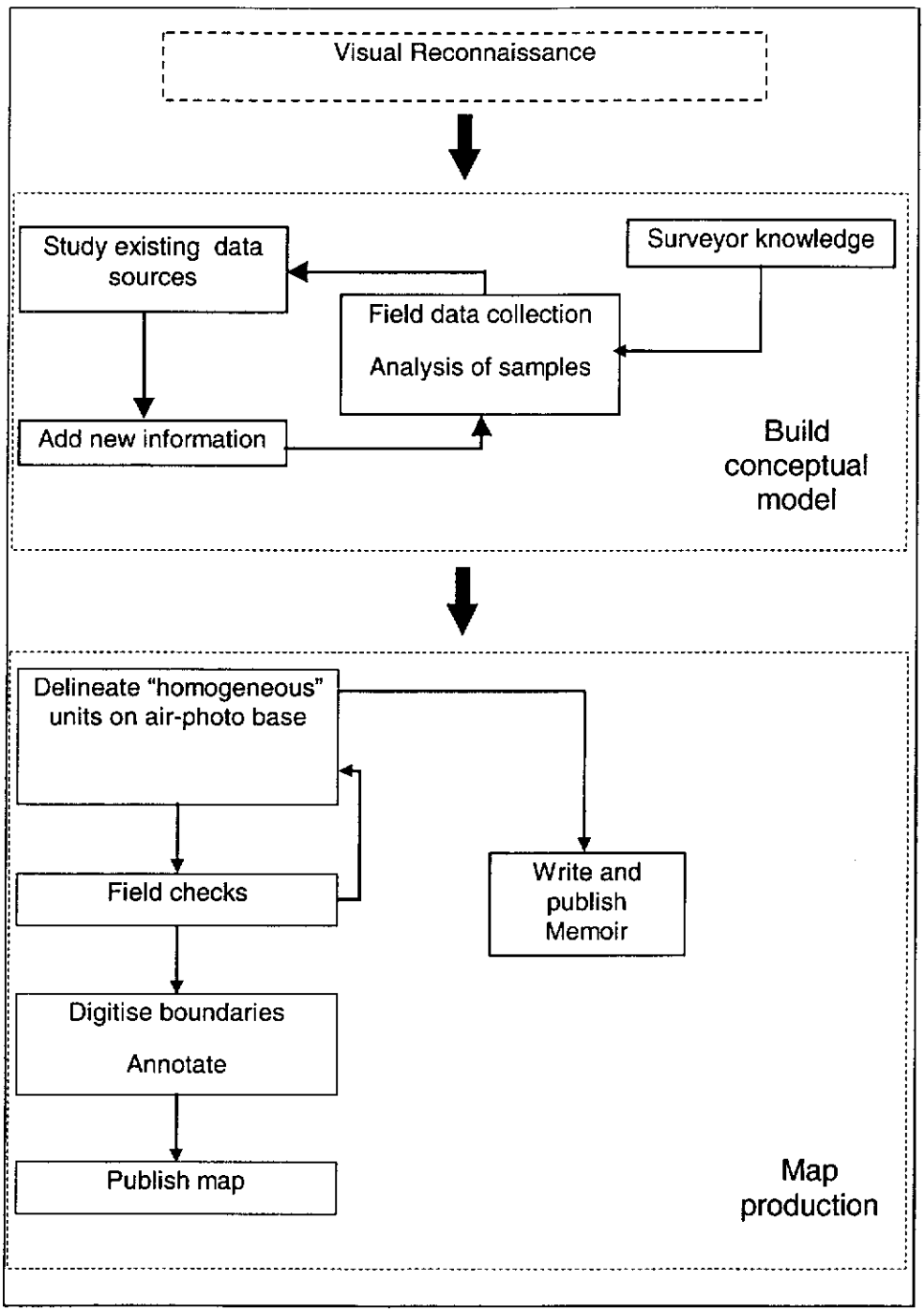

Figure 6.1 A schematic representation of the soil mapping process

6.3.2 Quantifying the process

The transition to the use of quantitative information as input to the model imposes a greater change on the surveyor. A prime example is the ability to use digital terrain data rather than directly interpreting a stereo view of the landscape. The surveyor has always known the relationships between, for example, slope and organic matter accumulation. It is generally assumed that organic matter will accumulate more readily in areas of low slope. Now, the opportunity is offered to quantify the qualitative statements in that relationship expressed by the terms 'more readily' and 'low'.

There are two stages to this quantification. It is first necessary to determine parameters that describe states such as 'low'. Once that is done, a mechanism is required to develop and describe relationships between this information, or evidence, and the attributes being mapped.

The development of such relationships will be assisted by the presence of datasets from the fieldwork stage of the mapping process. If both the attributes to be mapped and the quantitative variables being used to predict them are measured in the field, we can readily develop coincidence tables. These are an expression of the joint probability distribution of the attributes and the predicative information. Some predicator variables, such as derived topographic indices, are difficult to measure in the field. In these cases, a digital map of those predictors can be queried in order to extract the values of that variable which occur at each of the fieldwork sample points.

6.4 Paradigms for a quantitative process

A re-examination of the traditional method shows it to be broken down into two essential parts: model building and map production (Fig 6.1). The first takes place in the mind of the surveyor, aided by sketches, tables etc. which are generally organised in notebooks. The second stage, traditionally, takes place in a cartographic drawing office, now generally aided by GIS. The process can therefore be described as having a knowledge-intensive 'expert system' phase, followed by a more mechanical representational phase.

A major step in quantifying and formalising the process is to use the inherent power of GIS to either replace or augment parts of the model building process. It is therefore worth reconsidering the four methods suggested by Ferrier and Wadge (1997) for the integration of expert systems and GIS. To recapitulate, those are:-

a) Construction of a fully integrated expert GIS;
b) Enhancement of an expert system with GIS tools;
c) Development of an interface between an GIS and an expert system;
d) Enhancement of the GIS with expert reasoning facilities.

The natural resource assessment community already uses GIS for cartographic compilation. Their investment, in both software and skills, may be somewhat specific to existing proprietary systems. The complete development of a fully integrated expert GIS seems unattractive, since it would erode the value of this investment.

Methods 'b)' and 'c)' both require an existing expert system. The analysis of expert system methods carried out in preceding chapters tends to point to the use of a system based on probability. Such systems generally have interfaces designed more for use as a question and answer session rather than allowing the input of spatial representations of knowledge. This requires the special characteristic of being able to work with thematic maps as both input and output, which is not generally found in expert system 'shells'. Such a system would need to be simply interfaced to a GIS and be reasonably intuitive, thereby enabling its use by land resource mapping professionals rather than by computer scientists.

Considering this, the fourth paradigm seems the most attractive. It allows us to take existing expert system *methods* and construct a system that is uniquely designed for natural resource assessment. A number of functions of the overall system will require the use of standard GIS tools such as geo-coding, classification of data and map algebra. Rather than trying to duplicate these functions within the expert tool it is best to rely on existing GIS packages to provide those services.

Summarising the discussion of this and the preceding section, the soil mapping process could be quantified and formalised using a system which:-

a) Uses probability to represent knowledge;
b) Is able to synthesise information from multiple disparate data sources;
c) Is designed to map soil (or other resource) attributes rather than soil types;
d) Interfaces readily to GIS;
e) Provides an intuitive tool for soil surveyors;

The remainder of the work reported in this thesis covers the specification and development of such a system.

6.5 Software and hardware considerations

As a prerequisite, a system to partially automate the soil survey process must provide functional resources that accurately replicate the parts of the process being automated. However, in order to ensure adoption, those automated processes must be as compatible as possible with existing methods and skills bases. That means essentially two things: user friendliness and compatibility with existing software.

If potential users follow the 'user profile' described in Section 6.2, they are likely to include both land resource surveyors with a range of depth of experience and GIS staff. The overall method and its user interface must be one with which they will readily identify with little effort. It must, therefore, seem intuitively similar to processes with which they are familiar. Additionally, in order to capitalise on existing investments in software and training, it should be as compatible as possible with the GIS software currently in use.

The choice of language in which any expert system tool is written and the platform on which it is run will similarly be governed by the processing environments currently in place amongst the intended user community. It is, therefore, worth considering the systems commonly in use in Australia and the implications of that in terms of platform and language selection.

6.5.1 Common systems in use.

The Geographic Information System most widely used in Australia, at least by those agencies whose mandate is to map natural resources, is ARC/INFO. ARC/INFO is an extremely powerful suite of routines originally designed to run on workstations under the Unix operating system. An earlier PC version of the software has been abandoned in favour of a version of ARC/INFO for Windows NT. Also from the same software company is a relatively new product, ArcView.

Although available for both PC and Unix platforms, ArcView's main market share is in the Windows based PC environment. The software is capable, with add on modules, of processing in both raster and vector data and is almost, but not quite, as powerful as its ARC/INFO 'parent'. The two systems have many data structures in common.

An increasing trend in organisations is towards the use of ArcView for day-to-day work by specialists in various disciplines supported by a GIS 'service group' who maintain corporate spatial databases using ARC/INFO. In such organisations, a local network enables seamless sharing of data resources between the Unix and PC systems.

Although less widely used, another major software suite found in natural resource mapping is the Intergraph Microstation GIS Environment (MGE). Like ARC/INFO, this system has its origins on Unix platforms, but its use under Windows NT is now becoming more widespread.

6.5.2 Language

Most geographic information systems have an associated scripting language. These languages vary in complexity, but are designed to enable, at a minimum, the enhancement of the processing capabilities through the use of batch scripts. At the other extreme, some of them provide complex functions that allow direct access to the proprietary data formats of the GIS concerned. They typically have a facility to extend the Graphical User Interface (GUI) of the host GIS.

It would be possible to develop an expert system tool using the scripting capabilities of any of the common systems described in Section 6.5.1. Choosing to do so would, however, severely restrict the range of potential users. It is therefore desirable to identify an alternative, neutral language. The choice is to some extent dependant on the development platform to be used. In order to satisfy the user friendliness requirement, a graphical interface is necessary. The development language chosen must, therefore, be capable of easily constructing such an interface.

6.5.3 Linkages

As indicated above, a range of geographic information systems are in use by the soil mapping community. In order to gain wide acceptance, the expert system tool must be able to interact with a number of those systems. This militates against the design of a tool that directly manipulates system proprietary data files, although some strategy for such manipulation is necessary for two reasons. Data files will need to be queried in order to 'mine' any knowledge inherent in the data. Similarly, standard GIS tools such as map algebra need to be accessible.

One solution to this problem is to create, for each individual GIS, a simple set of routines to customise the necessary functions. These routines can be written in the native scripting language of the GIS and can communicate with the expert system tool using a simple set of interchange files. If handled appropriately, such a strategy would permit cross platform operations.

Given an appropriate network structure, it becomes possible to have an expert system tool operating on one platform and exchanging data and commands with a GIS operating on another platform. Modern network transfer protocols are capable of handling any file format conversion that may be required.

6.5.4 Language and platform choice

A choice needs to be made between developing in a Unix or PC environment. Given the network cross-platform capabilities described above, either is a potential candidate. However, there is currently a move by the larger GIS vendors away from Unix systems towards Windows NT platforms. This in itself suggests a 32 bit PC environment as being the best, giving the ability to run under either Windows 95, 98,

or NT. Other compelling arguments in favour of such an environment are the ubiquity of machines running such systems and the increasing use of GIS packages such as ArcView on laptop computers.

A number of languages are available for development in a 32 bit PC environment which exhibit a range of capabilities. In practical terms, the choice comes down to one of the Microsoft Visual products, either Visual Basic or Visual C++. Each of these is capable of providing an environment in which to develop software with 'forms-driven' graphical interfaces, thereby assisting with meeting the user-friendly criterion. Being a truly 'object-oriented' language, Visual C++ could claim some slight technical advantage over Visual Basic. However, Visual Basic has an object-oriented style of syntax and is, more importantly, easier to use. Unlike earlier Basic language interpreters, it allows the creation of executable files for distribution. This, coupled with its relative ease of use, makes it the preferred choice of programming language.

6.6 Functional stages of a quantitative process

Re-examining the graphical representation of the soil mapping process, shown in Figure 6.1, an alternative diagram is proposed for a quantitative method. Figure 6.2 shows the process divided into three sections; model building, data processing and map production. We can examine each of these in turn to see how they relate to the original method. It should now be assumed that our intention is not to map the soils of an area but to map attributes or properties of those soils.

6.6.1 Model building

The process of landscape familiarisation through fieldwork described in the previous chapter is still required. At the same time as this is taking place, the surveyor must be assessing the availability of digital data which describes landscape parameters and other indicators. These may take several forms, including digital terrain models, reconnaissance mapping, remotely sensed data and geo-referenced sample site data.

The model building process now includes the assembly of these data sets in GIS, as well as the consideration of landscape development process. The conceptual model can now be quantified and its inherent knowledge defined. This model will, to some

extent, feed off the available data in a knowledge editing process. Some data sets may be of variable or dubious quality, and the knowledge editing process needs to be able to handle such circumstances.

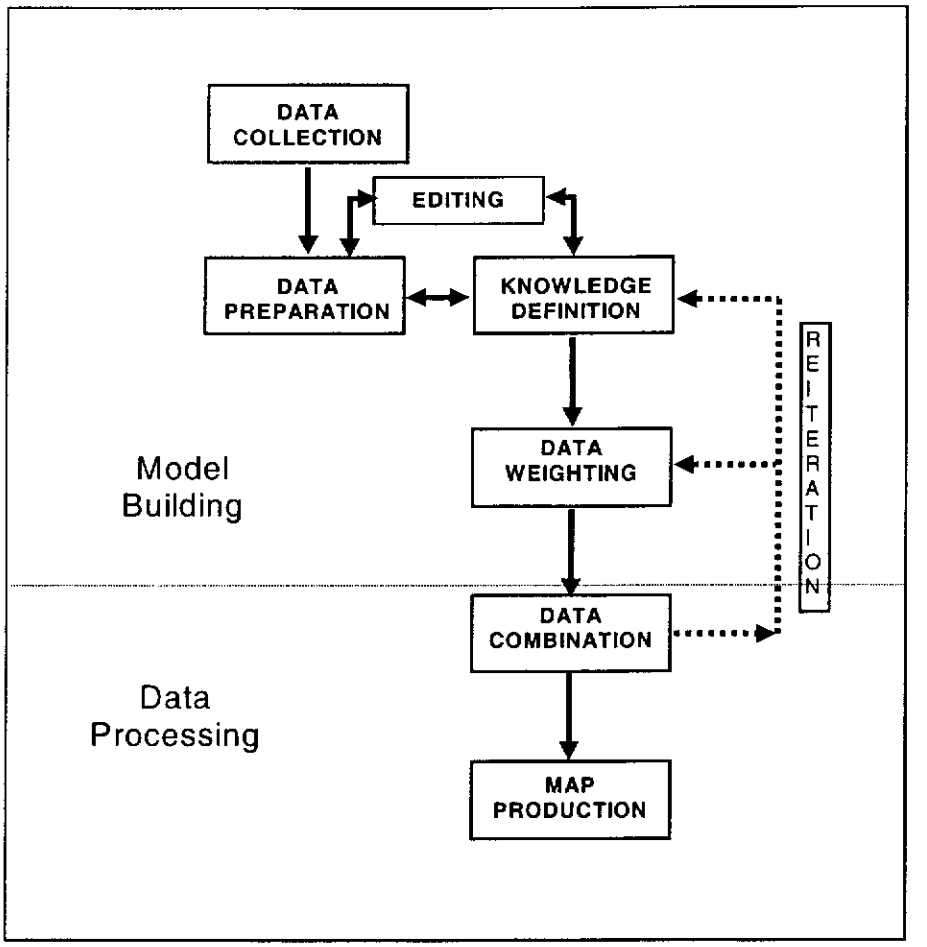

Figure 6.2 An alternative view of the soil mapping process

The various pieces of data used as evidence to predict the attribute of interest will perform that prediction with different degrees of certainty. A representation of knowledge about those different degrees, or weights, must form part of the model building process.

6.6.2 Data combination and map production

Much data processing takes place as part of the model building process. This develops a series of weighted relationships between our various input data and the attribute being mapped. All that remains now is the combination of the individual estimates of the attribute being mapped, based on individual relationships, into one

estimate for each location in space. This is the functional equivalent of surveyor's delineation of boundaries on a base map. As with that delineation, this will generally be an iterative process.

A cartographic representation or map production step can follow the data combination phase. This may produce an output similar to that discussed in Section 2.4 and illustrated in Figure 2.2.

6.7 Implementation

The process described in Section 6.5 involves the use of a custom written expert system tool linked to a GIS. Figure 6.3 shows how the tasks may be partitioned between the two pieces of software. To simplify the diagram, the iterative phase has been removed. The majority of the model building is handled by the expert system, with the GIS being responsible for data preparation and processing. Data collection is shown as split between the two systems in recognition of the fact that some existing data may already be digital, whilst other data will be collected by fieldwork.

There is a requirement for information interchange between the systems at two points. These are between data preparation and knowledge definition, and between the data weighting and data combination phases. The information that will be passed through these links is not high volume spatial data, but summary data about that spatial data. If the expert system tool is probabilistic in nature, that summary data will take the form of tables of probabilities.

6.8 Summary

A procedure has been presented by which the soil mapping process may be quantified and to some extent formalised. It involves the construction of expert system tools to operate synergistically with a GIS. Following a consideration of potential platforms and languages, Microsoft Visual Basic under a 32 bit PC operating system is selected as the development environment which offers the greatest flexibility for interfacing to 'host' GIS. Two-way communication between expert system routines and a variety of GIS can be constructed in the native scripting languages of those systems. The following chapters discuss in detail the design, construction and operation of such expert system tools.

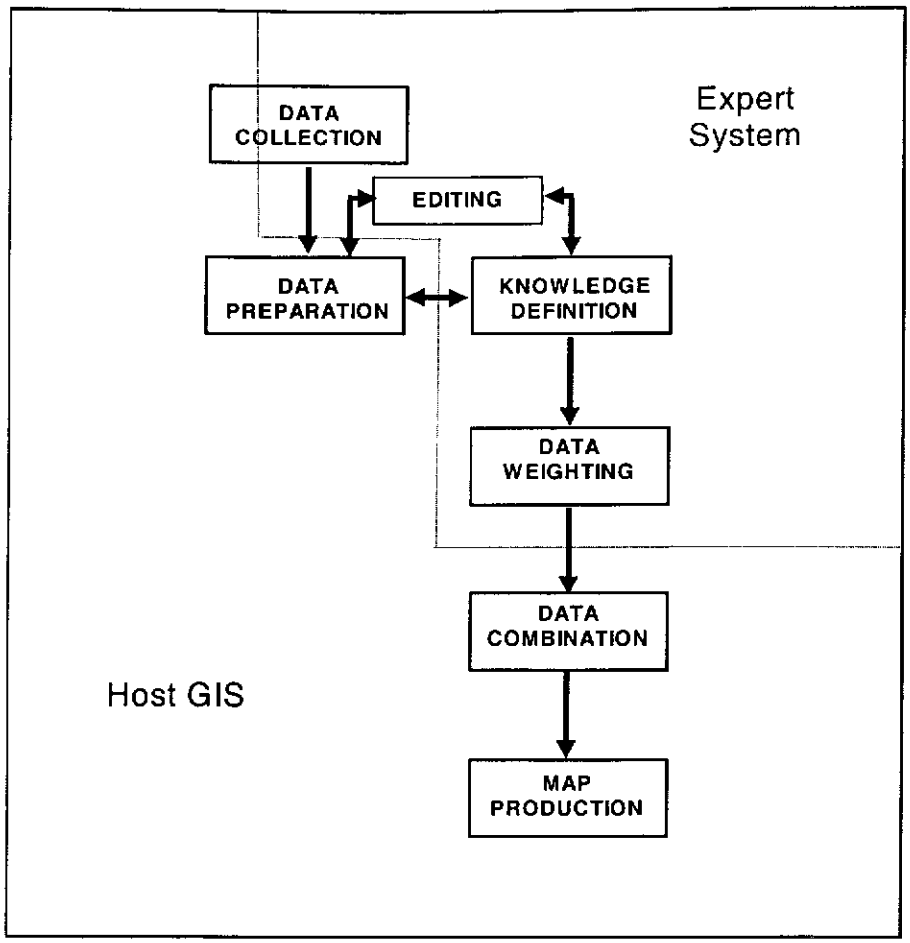

Figure 6.3 Partitioning tasks between a GIS and an expert system

Chapter 7

EXPERT SYSTEM ALGORITHMS FOR A QUANTITATIVE SOIL MAPPING SYSTEM

The preceding chapter described a process of quantification and automation for soil and natural resource mapping. It was suggested that this might be achieved using a combination of an expert system and a GIS. The expert system will handle the model building part of the process with the GIS being responsible for data preparation and combination. In Chapter 4, structures known as Bayesian networks or conditional probability networks (CPN) were described. These networks provide a useful mechanism for the manipulation of multiple evidence that either supports or contradicts some proposition. This is essentially what is involved in the model building and editing process which is described in the previous chapter. This chapter describes algorithms for the application of Bayesian networks in soil mapping. A software implementation of these algorithms will be described in Chapter 9.

7.1 A simple Bayesian network for soil mapping

Referring back to Section 4.4.2, we find a definitive description by Jensen (1996, p. 18) of a Bayesian network as having the following characteristics:-

a) A set of variables and a set of directed edges between them;
b) A finite set of states exists for each variable;
c) For each variable **A** with 'parents' $\mathbf{B_1}$ .....$\mathbf{B_n}$ there is a conditional probability table $\mathbf{P(A|B_1,B_2.....B_n)}$;
d) The variables and the edges form a directed acyclical graph (ie there is no feedback).

We can use this list of descriptive characteristics to examine the way in which a Bayesian network could be constructed to assist a soil surveyor.

7.1.1 A set of variables and edges

A Bayesian network requires as its starting point a hypothesis. We shall take, as an example, the presence of high levels of pisolitic gravel. For now we will ignore such matters as a definition of 'high levels' and concentrate on the concept involved. This

then is our hypothesis: 'That pisolitic gravel exists at high levels.' In spatial terms, we wish to produce a map showing areas where the hypothesis is true.

Using the traditional mapping method, a surveyor would map a homogenous unit, perhaps named Gravelly Soils. This unit would contain all those areas for which gravel content is high. The boundary of that soil unit would be drawn after consideration of a number of factors such as position in the landscape, soil colour as observed in air-photos, and fieldwork. From the standpoint of a Bayesian network the various factors considered are regarded as evidence variables. The logical links between them and the hypothesis are edges in the network.

A set of pieces of evidence and a hypothesis can be taken as providing the required set of variables. We can therefore depict a simple network as follows (Figure 7.1)

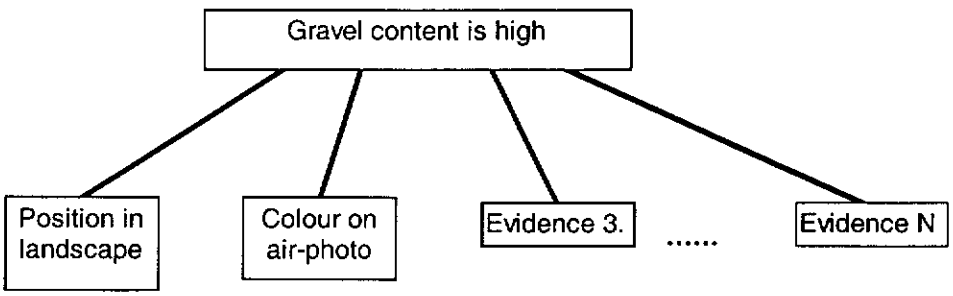

Figure 7.1 A simple Bayesian network

7.1.2 A finite set of states

The hypothesis that high levels of organic matter exist represents only one state of a variable. There will exist a contrary state, the absence of high organic matter content. This set of two states is the minimum that can exist for any variable. Similarly, the evidence variables will possess a minimum of two states and frequently more. The number of states, however, will be finite. An evidence variable that is continuous in nature can be rendered discrete by the definition of a set of bounds in its attribute space. The states of a variable are analogous to the classes in a digital map and both terms may be used in a discussion of the concepts.

Some nicety in semantics is required when designing a network. The variable 'High gravel content' has two states: True or False. The assignment of a particular soil sample to one of these states depends on the prescribed cut-off which defines 'high' as well as the actual gravel content of the sample. It is suggested above that there must be a mutually exclusive state of 'absence of high gravel content'. This is, subject to suitable definition of the terms 'high' and 'low', exactly the same as in a two state system of 'High gravel content' and 'Low gravel content'.

In practical cases this will almost certainly be extended to a multi-state variable 'Gravel content' with each state covering a range of, for example, 20 percent. This in turn can be reduced, for any one of those states, to a two-state variable where that state is either true or false.

Table 7.1 illustrates the point with some arbitrary state cut-offs. The variable named 'High Grav' has two states, labelled TRUE and FALSE, These correspond to the states HIGH (>50 percent) and LOW(<50 percent) for a variable named 'Grav.' The same variable is also shown with five states. The last column shows a variable called 'Grav 20-40 percent' which has only two states, even though the FALSE state is disjoint in attribute space.

<table>
<tr><th>Variable name</th><th>High Grav</th><th>Grav</th><th>Grav</th><th>Grav 20-40%</th></tr>
<tr><td rowspan="5">States</td><td rowspan="2">TRUE</td><td rowspan="2">HIGH</td><td>80-100%</td><td rowspan="3">FALSE</td></tr>
<tr><td>60-80%</td></tr>
<tr><td rowspan="3">FALSE</td><td rowspan="3">LOW</td><td>40-60%</td></tr>
<tr><td>20-40%</td><td>TRUE</td></tr>
<tr><td>0-20%</td><td>FALSE</td></tr>
</table>

Table 7.1 Different representations of a variable

7.1.3 Existence of conditional probability table

A conditional probability table will only exist if there is some causal or logical connection between the hypothesis and each piece of evidence. That restricts admissible evidence to that which either has a causal process impact upon the hypothesis state or is a manifestation of it. Merely statistical relationships based on apparent mutual abundance should not be used.

Such relationships may easily be postulated where none exist, frequently by misguided analysis of data. History is littered with such fallacies; for example, the long held belief that "foul swampy air" caused malaria. The real connection in this case is a third variable, the mosquito that carries the disease and favours a habitat that is "foul and swampy". Analogous situations could be found in soil mapping. Care must be exercised in the construction of networks, particularly when data alone is used to derive them.

7.1.4 A directed acyclical graph

This property of a Bayesian network is almost always going to be satisfied in a network designed for soil mapping, since feedback rarely, if ever, exists in that context. Feedback can be illustrated by an example from ecology. A model for determining population density of some species will consider (as evidence) habitat factors such as availability of food resources, water, and preferred climatic conditions. However, there is a degree of feedback in such a system. There is a two-way link between food availability and population, which can cause populations to peak and then decline as resources are overused.

In soil science there are examples of processes which effectively reach a steady state and plateau. Processes of both physical and chemical weathering may achieve a state of equilibrium, as long as conditions remain the same. The system is not closed and there is no feedback.

7.2 Parameters for a simple network

To examine the parameters required for the operation of a Bayesian network, we can reduce it to a single piece of evidence (**E**) supporting a hypothesis (**H**). We wish to calculate the posterior (conditional) probability that the hypothesis is true in the light of the evidence. We denote this posterior probability as **P(H|E)**. We must also be aware of the converse situation, that the hypothesis is false (**!H**). We can use Bayes' rule (Equation 3.12) to calculate **P(H|E)** as follows:-

$$\mathbf{P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E \mid H) \times P(H) + P(E \mid !H) \times P(!H)}} \qquad (7.1)$$

For a plain language explanation, we can refer back to the gravel content example. **P(E|H)** is the probability that some evidence exists, given that a location of interest has high gravel content, and **P(E|!H)** the probability that some evidence exists, given that the location does not have high gravel content. The ease with which this may be determined depends, to some extent, on the nature of the evidence. As discussed in Section 3.3.4, the definition of conditional probability, restated here in its general form as Equation 7.2, may be used to relate conditional probability to joint probability. This requires the specification, in this case, of a probability for the variable **B**.

$$\mathbf{P(A \mid B) = \frac{P(A,B)}{P(B)}} \tag{7.2}$$

In summary, in order to use a Bayesian updating calculus in a correctly specified network, the following parameters are required:-

a) Prior probability of the hypothesis;
b) Prior probability of each piece of evidence;
c) Joint probability between evidence and hypothesis.

All of these refer to probability distributions, with each distribution having as many members as there are states in the appropriate evidence or hypothesis. We must now consider means by which such parameters may be estimated in the context of resource assessment.

7.3 Methods of parameter estimation

There are essentially two ways in which parameters for a network can be derived. One is by examination of sample data, the other is by expert assignment. Each of these has potential problems.

7.3.1 Examination of sample data

If the sample size is very small, sample data may be unrepresentative. Even if the sample size is large, there may be biases due to the spatial distribution of the samples. In any event, the sample is likely to represent only a small fraction of the total area being mapped. For soil surveys being conducted for publication at a scale

of 1:100:000, a typical sampling density would be one point every square kilometre. Each sample will be taken from an area of about ten metres square (100 m^2.). The entire sample set, therefore, represents only 0.01 percent of the total area. This is compounded by the fact that the survey is unlikely to be carried out on a regular grid basis.

This is particularly relevant in the case of soil attributes whose occurrence is rare. For example, high levels of organic matter are rare in landscapes such as Western Australia, but not non-existent. It is quite possible that no sampling scheme of a study area will produce any sample points that fall within an area having high organic matter. The impact of sampling strategies on the setting of parameters is discussed at greater length in Chapter 12.

7.3.2 Expert assignment

Although expert assignment can also contain biases, its main drawback lies in its use of inexact terminology. This is particularly true in the assignment of conditional probability values. An expert will initially use natural language to estimate the degree of support that a particular piece of evidence provides for a hypothesis. The natural language statements then need to be converted into numbers. It is relatively easy to suggest a scale for quantification of statements ranging, for example, from 'weak support' through to 'strong support.' PROSPECTOR, for example, did this through the use of a certainty scale between -5 and + 5 (Figure 7.2).

The real problem relates to the constraints placed on probability values. Whilst it may be simple to convert a 'strong support' statement to a conditional probability value of, say, 0.9, this value needs to be considered in relation to other parts of the equation. Of particular importance is its relationship to the prior probabilities of the evidence and hypothesis.

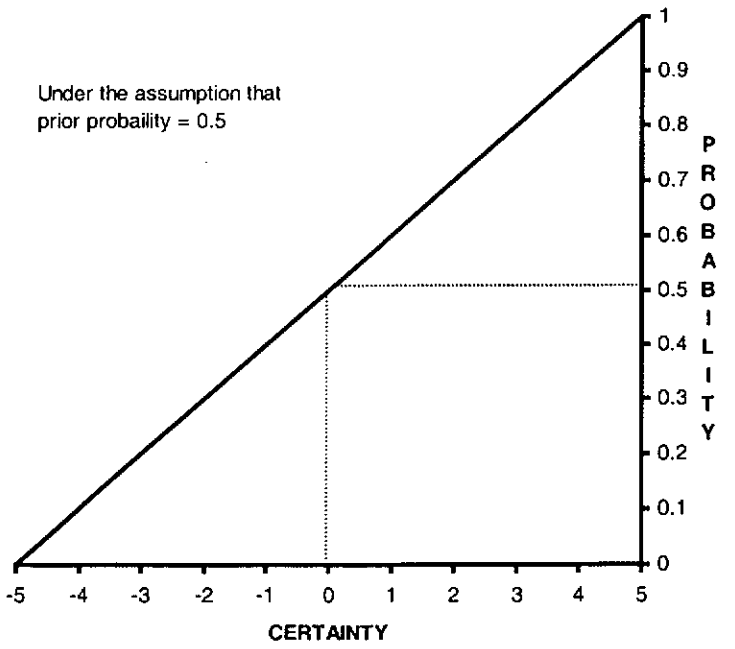

Figure 7.2 The relationship between certainty and probability in PROSPECTOR (After Katz, 1991)

Equation 7.2 shows the strict mathematical relationship between conditional and joint probability. The joint probability is constrained by the individual probabilities of the two variables. Figure 7.3 shows a Venn diagram for evidence **E** and hypothesis **H**. The shaded area is the joint probability. It is obvious from this that the joint probability cannot exceed the probability of **E** alone. In general terms for any two variables **A** and **B**, the maximum value for **P(A,B)** is whichever is the lesser of **P(A)** and **P(B)**.

Failure to consider this relationship at all times was one of the principal reasons for the problems of over-specification associated with systems such as PROSPECTOR (Rhodes and Garside, 1991). The following sections offer some suggestions as to how values may be arrived at for each of the required parameters.

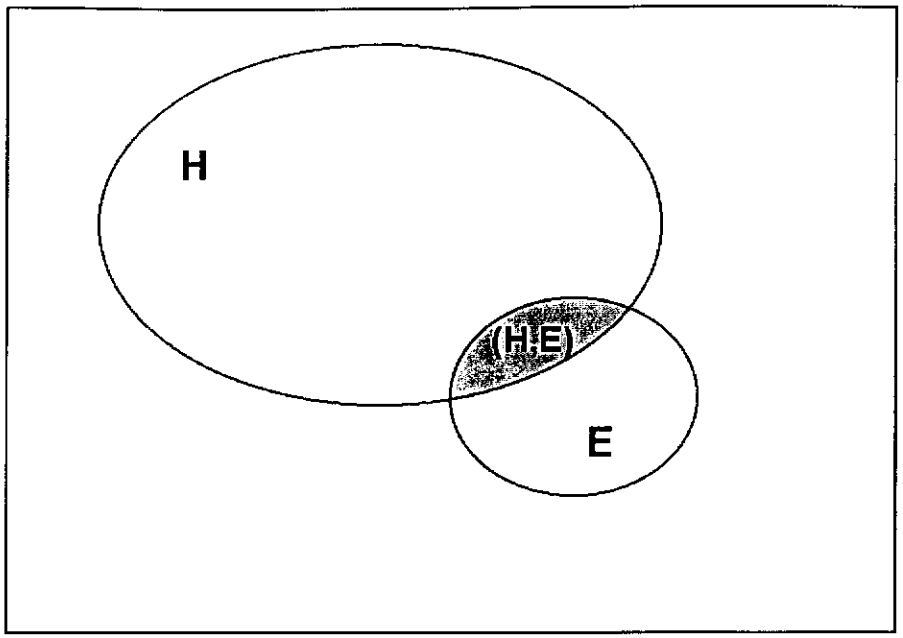

Figure 7.3 Venn diagram for evidence E and hypothesis H

7.4 Knowledge extraction and estimation of parameters.

The parameters in a Bayesian network can be regarded as expressions of knowledge about the specific physical system being described. That knowledge may be extracted either from data or from surveyor experience. Methods by which it may be extracted will be covered in more detail in a subsequent section, together with a description of software designed to assist that extraction. This section offers a brief overview of the methods in general. Although the examples given here generally refer to two state systems, they can be extended to more complex systems.

7.4.1 Prior probability of the hypothesis

The prior probability of the hypothesis is the probability distribution of the hypothesis without consideration of any evidence. This distribution may be given the symbol $P(H)$. In a context such as soil mapping, the prior distribution will vary spatially. However, in the absence of any evidence we can only assign values to it on a coarse spatial scale. In other words, we can assign only regional, or even global, general values. In the absence of any field work or 'ground truth' information, it may be necessary to assign this on the basis of experience from other, similar, areas. It is usual for fieldwork to be carried out, and that may be used to provide an estimate of an appropriate value.

A reasonable estimate of the values comprising $\mathcal{P}(\mathcal{H})$ can be obtained by examination of the site survey data. Returning to the gravel content example, let us presume that 100 sample sites within the area have been visited and a measurement made of the gravel content of the soil. We can further suppose that 27 of those sites had a gravel content which, for the purposes of this exercise, can be regarded as high. This provides a prior probability of 0.27 for high gravel content. This means that at any point in the study area, *in the absence of any other evidence*, there is a probability of 0.27 that gravel content will be high, with a probability of 0.73 that it will not be high. The capability of this for extension to a multi-state system is obvious.

An inherent problem in this method of estimation is that the sample data may be skewed or biased (as discussed in Section 7.3.1). In cases where a bias is suspected, there may be some merit in using expert opinion to override the sample-based estimate.

7.4.2 Prior probability of the evidence

The prior probability of the evidence is also a probability distribution and may be symbolised $\mathcal{P}(\mathcal{E})$. It describes the prevalence of the evidence and is always obtainable from data. It is most easily considered in a situation where the evidence is binary. Here the evidence has only two states, present or absent. That is to say, we are assuming some spatial distribution of evidence such that in parts of our area of interest this evidence exists and, in all other parts, it does not. We can estimate the values in $\mathcal{P}(\mathcal{E})$ from the magnitudes of the areas in which the evidence is found, and in which it is not. Again, this is capable of being extended to cover multi-state variables

7.4.3 Joint or conditional probability

The distributions under consideration here are the probability of the evidence given the hypothesis, $\mathcal{P}(\mathcal{E}|\mathcal{H})$, or the probability of the hypothesis given the evidence, $\mathcal{P}(\mathcal{H}|\mathcal{E})$. These can be related, using Bayes' rule and previously derived parameters, to the joint distribution of the hypothesis and evidence, $\mathcal{P}(\mathcal{H}.\mathcal{E})$. It is possible for a surveyor to make an estimate of conditional probability based on experience and

observation. In that case, it is necessary to provide some mechanism to check the consistency of such an estimate with prior probability estimates.

Similar checking is also required in the case of an estimate derived from data. If working from data, the parameter estimated is most likely to be the joint probability. One strategy for that estimation is to examine the data and count the number of times each possible combination of evidence and hypothesis occur. For a two-state hypothesis (True/False) and a two-state piece of evidence (Present/Absent) this produces a simple matrix of joint probabilities. It is necessary to ensure that the conditions referred to in Section 7.3.2 and illustrated in Figure 7.3 are fulfilled.

Unbiased sampling is critical to the success of such a method. In extreme cases a skewed sample could lead to the supposition that there was no relationship between one state of the evidence and either or both hypothesis states. Whilst this may in fact be the case, a mechanism must be provided to enable this to be checked and for the values to be overridden in the light of expert knowledge.

7.5 Uncertainty in evidence

Evidence data used in a Bayesian network is essentially operated on at the level of some finite geographic element. The most convenient way to do this is to work with the individual grid cells of a raster representation of a map. That map will have been derived from one of a number of sources, such as classified remotely sensed data, digital photogrametry, or digitised paper maps. Whatever the source, the raster representation will contain some degree of error. There will be two main components of this error, namely errors of classification and errors of location. For the purposes of describing a mechanism to cope with these errors they can be reduced to a combined 'map purity' parameter. Due to the nature of such error the discussion now proceeds to describe the case of a piece of evidence having more than two states. That is to say a grid cell whose probabilities we wish to manipulate using a Bayesian network which comes from a map comprised of a number of classes.

7.5.1 Determining map purity

For any evidence map, a table may be constructed which relates map assigned values to actual occurrences in the field. The values in this table may either be derived from sample data or assigned by an expert. Any one entry in the table represents the probability that if the map assigns a grid cell to one of these classes the class actually occurs in the field. This may be referred to as a map purity table.

If field sample data are used, the process of estimation is analogous to that described for the estimation of joint probabilities of evidence and hypothesis. In this case, the digital representation of the evidence is examined at each point for which a sample exists and the value, or class, on the map compared to that measured in the field.

Whatever the method of derivation, the result is a table having as many columns and rows as there are states in the evidence. Each column will show the conditional probability distribution that the evidence is in the state under consideration - given the evidence represented by the data source. Table 7.2 shows sample conditional probabilities for a three-class map.

Field Class	Map Class		
	1	2	3
1	0.95	0.1	0.05
2	0.05	0.8	0.05
3	0.00	0.1	0.9

Table 7.2 Conditional probabilities for a three class map

Considering the case of a grid cell mapped as Class 2, this table indicates that at that location in space there is an 80 percent chance that Class 2 actually occurs. Classes 1 and 3 each have a 10 percent chance of occurring at such a location. The distributions in such tables are not necessarily symmetrical and some cells may be zero. The exact nature of the distribution depends on the data type. Data with its origin in more continuous sources such as slope, aspect or elevation may have symmetrical or even circular distributions. Data whose origin is categorical, a geology map for example, will have a distribution governed by the degree of confusion possible between the different states, or classes, of the data.

7.5.2 Effect of map purity on evidence prior probability.

Since the Bayesian network contains a representation of the surveyor's conceptual model, the relationships portrayed by it will have been developed on a basis of true observations. That is, with the assumption that input data are 100 percent accurate. The model, therefore, assumes it is working with the real evidence as it exists in the field rather than with a map. The procedure described in Section 7.4.2 for determining evidence prior probabilities by measurement of map areas produces a probability distribution for the *map* evidence rather than for real *field* evidence. A direct determination of prior probabilities from the occurrence of evidence states in sample data is beset by problems of sampling bias, and a systematic measurement of large samples is not practicable. A procedure for converting the map class estimate of prior probability to a field class value is therefore required.

In the case of a multi-state piece of evidence, the prior probabilities represent a probability distribution which we can refer to as $\mathcal{P}(\mathcal{E})$ for the real or field situation and as $\mathcal{P}(\mathcal{E}')$ for the map data. The map purity values described above in Table 7.2 are then the conditional probability distribution $\mathcal{P}(\mathcal{E}|\mathcal{E}')$. The process of converting the values in $\mathcal{P}(\mathcal{E}')$ generated by an area count on the map can be illustrated with reference to a two class case where **E** has two states **A** and **B**.

From a cell count of the map we know the probabilities of occurrence **P(A')** and **P(B')** of map classes A B respectively. From the map purity table we also know the following:-

P(A|A')	Conditional probability that field class is A if map class is A
P(B|A')	Conditional probability that field class is B if map class is A
P(A|B')	Conditional probability that field class is A if map class is B
P(B|B')	Conditional probability that field class is B if map class is B

We also know from the total probability rule that, under the assumption that A' and B' are mutually exclusive and are exhaustive:-

$$P(A) = P(A,A') + P(A,B') \tag{7.3}$$

The joint probabilities in Equation 7.3 can be calculated by inverting the definition of conditional probability. That is:-

$$P(A,A') = P(A|A') * P(A') \tag{7.4}$$

Hence Equation 7.3 expands to;

$$P(A) = P(A|A') * P(A') + P(A|B') * P(B') \tag{7.5}$$

A similar relationship can be derived to solve for **P(B)**. **P(A)** and **P(B)** may differ from the actual prior probabilities of those classes. Generalising this for more than two classes we can derive:

$$P(E) = \sum_{E'} [P(E \mid E') * P(E')] \tag{7.6}$$

This operation needs to be performed for each individual class within the evidence. It is essentially a matrix operation that does not, at this stage, have a spatial context.

7.6 Updating the hypothesis based on map evidence

The action of taking the evidence proffered by a particular digital map and converting it into a map of the spatial distribution of an updated hypothesis is a spatial operation. It may, however, be regarded as the combination of a number of essentially non-spatial operations on individual grid cells and may be illustrated by a single grid cell. However, it is first necessary to remark on some constraints to the relationship between evidence and hypothesis.

7.6.1 Joint probability distribution

The Venn diagram in Figure 7.3 is redrawn as Figure 7.4 to show a two-state hypothesis **H** which is supported or refuted by a piece of evidence with two states **E1** and **E2**. Under the total probability rule, **E1** and **E2** must be mutually exclusive and

exhaustive. Similarly all the area outside the hypothesis **H** must be occupied by the contradictory hypothesis **!H**.

If a joint probability table is used to describe the relationships between these evidence states and the hypothesis (and its converse), then certain constraints apply. The sum of the joint probabilities of any one evidence state and all the hypothesis states must equal the probability of occurrence of that evidence state. Similarly, the sum of the joint probabilities of any one-hypothesis state and all the evidence states must equal the probability of occurrence of that hypothesis state. Due to the direct relationship between joint and conditional probability, the same constraint applies to a table, or distribution, of conditional probability. The only difference is that the members of a conditional probability distribution must sum to unity.

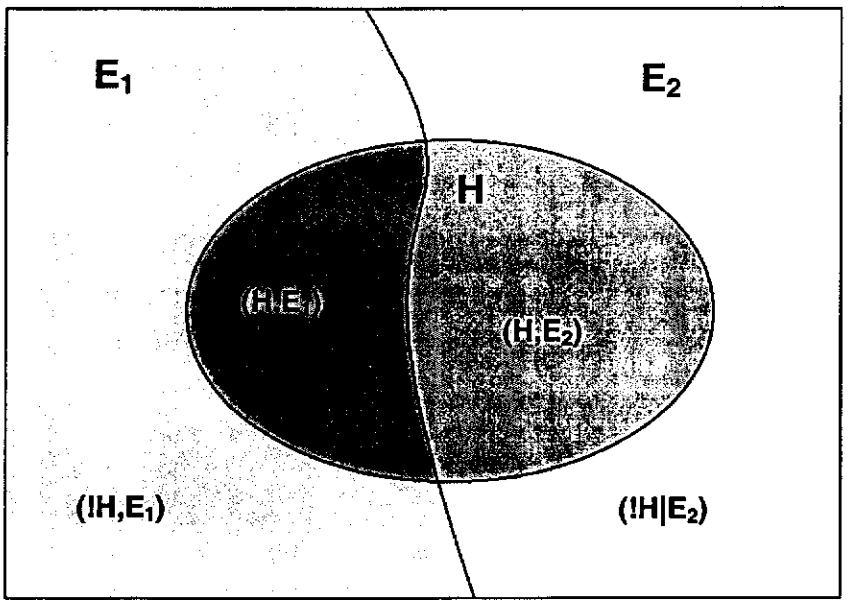

Figure 7.4 Venn diagram showing joint probability relationships

An example of a joint probability table is provided in Table 7.3. This is for the three-state evidence previously described in the discussion of map purity. It may be noted that the rows all sum to the prior probability **P(H)** of each hypothesis state and that the columns sum to the probabilities **P(E)** of each of the evidence states.

Hypothesis Class	P(H)	Field Class		
		1	2	3
1	0.67	0.135	0.33	0.205
2	0.33	0.25	0.02	0.06
Sum	1.00	0.385	0.35	0.265

Table 7. 3 Joint probability distribution for two-state hypothesis and three-state evidence layer

7.6.2 The map as evidence

We can now consider what happens, on a cell by cell basis, as a map is taken as evidence. A grid cell assigned to a state by the map has a number of associated parameters. By virtue of its map class membership, it has a conditional probability distribution of membership of all possible real, or field, classes such as that shown in any column of Table 7.2. Each of those field classes, in turn, has a joint distribution across all admissible states of the hypothesis. Such a distribution is exemplifed by one of the columns of Table 7.3. In addition, we have a prior probability of occurrence of that particular class on *the map*. In order to populate a table such as Table 7.3, that must be converted to a prior probability for that class *in the field*. To do this we use the procedure outlined in Section 7.5.2. and considered in more detail below.

The map area denoted by that grid cell will also have a prior probability distribution $\mathcal{P}(\mathcal{H})$ across all states of the hypothesis. The task of the Bayesian network, simplified in this case to only one piece of evidence, is to update that probability distribution to $\mathcal{P}(\mathcal{H}|\mathcal{E}')$,where $\mathcal{E}'$ indicates that we are using map evidence.

Using the defining relationship of conditional probability, we know that:-

$$\mathbf{P(H|E)} = \frac{\mathbf{P(H,E)}}{\mathbf{P(E)}} \tag{7.7}$$

For any one grid cell, **P(H|E)** may be written long hand as :

P(Hypothesis = Class j | Field class is Class i)

Since we are working from map data the quantity of interest is :-

P(Hypothesis = Class j | Map class is Class i)

The uncertainty in class membership has been quantified as a distribution describing the probabilities of occurrence of the field classes at that location. By analogy with Equation 7.6 and using the total probability rule, we can calculate the probability that a particular hypothesis class exists at that location by summing the contributions made to it by all of the possible field classes. That is the sum over the field classes of:-

P(Hypothesis = j| Field class = i) * P(pixel is a member of field class i)

In this case, the second term in this expression is in fact :-

P(Pixel is in field class i | pixel has the value given in the map)

which is the Map purity value. We can therefore summarise as:-

$$P(H \mid E') = \sum_{e} [P(H \mid E) * P(E \mid E')] \qquad (7.8)$$

Equation 7.8, forms the central calculus of a Bayesian 'expert system' which may be used to map probabilities of occurrence of some attribute (the hypothesis) based on a number of pieces of uncertain evidence.

7.6.3 A graphical representation of the calculus

The tree diagram in Figure 7.5 shows an example of this calculus using the values in Tables 7.2 and 7.3. In order to understand this diagram it is first necessary to convert the joint probability distribution shown in Table 7.3 to a conditional probability distribution. This is achieved by dividing though by the prior probability of the appropriate evidence class (Equation 7.7). The results of this are shown in Table 7.4.

Hypothesis Class	P(H)	Field Class		
		1	2	3
1	0.67	0.35	0.94	0.77
2	0.33	0.65	0.06	0.23
Sum	1.00	1.00	1.00	1.00

Table 7. 4 Conditional probability distribution for two-state hypothesis and three-state evidence layer.

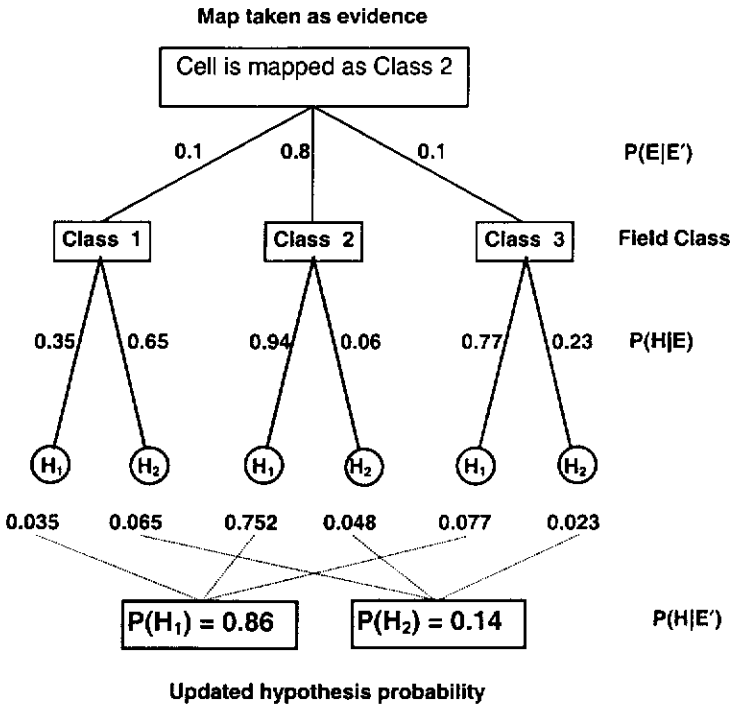

Figure 7.5 A graphical representation of the process of taking a map as evidence

The values shown in Figure 7.5 are representative of any grid cell in a digital map. If we consider a cell which has been mapped as belonging to Class 2, then, using the map purity figures in Table 7.2, it has a probability of 0.8 of truly belonging to that class. There is a probability of 0.1 that it belongs to each of the other two classes.

If this cell truly belongs to Class 2 then, using the values in Table 7.4, the probability of it of being assigned to hypothesis state 1 is 0.94. There is a concomitant probability of 0.06 that it belongs to hypothesis state 2. However, the imprecision

represented by the map purity table modifies these values to 0.752 and 0.048 respectively. Similarly, there are other measures of its probability of membership of the two hypothesis classes derived from the probabilities that it really belongs to one of the other two field classes. The overall probabilities of this grid cell belonging to each hypothesis class are then found by summation of the contributions from each of the 'streams'.

7.7 Summary

This chapter has described the means by which a Bayesian network can be harnessed to provide the expert system component of a quantitative soil mapping system. Algorithms and methods have been described for obtaining the parameters and probability distributions necessary to populate a Bayesian net. The problems of using inexact data have been discussed and a practical method of handling it described. The process of taking a map as a piece of evidence has been described with detailed discussion of the calculations performed for a particular grid cell. The next chapter will discuss detailed algorithms for the GIS component of a soil mapping system.

Chapter 8

GIS INTERFACING AND DATA COMBINATION

The quantitative soil mapping system described in Chapter 6 has two components. One is an expert system tool, based on Bayesian network algorithms which were described in the preceding chapter. In addition, there are a number of tasks which require manipulation of data in the spatial domain. Geographic information systems may be harnessed to perform these operations, either directly or through the use of their scripting languages. This chapter describes the algorithms and procedures required to carry out these spatial tasks. A number of the GIS tasks use common algorithms which are encoded as a standard part of most commercial GIS. It is not intended to discuss those standard algorithms in detail. However, where a choice exists between methods of performing the same task, a discussion is provided as to which is most suitable.

The overall assemblage of expert system tool and GIS linkages that has resulted from this work is called the Expector method. The name was chosen partly to acknowledge the impetus provided to its development by the PROSPECTOR method and partly because the system represents the *expectations* of the soil surveyor. The system is described in detail in Chapter 9, but reference is made in this chapter to some of its file formats in order to illustrate the linkage methods employed.

In Expector, some of the more advanced tasks are coded in both ARC/INFO and ArcView as interface modules to the expert system. However, at the most basic level, the general algorithms consist of a series of the 'primitive' operations which are at the core of any GIS possessing map-algebra capabilities. They are, therefore, capable of being implemented in almost any GIS.

By referring back to Figure 6.3 (p. 66), it can be seen that the major tasks to be performed in a GIS are those of data preparation, combination and presentation. In addition to these, this chapter covers tasks involved in the extraction of knowledge from spatial data. The transfer of that knowledge between the GIS and the expert system is also covered here. We begin with the preparation of spatial data.

8.1 Data preparation - general considerations

The data to be used for soil mapping will essentially be of two types. There will be site sample data and extensive map data covering the entire area of interest. The extensive data could be in one of several forms. These include categorical data, typically expressed as a choropleth map, and continuous data, represented either as a raster or as a contour map in vector form. Each of these various data types needs to be made ready for interface to the expert system tool. The first task is to ensure that all the spatial data are in a common spatial reference system.

8.1.1 Geo-coding of site and map data

It is a pre-requisite for its use in a GIS that any data presented for analysis has a co-ordinate reference. It is desirable that the work be conducted in a linear metric co-ordinate framework. Appropriate grid systems, therefore, include projections such as the Australian Map Grid (AMG) or Universal Transverse Mercator (UTM). The necessity, or otherwise for transformation of data to this common reference framework will depend on its provenance. Original data may have been collected in one of these reference systems, some other metric or imperial grid system, or in geographical co-ordinates (latitude and longitude).

Recently collected site data is likely to have been geo-coded using Global Positioning System (GPS) techniques. The data source needs to be checked to ensure that the geodetic parameters (spheroid, datum, etc.) used in the GPS surveying are in accord with those chosen for the overall analysis framework. If they are not, then a data transformation needs to be undertaken. Most GIS have facilities for such transformations.

If the site sample data is less recent, its co-ordinates may have been derived either by traverse surveying or by the digitising of spot locations from aerial photography. In these cases, reference should again be made to the original sources to check geodetic and projection parameters, with transformations being performed as necessary.

In some extreme cases, site sample locations may be described in some form other than a co-ordinate system. Although procedures such as the conversion of cadastral lot descriptions to co-ordinates, and the use of ortho-photos to locate points from

descriptions can be applied, the spatial accuracy of such data is likely to be questionable. The use of the spatial properties of such data needs, therefore, to be undertaken with care.

8.1.2 Co-registration of data

Extensive data sources include remotely sensed data, digital elevation models and digitised paper maps. Remotely sensed data will generally have been geo-referenced using the facilities provided by digital image processing software. Again it needs to be checked to ensure that geodetic and projection parameters conform to those chosen for the rest of the analysis. The same applies to digital elevation models. Digitised paper maps and similar sources may have originally been generated in geographical co-ordinates and may also require transformation.

The Expector system has been designed to use raster data. At an early stage of data preparation, consideration needs to be given to the spatial resolution of the data. Given that the data will have come from a variety of different sources, the Expector has the capability to handling data of varying cell sizes. It must, however, be recognised that in any data combination process the coarsest resolution data will have a significant contribution to the result. Nevertheless, higher resolution data which provides significant evidence will have the effect of 'breaking up' coarser, more patchy data. Some of the operations required to prepare the data will involve the adjustment of cell size by resampling.

Data should also be co-registered so that each raster has a common origin. It is also advisable that, where multiple resolution data sets exist, there is a factorial relationship between their cell sizes. For example, remotely sensed data are frequently re-sampled, during rectification, to 25m pixels with their origin on an even hundred in both easting and northing. It would be appropriate to rasterise a coarser scale digital geology map at a resolution of 100m, again with an origin on the even hundreds.

It is also necessary that all evidence data sets have the same spatial extent. Where data are missing for part of the area, a 'no-data' value, appropriate to the GIS being

used, should be set. The software contains a mechanism for handling such missing data areas (see Section 8.5.2).

8.2 Preparation and use of extensive evidence data

The expert system component described in the previous chapter requires categorical raster data input. The data that a soil surveyor wishes to use may be available in a number of alternative forms. These include vector polygon data, continuous raster surfaces, surfaces represented as vector contours, and transect data from sources such as airborne geophysical profiling instruments. All of these data forms require conversion to categorical raster data.

8.2.1 Selection of categories

Before converting any of this evidence data to a categorical raster format, the number and size of those categories need to be defined. Each category represents one state of the evidence. The number of states used for a piece of evidence, and the bounds between them, may vary according to the hypothesis attribute. That is, one attribute may be more sensitive to changes in state of a particular evidence variable than another. For example whilst organic matter may only begin to accumulate on slopes lower than, say 2 percent, ironstone nodule content may be sensitive to slope class breaks at, say 1 percent, 3 percent, 5 percent, and 7 percent.

Care needs to be exercised in the choice of class breaks if an evidence layer is to be used for multiple hypotheses in turn. Consideration also needs to be taken of the number of states in the hypothesis variable. It may be more efficient to create, in the example above, two evidence layers representing slope. One for use in the prediction of organic matter content, with two states (0-2 percent and >2 percent). The other for prediction of ironstone nodules, with 5 states (0-1 percent, 1-3 percent, 3-5 percent, 5-7 percent and >7 percent).

8.2.2 Preparing existing raster data

Some existing raster data, such as maps of landcover classes, will already be categorical. Other data may be stored as a continuous raster representation of a surface. Both types of datasets will require reclassification to conform to the states determined for that variable. Reclassification is a common GIS task.

For continuous data, a lookup table is constructed, often using an on-screen form. The lookup table matches ranges of values in the continuous data to categorical classes. Similarly, for data that is already categorical, the lookup table reassigns the classes to those appropriate for the analysis. The principal difference is that, in the case of continuous data, the classes will be formed from homogenous values, whereas with categorical data, assignment is made on the basis of the *meaning* of the class label rather than the absolute value. In both cases, it is good practice to save the lookup table used for the transformation. It can prove useful both as a record of the classification and as a template for further, similar, reclassifications.

8.2.3 Vector polygon data

The treatment of vector polygon data is similar to that of categorical raster data, since some degree of reclassification using a lookup table may be required. An additional step, for which most GIS provide standard tools, is the rasterisation itself. The comments regarding transformation, cell size and co-registration discussed in Section 8.1.2 are particularly relevant here.

Commercial GIS generally offer a choice in the treatment of cells whose position falls on the boundary between categories. Typical algorithms assign cell values to one of the following categories:-

a) The category which occupies most of the grid cell,
b) The category in which the centre of the grid cell falls,
c) An area weighted mean of the possible categories.

When working with thematic categorical data, option '**c**' is not appropriate since it will introduce cells with spurious, possibly non-integer, values. Options '**a**' and '**b**' are both entirely suitable. It is up to the individual analyst to decide which best suits the data. Consideration should be given to the scale of the original survey, the size of the smallest polygons and the intended raster grid cell size.

8.2.4 Using other extensive data

Whilst there are a number of other data types which require conversion, the most common will be data in the form of contour strings or linear profiles. Both have in

common the fact that the data are better represented in one direction than in the other.

As an example, profile lines from a geophysical profiling instrument, may be two hundred metres or more apart, whilst the sampling rate along the lines may have resulted in a data point every thirty metres or so. Most GIS will provide algorithms for generating surfaces from point data. Simple interpolation methods that use distance weighting are inappropriate for such data, due to the generally non-isotropic nature of its spatial distribution. More sophisticated methods including spline interpolators or geo-statistical techniques such as kriging, are more appropriate for these data types. Davis (1986) provides guidance in the selection of techniques.

Vector contour data which has been digitised from map compilation sheets using a line-following scanner has a particularly high sample rate along the lines and it may be necessary to weed out some of the data points in the along contour direction prior to using a spline interpolator. Similarly, point data that are to be used as evidence will require conversion to a raster representation of a surface. The interpolation method required will depend on the nature and spacing of the sample points, and each case will require separate consideration. Davis (1986) and Burrough (1986) provide discussions of various interpolation routines and their appropriateness to particular circumstances.

8.2.5 Derivation of indices

During the model building phase within the expert system, some evidence variables may be envisaged which are not direct categorisations of existing data. These may include constructs such as compound topographic (wetness) index, indices of solar irradiation, or other terrain-based attributes such as local relief. All of these require a certain degree of processing within GIS. Once the indices have been calculated the grid data sets should then be categorised appropriately.

8.3 Preparation and use of site sample data

As indicated in Chapter 7, site sample data will be used in three ways: firstly to determine the prior probability distribution of the hypothesis (the attribute being mapped), secondly as an aid to constructing joint probability distributions with the

various evidence variables, and thirdly to determine the 'purity' of each of the evidence variables. These are all ways in which knowledge is extracted from the data.

8.3.1 Prior probability of the hypothesis

As described in Section 7.4.1, the prior probability distribution of the hypothesis may be estimated from the number of occurrences of each hypothesis state present in the site data. If the hypothesis attribute is a categorical variable, this presents no problem, but continuous variables must be rendered discrete. This requires the selection, by an expert, of the range of each state within the variable. Once this is done, the prior probability distribution can be determined by ordering and counting the data points which occur in each category. This task can either be performed using the database manipulation functions of the GIS or in a separate database or spreadsheet. The procedure is simplified if a data file is constructed which contains only four columns. The record for each point should comprise an identifier, an easting, a northing and the appropriate hypothesis state, expressed numerically.

Even if the site data is not spatially referenced, or if the accuracy of the referencing is suspect, it can still be used to develop estimates of abundance for use as hypothesis prior probabilities. This is possible because the estimation of the prior distribution is aspatial and is generally a regional distribution that will be modified for each grid cell as the evidence is considered.

8.3.2 Determining the joint probabilities

Where accurately spatially referenced site data are available, they may be used to determine the coincidence between the hypothesis states and the evidence states, at least as represented by that sample. This is done by querying each of the categorised evidence layers at each sample point in turn. Most GIS have facilities for on-screen query of individual cell values. In addition, such queries can usually be initiated by entering co-ordinate pair values at the command line. Neither of these methods is practicable for a reasonably sized sample data set.

The algorithm for this procedure, for any one evidence variable, proceeds as follows:-

- Open hypothesis data file (see Section 8.3.1.for format),
- Open an output file,
- Read a record (identifier , co-ordinates and hypothesis state),
- Extracted the co-ordinate values from record,
- Query the GIS data layer to determine the evidence category at that point,
- Create new record comprising point identifier, hypothesis state, evidence state,
- Write record to output file,
- Read next record from input file,
- Repeat until all input data read,
- Close files,

This procedure has been coded in the scripting languages of ARC/INFO and ArcView as part of the Expector package. Although standard spreadsheet tools may be used to examine the output and to calculate joint probabilities, the Expector software provides tools for this as part of the knowledge editing process.

8.3.3 Prior probabilities of the evidence

This probability distribution is, essentially, a numerical expression of the histogram of the evidence data classes. In some GIS, this histogram is stored as a table in an associated database; in others, the information needs to be extracted from the raw data. A procedure is, therefore, required to either report the values from the data base table, to decode a listing of the histogram produced by the GIS, or to derive the area counts directly from the raw data file. As part of the Expector software package, procedures have been written to access the database tables associated with ARC/INFO and ArcView grid data files and write the values to a separate file. The algorithms necessary to decode other GIS-specific histogram files and data sets will vary from system to system and are, therefore, not discussed in detail here.

8.4 Data and knowledge exchange

Referring again to Figure 6.3 (p. 66), there are two main points at which interchange is required between the GIS and the expert system tool. These are the passing of

knowledge from the GIS as part of the knowledge definition and editing stage and the passing of data weights back to the GIS prior to a data combination stage.

8.4.1 Passing knowledge

The knowledge that is to be passed between the GIS and the expert system tool is in the form of probability distributions. The prior probability distribution of the hypothesis has been determined using the procedure described in Section 8.3.1. This is a one dimensional distribution with as many members as there are states in the hypothesis. Since this has been determined outside the GIS, usually in a spreadsheet, facilities have been provided for entering the values directly into the expert system tool.

The joint probabilities of the hypothesis and evidence have been determined by the method in Section 8.3.1 as a coincidence table. Table 8.1.a) is a direct extract from a coincidence table and does not contain a header line. The first column is a numeric identifier for the sample site, the second column is the hypothesis state at that point and the third is the evidence state at that point. The Expector software reads this table and converts it to an initial joint probability table which can then be edited as required.

```
1,2,4
2,1,3
3,1,3
4,1,2
5,1,2
6,1,3
...,...,..
...,...,..
217,2,4
```

a)

```
"Value","Count"
1,1186
2,23388
3,27104
4,13697
5,8842
```

b)

Table 8.1 Extracts from Expector data interchange files.
a) Coincidence table showing point ID, hypothesis state and evidence state
b) Evidence probability distribution

The interchange file format has been deliberately constructed as a comma delimited ASCII file in order to enable the easy construction of interfaces to GIS packages. The evidence prior probability distributions are also passed to the Expector software as comma delimited files, an example of which is shown in Table 8.1.b). The two columns in this file refer to the unique values present in the evidence data set and a count of the number of cells in each category. To enable their files to be read by the expert system, interfaces to other GIS must use exactly the same format.

8.4.2 Passing weighted data back to the GIS

As described in Section 7.6, for each evidence layer the expert system tool first produces an updated probability distribution for the hypothesis over the evidence states. This distribution $P(\mathcal{H}|\mathcal{E}')$, is two dimensional. Table 8.2 shows an example of such a distribution; again this is a direct extract from the expert system data file.

```
value,pheq1,pheq2
1, 77069, 22930
2, 65000, 34999
3, 49050, 50950
4, 16950, 83050
5, 2500, 97500
```

Table 8.2 Expector data file representing an updated probability distribution

This is a distribution of a two-state hypothesis across a five-state evidence layer. The first line in the file contains column labels and is followed by as many lines as there are states in the evidence. Each of those lines comprises the category value for the state then the values of $\mathbf{P(H_j|E_i)}$. For any hypothesis state **j** the distribution thus reads down the column. The values in this file are probabilities and they should lie between 0 and 1. Since some GIS have difficulty importing floating point data, they have been scaled into the range 0 to 100,000. Although this may cause some apparent loss of precision in the values, it is debatable whether changes in probability values as small as 0.00001 are meaningful. In common with the files that transfer knowledge into the expert system tool, a comma delimited ASCII structure has been used for maximum portability.

In order to more fully understand the values in Table 8.2, we may take this as being an example of the joint distribution between a slope layer, categorised in to five classes and a 'high organic matter' hypothesis with two states, present (pheq1) and absent (pheq2). The first line of the table then reads as a mathematical expression of the belief that if the map assigns a pixel to the low slope category, then high organic matter is more probable than low organic matter in the ratio 77:23. Other lines reflect belief ratios for other slope classes.

8.5 Combining probabilities from several evidence layers

Once the individual updated values for the hypothesis probability distribution of each of the evidence layers have been passed back to the GIS, it is necessary to combine them into one distribution for the hypothesis. So far the expert system tool has essentially operated a-spatially on the individual evidence classes. Since the class boundaries in physical space differ between evidence layers, this combination needs to be carried out on an individual cell by cell basis throughout the grids involved. The process involved uses map algebra and the exact detail will vary according to the implementation of map algebra in the particular GIS in use. The following describes the mathematical basis behind the combination and is followed by a description of the implementation of the algorithm in ARC/INFO.

8.5.1 General mathematical principles

We must consider the situation of a grid cell for which we have multiple evidence streams supporting a particular hypothesis. For the sake of simplicity we will begin by considering a two-state hypothesis **H** with states $\mathbf{H_1}$ and $\mathbf{H_2}$, and two streams of evidence which we will call $\mathbf{E_1}$ and $\mathbf{E_2}$.

$\mathbf{E_1}$ and $\mathbf{E_2}$ are map variables and the values of parameters in this discussion represent the actual values pertaining at the particular grid cell based on that cell's membership of a particular class on each of the evidence maps. We have previously calculated, in the expert system tool, values for the following parameters.

$\mathbf{P(H_1|E_1)}$ The probability of hypothesis state 1 given evidence layer 1

$\mathbf{P(H_1|E_2)}$ The probability of hypothesis state 1 given evidence layer 2

$\mathbf{P(H_2|E_1)}$ The probability of hypothesis state 2 given evidence layer 1

$\mathbf{P(H_2|E_2)}$ The probability of hypothesis state 2 given evidence layer 2

From these we wish to calculate two pooled values. These are :-

$P(H_1|E_1,E_2)$ The probability of hypothesis state 1 given both evidence layer 1 and evidence layer 2, and

$P(H_1|E_1,E_2)$ The probability of hypothesis state 2 given both evidence layer 1 and evidence layer 2.

We need only consider the calculation for one hypothesis state, since the other proceeds analogously. The first part of the following derivation is after Cohen (1985).

We know from Baye's theorem (Equation 3.11) that for any one hypothesis state H_1 and the event of taking any one piece of evidence E:-

$$P(H_1 \mid E) = \frac{P(E \mid H_1) \cdot P(H_1)}{P(E)} \tag{8.1}$$

If H_1 and H_2 are mutually exclusive then we also know that:-

$$P(E) = P(H_1, E) + P(H_2, E) \tag{8.2}$$

However, we also know, from the definition of conditional probability, (Equation 2.3) that:-

$$P(H, E) = P(E \mid H) \cdot P(H) \tag{8.3}$$

So, from Equations 8.2 and 8.3:-

$$P(E) = P(E \mid H_1) \cdot P(H_1) + P(E \mid H_2) \cdot P(H_2) \tag{8.4}$$

Substituting Equation 8.4 into Equation 8.1 gives :-

$$P(H_1 \mid E) = \frac{P(E \mid H_1) \cdot P(H_1)}{P(E \mid H_1) \cdot P(H_1) + P(E \mid H_2) \cdot P(H_2)} \tag{8.5}$$

If we now replace the event of considering one evidence layer only with the event of considering two evidence layers E1 and E2 then (under the assumption that the evidence layers are independent) we can now rewrite Equation 8.5 as:-

$$P(H_1 \mid E_1, E_2) = \frac{P(E_1, E_2 \mid H_1) \cdot P(H_1)}{P(E_1, E_2 \mid H_1) \cdot P(H_1) + P(E_1, E_2 \mid H_2) \cdot P(H_2)} \tag{8.6}$$

Equation 8.6 can be generalised to provide a means of calculating the pooled posterior probability for any member $\mathbf{H_j}$ of a suite of **n** hypothesis states given the event of taking into consideration **m** evidence layers. That is:-

$$P(H_j \mid E_1, E_2, \ldots\ldots E_m) = \frac{P(E_1, E_2, \ldots\ldots E_m \mid H_j) \cdot P(H_j)}{\sum_{j=1}^{n} P(Hj) \cdot P(E_1, E_2, \ldots\ldots E_m \mid Hj)} \tag{8.7}$$

The solution for $P(H_j \mid E_1, E_2, \ldots\ldots E_m)$ in Equation 8.7 does not quite suit our purposes for two reasons. Firstly, it does not use the individual $P(H_j|E_m)$ values provided by the expert system tool and secondly, evaluation of the denominator requires that we know the conditional probabilities of all possible combinations of the states of the **m** evidence layers with all **j** states of the hypothesis. In a moderately complex problem this rapidly becomes unattainable and will be dealt with first.

In order to proceed we need to assume that the individual evidence layers are conditionally independent. Conditional independence and the operational ramifications of this assumption are discussed at greater length in Chapter 12. This assumption of evidence data layers as being conditional independence of data may be stated, for two evidence layers, as (Cohen,1985):-

$$P(E_1, E_2 \mid H_1) = P(E_1 \mid H) \cdot P(E_2 \mid H) \tag{8.8}$$

Generalising this assumption for more than two cases we can now rewrite equation 8.7 as:-

$$P(H_j \mid E_1, E_2, \ldots\ldots E_m) = \frac{P(H_1).\prod_{i=1}^{m} P(E_i \mid H_1)}{\sum_{j=1}^{n}\left\{P(Hj).\prod_{i=1}^{m} P(E_i \mid Hj)\right\}} \qquad (8.9)$$

Equation 8.9 is the standard method used for updating of probabilities in Bayesian networks Cohen, 1985, p. 30). It calls, however, for the derivation of estimates of individual values of **P(E|H).** This is the probability that the *evidence exists* given that the *hypothesis is true*. From the point of view of a soil surveyor, this is a less intuitive value than the probability that the *hypothesis is true* given the fact that the *evidence exists*. For that reason the expert system tool described in Chapter 7 was designed to provide the latter value. However we know from Bayes' rule that :-

$$P(H_j \mid E_i) = \frac{P(E_i \mid H_j) \,.\, P(H_j)}{P(E_i)} \qquad (8.10)$$

This allows two courses of action. We can calculate individual values of $P(E_i|H_j)$, either in the expert system tool or as a first step calculation in GIS, or we can rewrite Equation 8.9 in such a way as to use the $P(H_j|E_i)$ values. The derivation now departs from that of Cohen (1985) and other standard works.

Returning to the two evidence layer situation we restate Equation 8.6:-

$$P(H_1 \mid E_1, E_2) = \frac{P(E_1, E_2 \mid H_1) \,.\, P(H_1)}{P(E_1, E_2 \mid H_1) \,.\, P(H_1) + P(E_1, E_2 \mid H_2) \,.\, P(H_2)} \qquad (8.11)$$

Rewriting the numerator under the assumption of conditional independence expressed in Equation 8.7 gives:-

$$P(H_1 \mid E_1, E_2) = \frac{P(E_1 \mid H_1) \,.\, P(E_2 \mid H_1) \,.\, P(H_1)}{P(E_1, E_2 \mid H_1) \,.\, P(H_1) + P(E_1, E_2 \mid H_2) \,.\, P(H_2)} \qquad (8.12)$$

Inverting Equation 8.10 then gives:-

$$P(E_i \mid H_j) = \frac{P(H_j \mid E_i).P(E_i)}{P(H_j)} \tag{8.13}$$

Substituting Equation 8.13 into the numerator of Equation 8.12 gives :-

$$P(H_1 \mid E_1, E_2) = \frac{\left\{P(H_1).\frac{P(H_1 \mid E_1).P(E_1)}{P(H_1)}.\frac{P(H_1 \mid E_2).P(E_2)}{P(H_1)}\right\}}{P(E_1, E_2 \mid H_1).P(H_1) + P(E_1, E_2 \mid H_2).P(H_2)} \tag{8.14}$$

By analogy with Equation 8.13 :-

$$P(E_1, E_2 \mid H_j) = \frac{P(H_j \mid E_1, E_2).P(E_1, E_2)}{P(H_j)} \tag{8.15}$$

The denominator of Equation 8.14 can now be written as:-

$$\left\{P(H_1).\frac{P(H_1 \mid E_1, E_2).P(E_1, E_2)}{P(H_1)} + P(H_2).\frac{P(H_2 \mid E_1, E_2).P(E_1, E_2)}{P(H_2)}\right\} \tag{8.16}$$

Simplifying Equation 8.16 and collecting terms we now rewrite 8.14 as:-

$$P(H_1 \mid E_1, E_2) = \frac{P(H_1).P(E_1).P(E_2)\left\{\frac{P(H_1 \mid E_1)}{P(H_1)}.\frac{P(H_1 \mid E_2)}{P(H_1)}\right\}}{P(E_1, E_2)\{P(H_1 \mid E_1, E_2) + P(H_2 \mid E_1, E_2)\}} \tag{8.17}$$

However under the assumption of conditional independence:-

$$P(E_1, E_2) = P(E_1).P(E_1) \tag{8.18}$$

Equation 8.17 now simplifies to:-

$$P(H_1 \mid E_1, E_2) = \frac{P(H_1).\left\{\frac{P(H_1 \mid E_1)}{P(H_1)} \cdot \frac{P(H_1 \mid E_2)}{P(H_1)}\right\}}{\{P(H_1 \mid E_1, E_2) + P(H_2 \mid E_1, E_2)\}} \tag{8.19}$$

This can be generalised for multiple hypotheses and multiple evidence layers as:-

$$P(H_1 \mid E_1, E_2, \ldots E_m) = \frac{P(H_1).\prod_{i=1}^{m}\left\{\frac{P(H_1 \mid E_i)}{P(H_1)}\right\}}{\sum_{j=1}^{n} P(H_j \mid E_1, E_2, \ldots E_m)} \tag{8.20}$$

Equation 8.20 now gives us a means of combining the individual P(H|E) values produced by the expert system tool.

8.5.2 ARC/INFO algorithm implementation

The combination of probabilities from several evidence layers has been implemented as a routine in Arc Macro Language (AML) as part of the Expector ARC/INFO interface. The algorithm assumes that there is available, for each evidence layer, a table (like Table 8.2) which gives the probability of each hypothesis state for each evidence state or class. The operation of the algorithm also requires a basic parameter file that contains the prior probability distribution for the hypothesis and a list of evidence layer to be used.

The algorithm starts by reading this parameter file and then, using the GIS native database abilities, joins the appropriate probability table (like Table 8.3) to the attribute table for each evidence layer. This has the effect of creating a virtual grid dataset for each hypothesis-state/evidence-state combination. Within each hypothesis state, a loop is initiated which, using map algebra, creates a grid dataset for each hypothesis state which is the product of the ratios of the individual evidence grid cell $P(H_i|E_j)$ values and $P(H_i)$.

Map algebra operates on a cell by cell basis and, where a no-data value is encountered in one of the evidence layers, the value of $P(H_i)$ is substituted for

$\mathbf{P(H_i|E_j)}$. This has the effect of giving the multiplier for that cell a value of one, indicating that the evidence has no effect there. This is then multiplied by $\mathbf{P(H_i)}$ to give the numerator of the right hand side of Equation 8.20. The grids representing these accumulated values are then summed in a loop which runs through each of the hypothesis states and performs the normalisation which completes the evaluation of Equation 8.20.

8.6 Data presentation

The result of the data combination process described in Section 8.5 is a number of grid data sets, one for each state of the hypothesis. Each of these shows the spatial distribution, as predicted by the evidence used, of the probability of occurrence of that hypothesis state. These grids are floating point datasets with values lying between 0 and 1. They may be presented to the user either as probability grids with a suitable display scale or they may be used to derive a further grid that shows the most probable hypothesis state.

8.6.1 Display of probability data

In order to cope with data scaling problems associated with the display of floating point data in some GIS, the Expector software package provide tools to re-scale the data into the range 0 to 100, thus expressing the probabilities as integer percentages. A colour table distributed with the software package has 100 entries covering this range to facilitate the display of probability data.

8.6.2 Derivation and display of most probable state maps

Conversion of the data to a most probable state map requires that the individual data sets pertaining to each hypothesis state be queried on a cell by cell basis to determine which has the highest value at each grid cell. Some GIS provide a standard function to perform this operation (for example, the UPOS function in ARC/INFO). For other systems, a tool needs to be constructed in the appropriate scripting language. The algorithm is detailed here.

- Create an 'index grid' of same extent and cell size as hypothesis grids,
- Set all 'index grid' cells to 0,
- Create an 'maximum grid' of same extent and cell size as the index grid,
- Set all 'maximum grid' cells to 0,
- Set a counter to 1,
- Compare probability grid for first hypothesis state with 'maximum grid',
- If hypothesis state grid cell value is greater than 'maximum grid' value, set 'index grid' value to counter value and set 'maximum grid' cell value to hypothesis state grid value,
- Else leave index grid value unchanged,
- Increment counter and repeat previous two steps,
- When all hypothesis grids compared, save 'index grid' and delete 'maximum grid'.

An implementation of this algorithm in Avenue, the scripting language of ArcView, is incorporated in the Expector software package. As indicated above, in ARC/INFO the procedure is accomplished by an in-built function.

8.7 Summary

A number of GIS processes are required to make spatial data usable in an expert system tool and to combine the results of calculations carried out by that tool. Processes common to the treatment of site sample data and extensive (map) data sets include geo-coding and co-referencing. In addition, extensive data may require reclassification. Following the definition of hypothesis states, site sample data may also require reclassification.

Once prepared, the site sample data may be used in knowledge extraction processes. The extracted knowledge includes the prior probability distribution of hypothesis states as well as sample joint distributions between the hypothesis and the various evidence variables. A number of procedures and file structures by which this knowledge can be passed to the expert system tool have been discussed, as have those for returning processed knowledge back to the GIS.

The combination of individual evidence layers in GIS has been described, firstly in its general mathematical principles and secondly as an algorithmic implementation. Although the mathematical principals used are based on standard Bayesian network calculus, they have been modified to use probability distributions that are more intuitively meaningful to a soil surveyor. Algorithms have also been presented for the display of combined data, both as probability maps and as most probable state maps. The next chapter gives a detailed description of the operation of the Expector natural resource mapping software.

Chapter 9

A DESCRIPTION OF THE EXPECTOR SOFTWARE

This chapter provides a discussion of the operation of the Expector software. It is not intended as a reference manual for the user. That forms a separate document which is contained on the Expector software distribution disk included with this thesis.

9.1 Components of the software

The overall design of the Expector software, as discussed in Chapters 7 and 8, has resulted in two main software components. One is a standalone application that performs the knowledge editing functions described in Chapter 7. The other is a GIS specific component which handles the data preparation, interfacing, and data combination tasks as described in Chapter 8.

The standalone software was written in Microsoft Visual Basic™ and provides a forms type interface. The GIS specific software has been written in Avenue and AML, the native scripting languages of ArcView and ARC/INFO. Simultaneously with the development of this software, a complementary interface to the Intergraph Microstation GIS Environment software has been written by others. This description of the software covers the standalone component and the ArcView interface.

9.2 An overview of the Expector process

The process of quantitative natural resource assessment was discussed at length in Chapter 6. Figure 9.1 shows a flowchart for the Expector implementation of that process. The following stages are defined:-

a) Knowledge definition,
b) Data preparation,
c) Knowledge extraction,
d) Data weighting (Knowledge editing),
e) Data combination,
f) Map preparation (Data display).

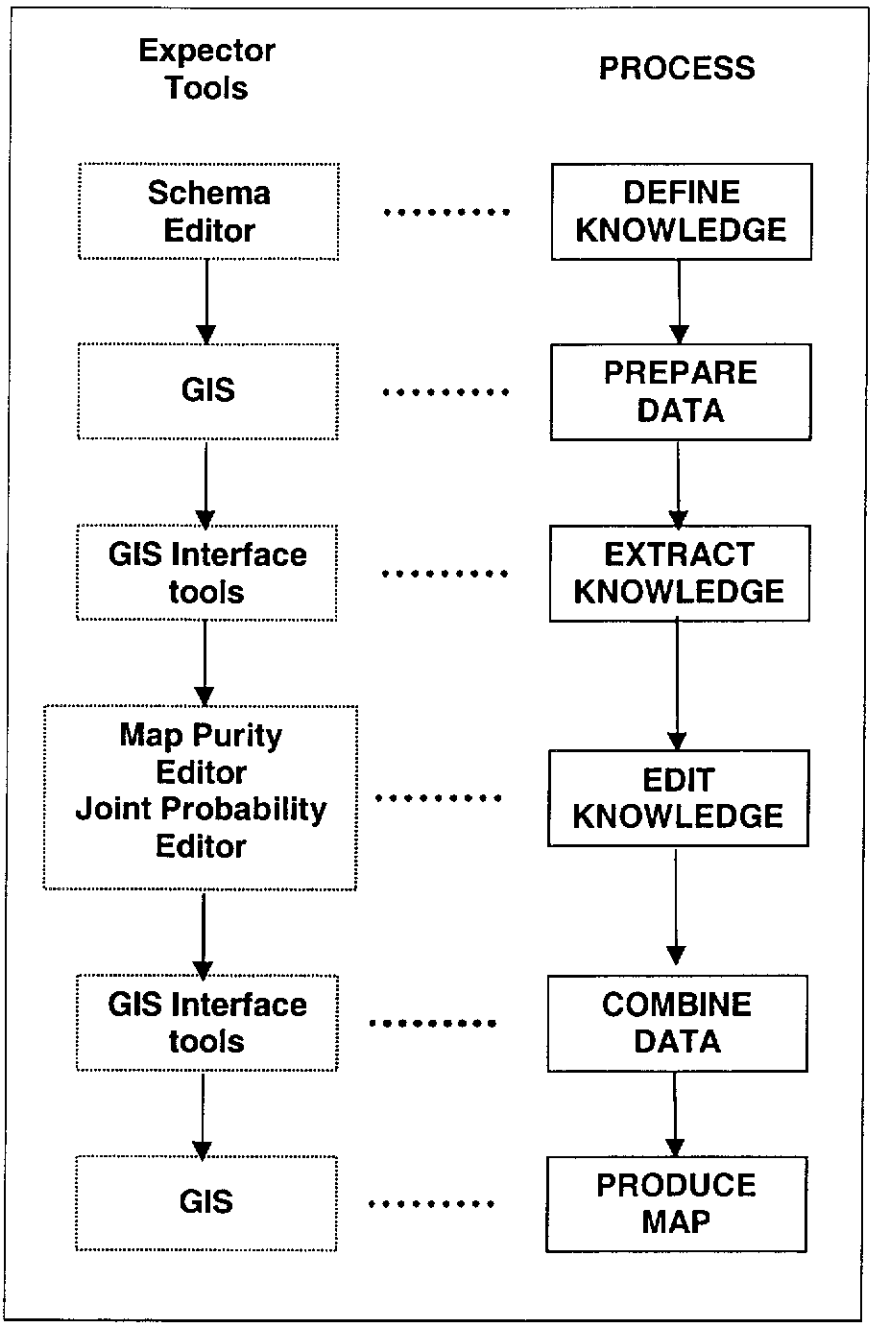

Figure 9.1 The Expector process

9.3 Knowledge definition

Knowledge definition is the process of defining the attribute to be mapped and assembling the evidence to be used in that mapping. The Expector software provides a schema editor that enables the problem to be set out in a graphical form. A separate schema should be created for each mapping project.

Before a schema is created, the attribute to be mapped (the hypothesis) should be defined and, where possible, fieldwork carried out to determine its likely probability distribution within the area of interest. Fieldwork and sampling are discussed in greater detail in Chapter 12. Similarly, evidence datasets relevant to the defined hypothesis need to be identified. Again, this may require some data collection.

The schema editor is the 'front page' for Expector and is displayed whenever the software is invoked. The software may be started either directly from MS Windows or indirectly through the GIS interface. Figure 9.2 shows a completed schema editor form.

The schema editor is divided into two parts; one dealing with the evidence and the other with the hypothesis, or attribute being mapped. Each contains a number of data entry fields, all of which must be completed. With the exception of the minimum number of hypothesis states, there are no default values. In addition to the interfacing provided in these two sections, some functions are supported by drop down menus.

9.3.1 The hypothesis section

The hypothesis section contains a number of boxes for user input, a 'spin button' control (a pair of increment - decrement arrows), and a bar graph display. The upper-most box is used to enter the name of the attribute being mapped. It is advisable that the name used be kept brief. It will be used as a basis for further file names, and some host GIS have a limit on file name length.

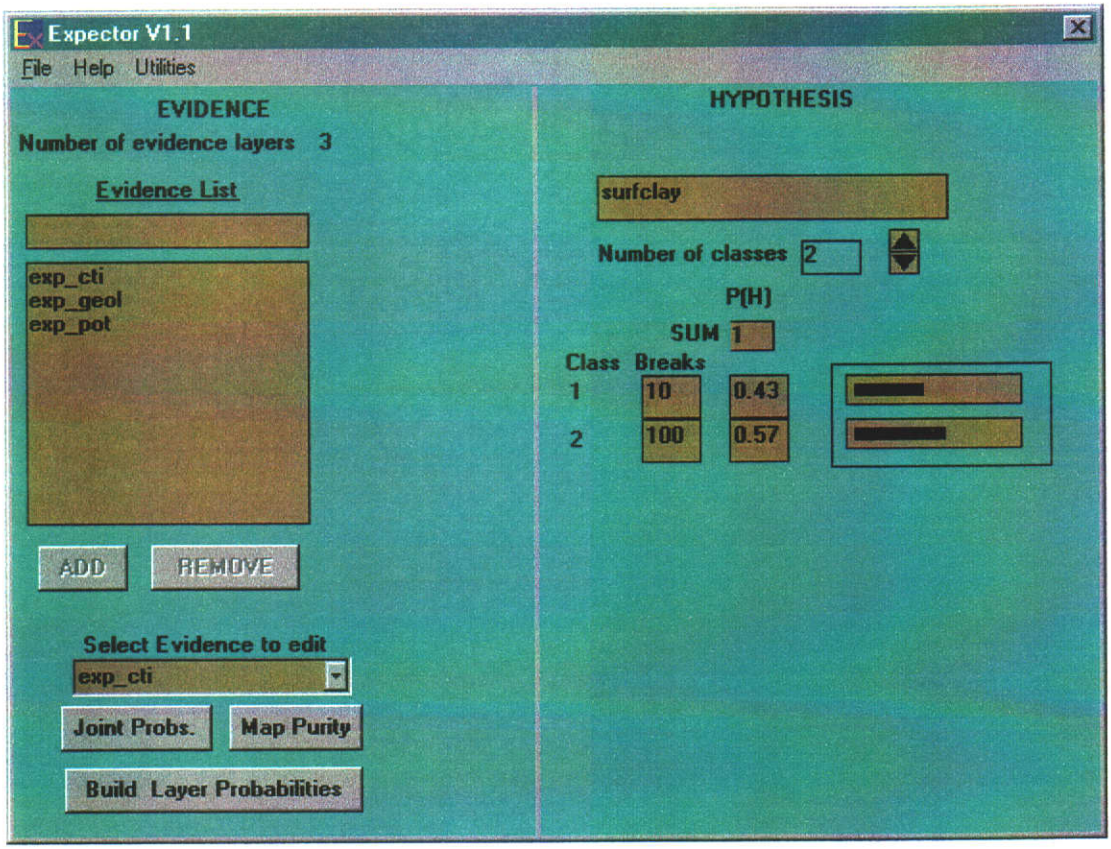

Figure 9.2 A completed schema

The 'spin control' is used to enter the number of hypothesis states. The minimum number of hypothesis states is two. As the number of states is incremented, the form will expand to accommodate them, up to a maximum of 18 states. This is, however, an unusually large number for most natural resource attributes and a more typical figure would be between four and six.

The remaining boxes are used to enter the upper bounds of each of the states and their prior probabilities. Prior probabilities will be determined either from sample data, using the method described in Section 8.3.1 or will be assigned on the basis of expert knowledge. It is a prerequisite that the sum of the prior probabilities be one. A bar graph display allows inspection of the distribution of these values and a sum check box is provided.

The hypothesis section of the form displayed in Figure 9.2 is showing information about a hypothesis named 'surfclay'. The hypothesis has two states, the upper bound values of 10 and 100 indicating ranges of 0-10 percent clay and 10-100 percent clay respectively. Prior probabilities of 0.43 and 0.57 have been assigned to these states.

9.3.2 Evidence section

The evidence section contains two main features. The first, occupying the upper part of the form, is a list box for the names of all evidence layers to be used. The second is a tool for selecting individual evidence layers for use in the knowledge editing process.

Evidence names are added to the list by typing their name into the data entry field above the box. Once a name is entered the **ADD** button becomes active. A mouse click on this button then places the name onto the list of evidence layers. To remove a name from the list, it must first be highlighted by dragging the cursor over it. This activates the **REMOVE** button. Clicking on that button deletes the name from the list. The names entered into these boxes should be the names of the raster datasets forming the evidence layers.

The selection box and the group of three command buttons located at the bottom of the evidence section are used to proceed to the knowledge editing stage. To ensure

correct operation of the knowledge editing algorithms, this process must be performed in a particular order. This is controlled by routines within the software which monitor the progress of the evidence layers through the editing process and selectively activate the command buttons. An evidence layer is selected for editing by highlighting its name in the selection box and selecting the appropriate action from the available command buttons. This procedure is discussed in greater detail in Section 9.6.

The evidence section of the form displayed in Figure 9.2 indicates that three evidence layers are being used in this analysis. These are named exp_cti, exp_geol and exp_pot. The exp_cti data layer is selected for knowledge editing

9.3.3 Drop down menus

Access to file saving and opening functions is provided through the **File** menu. Once the schema editor form has been filled in, it must be saved. As a fail-safe, the software will not allow entry to the knowledge editing process until the schema has been saved. A default extension of ***.exp*** is suggested for the schema file, which should be stored in the same directory as the evidence data layers and other files that Expector will use. It is important, for good data management, that a separate directory be used for each hypothesis attribute.

Expector generates a number of files that are used in the knowledge editing and transfer process. The **Utilities** menu provides access to a file management tool to assist in the maintenance of this knowledge base. This management tool is discussed in more detail in Section 9.10.

9.3.4 Editing an existing schema.

An existing schema may be loaded for editing by using the **Open** choice from the **File** menu. Care needs to be exercised whilst editing existing schemas since there is considerable co-dependency between many of the knowledge base files.

Evidence names may be added or subtracted without difficulty, but there are restrictions on the manipulation of the hypothesis information. The number of states of a hypothesis may not be changed. If such a change is necessary, a fresh schema

should be created. If it is intended that any of the evidence data be used again for the estimation of state values in a different hypothesis, then that schema should be created in a different directory. Evidence data sets pertaining to that schema should also be copied into that directory. This avoids the possibility of confusion between either differently categorised dataset representing the same basic evidence data, or between similarly named ancillary files containing different hypothesis - evidence joint probability data.

Within an existing schema, the hypothesis prior probability distribution may be changed. However, if prior probabilities are changed once the knowledge editing phase has started, there will be a miss-match between the values in some of the files in the knowledge base. If this situation occurs, an additional **Refresh Priors** button will appear on the schema editor form. Clicking on this button will initiate a routine to automatically adjust files, where possible, and to provide a warning of files that require further editing by the user.

9.4 Data preparation

Once a schema has been defined, evidence layer datasets will need to be prepared for knowledge extraction and combination. This data preparation takes place using a GIS and there is a degree of preparation required for both evidence and hypothesis data.

9.4.1 Preparation of evidence data layers

A general discussion of the data processing required at this stage is provided in Section 8.2 and will not be repeated here. It suffices to say that a categorised raster file must be prepared for each evidence layer and that all files should be co-registered in a common co-ordinate reference system. Grid datasets should be saved into the same directory or workspace as the Expector schema file and should be named to correspond exactly with the evidence layer names in the schema.

9.4.2 Preparation of hypothesis data

A site data file of observations of the hypothesis (such as that discussed in Section 8.3) should be prepared. This should be produced using a spreadsheet or text

editor as a comma delimited ASCII file and stored in the same directory or workspace as the schema file.

9.5 Knowledge extraction

Knowledge extraction is performed using interface routines specific to the host GIS. For the purposes of this discussion it is assumed that the GIS being used is ArcView, and that the reader is familiar with ArcView terminology. The Expector interface for ArcView is accessed through a series of custom buttons on the ArcView desktop (Figure 9.3). The leftmost two buttons in the interface are those associated with knowledge extraction.

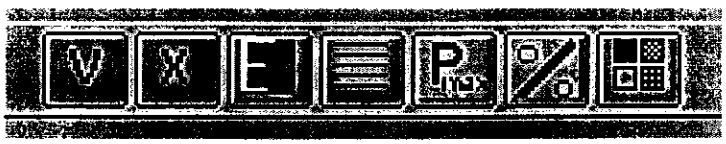

Figure 9.3 Control buttons for the ArcView interface

9.5.1 Determining evidence probability distributions

As described in Section 8.3.3, the prior probability distribution of the evidence layer is essentially a histogram of the cell values in the evidence dataset. Clicking on the **V** button in the interface starts a routine which directly accesses the database table associated with an evidence layer and writes the values to an ASCII file. This button is only active and useable if the active theme in the view is a grid dataset. A default filename comprising the active grid datasets name with the extension **.val** is used. This file is referred to as the Values file in subsequent discussion. Each evidence layer in turn should be made the active theme and this process carried out.

9.5.2 Determining joint probability distributions

In cases where a site sample data file exists, a routine is used which queries grid datasets at the locations of the sites and generates a cross-tabulation file. This routine is accessed through the **X** button in the interface. Again, this button is only active if the active theme is a grid dataset. The user is asked to select, using a file dialogue box, the appropriate site file for the layer under consideration. In general, there will only be one site file for a particular mapping project. A default filename,

comprising the active grid datasets name with the extension **.xtb** is used. This file is referred to as the Cross-tab file in subsequent discussion. Since this process also requires that each evidence layer in turn should be made the active theme, it can be carried out in tandem with the creation of the Values file.

At this stage, the user can use ArcView's Event Theme creation capabilities to display the point data over the evidence layers. This will give an appreciation of any bias which may be present in the sampling and which will have a bearing on the validity of the figures stored in the Cross-tab file.

9.6 Knowledge editing

Once a schema has been defined and all evidence data prepared, the user can proceed to build and edit the knowledge base. This involves building Map Purity and Joint Probability tables for each evidence layer. The Values files and Cross-tab files which have been created using the 'host' GIS may be used to seed this process.

As described in Section 7.5.2, Expector uses the Map Purity values to convert the relative abundance of the evidence map classes to the prior probability distribution for the corresponding field classes. It is, therefore, imperative that the Map Purity table for any evidence layer be completed before its Joint Probability table. Expector enforces this by denying access to the **Joint Probs.** button for each layer until the Map Purity table is complete.

9.6.1 The role of the knowledge base

It is implicit in the design of Expector that expert knowledge be used as much as possible to populate the knowledge base. This is due largely to the relatively small size of sample data typically available. However, in cases where a particularly rich dataset exists, a knowledge base derived from it may be used with little alteration. Such a choice is left to the expert judgement of the individual user. In either case, the operation of Expector has the effect of using the spatially dense evidence layers to extend the relationships contained in the knowledge base across the landscape being surveyed.

9.6.2 Building the Map Purity table.

The Map Purity table building and editing routine for a particular evidence layer is initiated by highlighting the name of that evidence layer in the selection box and clicking on the **Map Purity** button (Figure 9.2). If no Map Purity file yet exists for this evidence layer, the system will build one with a default set of purity values. Choices are offered regarding the way class labels (stored in an associated file) and prior probabilities are assigned. At this stage, the user is presented with a form containing three command buttons (Figure 9.4).

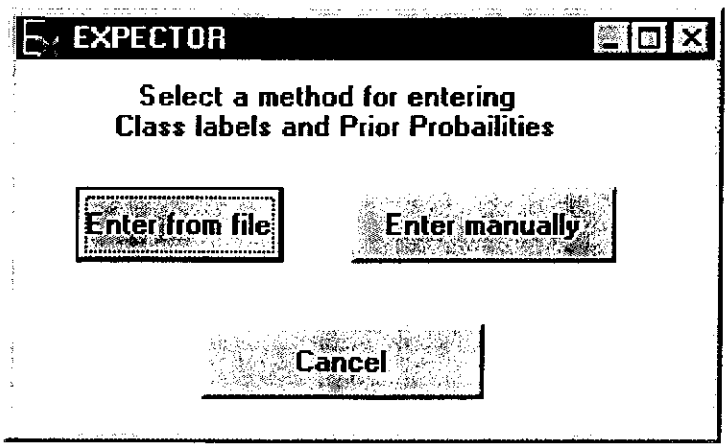

Figure 9.4 Choice box

The usual method of entering this information will be from the Values file created during the data preparation phase. Choosing this option will lead to a file opening dialogue, within which the Values file appropriate to the evidence layer is offered as the default choice. If, as is likely, this file has been created by one of the supplied GIS interfaces, the individual states or classes of the evidence will have numerical identifiers. Since it is preferable to work with descriptive labels in the knowledge editing tools, a further dialogue offers the choice of entering text names for each class. Whether or not this option is exercised, the states continue to be represented internally by the numeric class labels derived from the GIS.

In the unlikely event of the software being used with a GIS for which Values files cannot be created, the user can choose to enter probabilities and labels manually. The software provide a series of prompt boxes which request firstly the number of classes in the evidence layer and then, for each class, a label and the probability of

occurrence for that class. Classes will be given consecutive internal numbers, although the user may enter non-numeric labels. The probabilities, which should be derived from a cell count or histogram of the evidence layer, should sum to one.

Once the Prior Probability values and labels have been entered, the Map Purity Editor form will be displayed. If a Map Purity table already exists for the relevant evidence layer, the software proceeds directly to this point.

9.6.3 The Map Purity Editor

The Map Purity Editor is a graphical form that allows manipulation of the Map Purity table for the selected evidence layer. To recapitulate, this table holds the distributions that describe the conditional probability that each map class actually represents that class in the field. This matter was covered in detail in Section 7.5.1.

The form contains two main functional areas. (Figure 9.5) On the right is a grid which displays Map Purity values as a matrix whose size is determined by the number of classes in the input evidence layer. When the form initially appears, it contains figures which assume that the map is 100 percent accurate in all classes. This is reflected by the values of 1 on the diagonal and zeroes elsewhere. The figures in this grid cannot be edited directly.

Each column in the grid represents the probability distribution of a map class over the real field classes. The left-hand side of the form contains tools to edit individual columns. The figures in the line above the white portion of the grid (labelled FREQ) are the relative abundance of the various map classes. Rows and columns are labelled with alphanumeric labels, if the user has supplied them, or else with class numbers.

When the form is first displayed, the numerically first class is highlighted to indicate that it is the class being edited, and its values are displayed in the boxes on the left of the form. Any other class may be selected for editing by clicking on the relevant column in the grid. Values in the editable area may be entered either by typing in the box or by using the spin buttons. If the 'spin buttons' are used, the size of the spin

increment may be set to either 0.1, 0.05 or 0.01 by choosing the appropriate option box.

As changes are made to the class purity values in these editing boxes, the probability values are changed in the grid and new totals calculated. The conditional probability distribution down each column must sum to one. Figure 9.5 shows a form for which the editing is complete. The evidence layer shown here is named exp_cti and has five states or classes. The first state ('upland') is selected for editing.

Once the user is happy with the numbers in the matrix, a hard copy of the form may be generated as a screen dump using the **Print** button. Alternatively, if the figures are required in text form for entry into a report, they may be saved using the **Print to File** option from the **File** Menu.

When editing or viewing the probabilities for a particular evidence layer is complete, clicking on the **Done** button will offer an option to save the file and return the user to the main screen. The file may also be saved using options from the **File** menu. A default file name, based on the evidence layer name and the extension **.pee**, is assigned. This name should be used to ensure correct operation of the software. After saving the file, the user is returned to the main Schema Editor form. (Figure 9.2).

9.6.4 Building the Joint Probability tables.

Once the Map Purity table for an evidence layer has been created, the user can create a Joint Probability table. This process uses a similar form to the Map Purity Editor and is accessed using the **Joint Probs.** button on the main form.

The first time a Joint Probability table for an evidence layer is accessed, the system will build a new one. There is a choice of methods by which the initial values in that table are assigned. A dialogue box (Figure 9.6) appears offering the option of either entering values from a file or allowing the system to assign some estimated values.

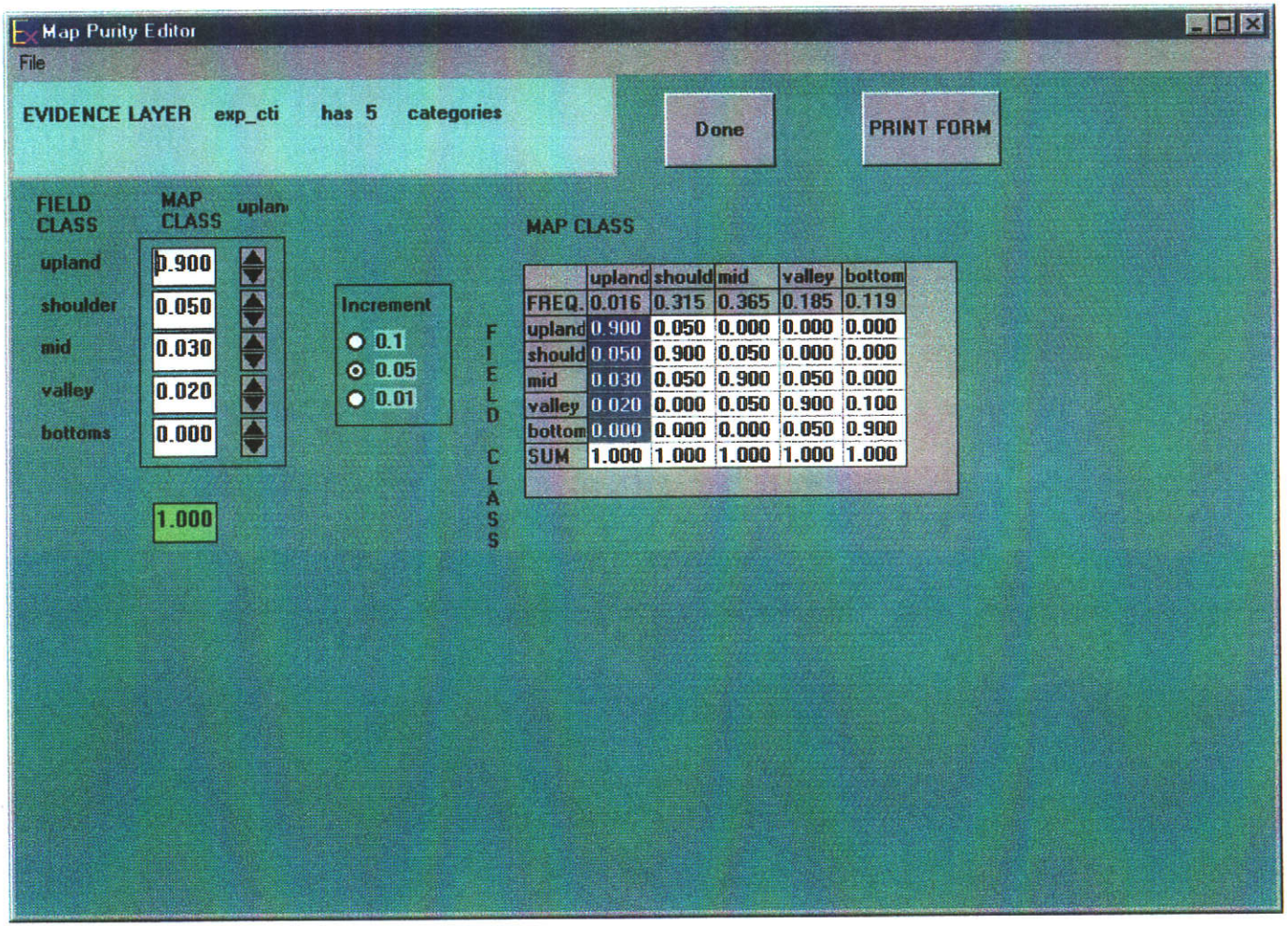

Figure 9.5 The Map Purity Editor

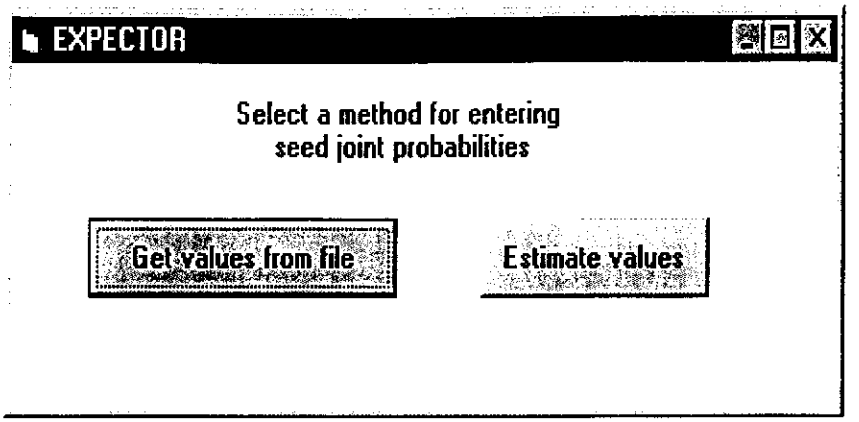

Figure 9.6 Dialogue box for seed joint probabilities

In most cases, the user will choose to enter values from a file. The file used will be the Cross-tab file created by the GIS interface, as described in Section 9.4. The file is selected using a typical Windows file dialogue box. If the user chooses to estimate values, the Joint Probability table will be populated by values calculated from the prior probabilities of the field classes and of the hypothesis states. Whichever method is chosen, the values in the table should be regarded as seed values only. They will require editing.

If values from a Cross-tab file are used, two points need to be recognised. Firstly, these initial values refer to the co-occurrence of evidence *map* classes and hypothesis states. Secondly, these figures have been derived from a sample dataset which may be biased. The Joint Probability table is intended to be a representation of the users expert knowledge about the co-occurrence of *field* classes in the evidence with the hypothesis states throughout the landscape, not just at the sample sites. For that reason, the initial values should be regarded as being, at best, a guide to relative magnitudes.

Using estimated values derived from the prior probabilities of the evidence classes and the hypothesis states in an analysis would render that particular evidence layer powerless. This is due to the fact that the effect of any one grid cell in an evidence layer on changes in the probability distribution of the hypothesis states depends on the ratio of its class conditional probability with the hypothesis prior probability. (c.f. Equation 8.20). If the estimated values are used this ratio will become unity,

causing the evidence layer to behave as if no data were present. However, in a reconnaissance situation and in the total absence of any site data, the estimated values at least provide some starting figures.

9.6.5 Editing the Joint Probability table

Once the seed joint probability information has been entered or if a joint probability file already exists, the user is presented with the Joint Probability Editor form (Figure 9.7). This form is generally similar to the Map Purity Editor. It presents the joint probability table in the form of a grid with row and column sums. As with the Map Purity form, the grid is not directly editable and an interface is provided to it through a series of edit boxes. Each column in the grid represents the probability distribution of one field class of the evidence layer across the several states of the hypothesis and should sum to the prior probability of that field class. The joint probabilities are converted to percentage values for display in the edit boxes.

When the form is first displayed, the class with the lowest category value is highlighted as being editable. To select any other class, the cursor is moved to the relevant column in the grid, and the mouse is clicked. Values in the edit boxes may be altered either by typing numbers in directly or by using the spin buttons. If using the spin buttons, the increment may be set to either 10 percent or 1 percent by choosing the appropriate option box. As changes are made to the percentage values in these editing boxes, the probability values are changed in the grid and new totals calculated. Column totals should all be 100 percent when editing is complete.

It is a requirement of the Bayesian network theory which underlies this software that the rows in this table should sum to the prior probability distribution of the hypothesis attribute and the columns should sum to the distribution of the evidence variable. The degree by which the values differ from the simple product of these distributions is a measure of the power of the individual evidence layer. The determination of the prior distribution of the hypothesis is therefore critical.

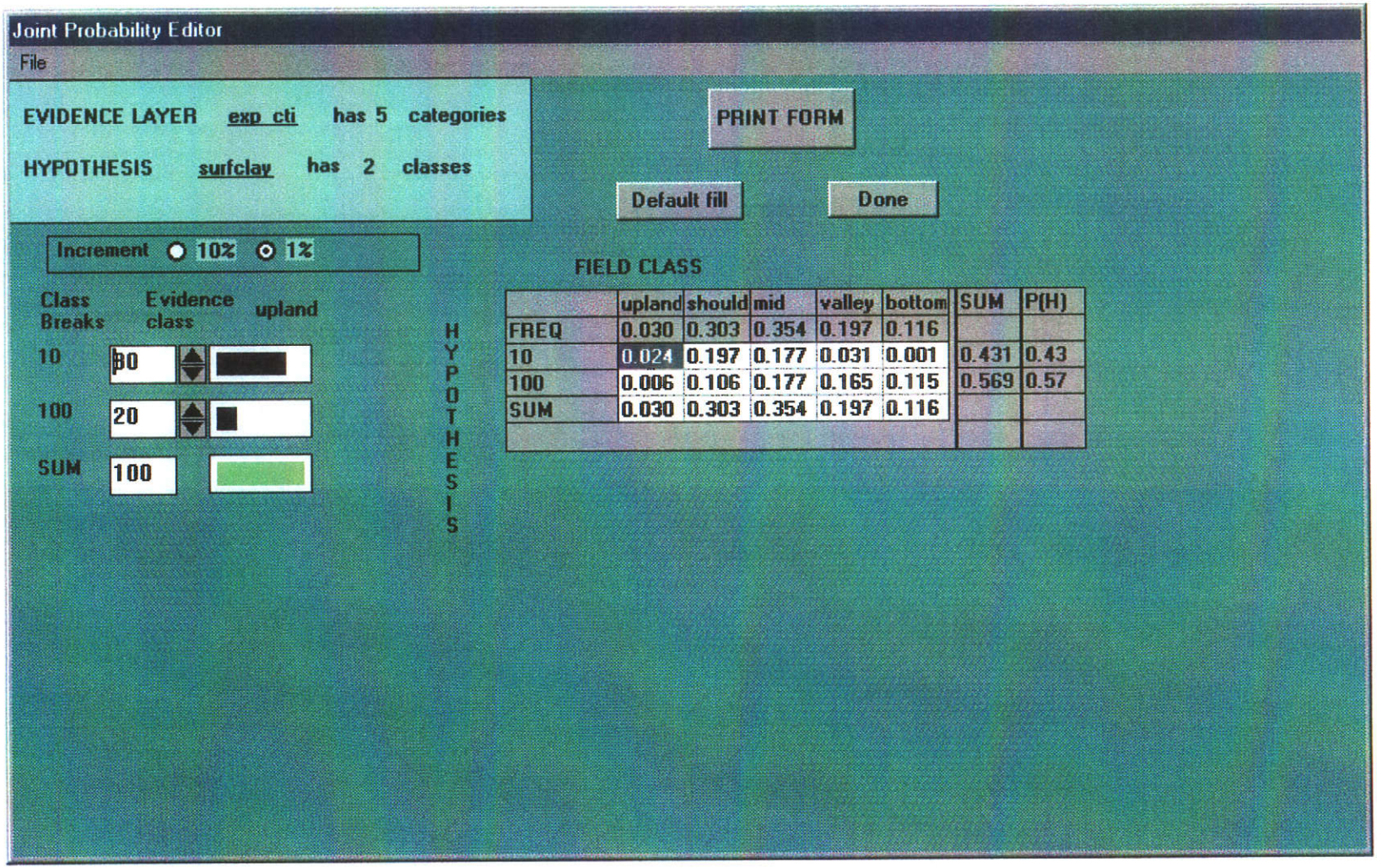

Figure 9.7 The Joint Probability Editor Form

If an evidence layer has only a few classes for which a strong defining relationship with the hypothesis attribute exists, there may be a further class which contains all other possible values of the evidence layer. An example of this is an evidence layer such as slope for which a relationship with the hypothesis is constant above a particular value. For cases such as these a **Default fill** button is provided. Clicking on this button performs the arithmetic of filling the selected column in the grid with values which cause the sum along rows to tally with the hypothesis prior distribution.

Once the user is satisfied that the table is completed, a screen dump of the values selected may be created using the **Print** button. Alternatively, if the figures are required in text form for entry into a report, they may be saved using the **Print to File** option from the **File** Menu.

In order to leave the Joint Probability form, the user should click on the **Done** button. A file saving dialogue is initiated, and a default file name based on the evidence layer name, with the extension **.phe** is assigned. The file may also be saved using the pull down **File** menu. After saving the file, the user is returned to the main schema editor form.

9.7 Building the probabilities for each evidence layer

Once the Map Purity and Joint Probabilities files for an evidence layer are created, the user must create a file to pass these probabilities back to the GIS. This file contains the posterior probability distributions for the hypothesis class across the mapped evidence classes. There is one such file per evidence layer. The mathematics behind this process are explained in Section 7.6 and illustrated by Figure 7.5.

Generation of the file is initiated simply by clicking on the **Build Layer Probabilities** button on the main form (Figure 9.2). The button will not be active unless both Map Purity and Joint Probability files exist for the evidence layer highlighted in the choice box. The user is responsible for ensuring currency of this file after any editing changes to either of the input files. The layer probability file is named by default using the evidence layer name and an extension of **.phq**. It is stored in the same directory or workspace as the schema file.

9.8 Data combination

Once all knowledge editing has been carried out and all Layer Probability files created, the task of data combination can be performed. This process executes the algorithm described in Section 8.5.2. and is carried out in the GIS. Using the ArcView interface, access to this routine is provided by clicking the fourth button from the left in Figure 9.3. For this button to be active, it is necessary that one of the evidence layers is an active theme in the view. A file dialogue box appears from which the user can choose the schema file that will drive the data combination process.

The process proceeds in the background following the steps of table linking, multiplication and normalisation described in Section 8.5.2. On completion a number of new grid datasets will have been created, one for each state of the hypothesis. These are floating point datasets representing probability surfaces whose values lie between zero and one. For each of them, a new view will be automatically created in the ArcView project. In those views the appropriate probability surface is displayed. The values are scaled into the range of zero to one hundred and a suitable colour table attached.

9.9 Additional features of the GIS interface

Expector's GIS interfaces provide additional capability for displaying probability grids and the ability to launch the Expector standalone routines from the ArcView desktop. Figure 9.8 again shows the buttons used.

Figure 9.8
Custom buttons in the Expector interface to ArcView

The **V** and **X** buttons are used to initiate the creation of Values files and Cross tab Files respectively. The button with the **Ex** symbol is used to launch the Expector software. The three rightmost buttons on the bar are used to access the display utilities.

The **Pmax** button is used to create a map of the most probable state of the hypothesis attribute using the algorithm described in Section 8.6.2. Clicking this button presents the user with a file dialogue box from which to chose the schema file pertaining to the required hypothesis attribute. On completion of the calculation the resulting grid dataset is displayed as a theme in the active view. The theme is given the name INDEX and is displayed by default with a colour table having six entries. If there are more than six states, it is the responsibility of the user to amend the colour table accordingly. It is also the responsibility of the user to save the dataset, if so desired, with a suitable filename.

The **%** button allows the display of any probability grids with a range of zero to one. These may be the result of previous runs of the software which the user wishes to display as a comparison with those generated by the most recent run. The grids are scaled into the range zero to one hundred and the appropriate colour table is attached.

The rightmost button, with the 'four-square' icon, enables simultaneous display, in several views, of an individual evidence layer and its associated layer probabilities. When clicked it activates a file dialogue from which the user should chose the layer probability file (extension .**phq**) for the evidence layer of interest. New views are then automatically created in the ArcView project, one for each state in the hypothesis attribute and one for the raw evidence layer states.

In each view that relates to a hypothesis state, a surface is displayed which shows the spatial distribution of the probability of that state, based solely on that one evidence layer. The view containing the evidence layer values is provided for reference. This display utility is particularly useful when examining unexpected results after data combination.

9.10 File utilities

Expector produces a number of knowledge base files associated with each evidence layer. It is inevitable that mistakes will be made during the course of an analysis session or that changes of opinions will occur. This results in the need to remove some of the knowledge base files.

To assist with maintenance of Map Purity and Joint Probabilities files, Expector provides a file utility tool. This is accessed by selecting **Manage Files** under the **Utilities** menu in the Expector schema builder and is only available if a schema is loaded. This gives access to a graphical form such as that shown in Figure 9.9.

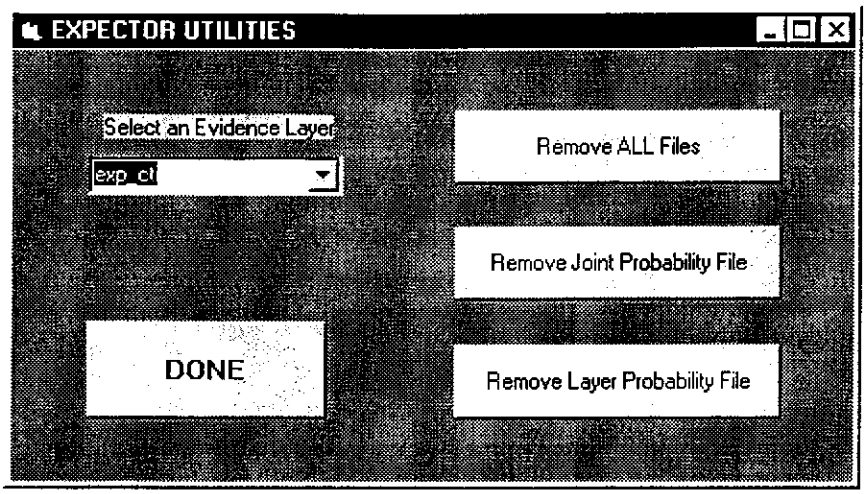

Figure 9.9 The Expector file utilities tool

The selection box contains a list of the evidence layers in the loaded schema. Once an evidence layer is selected, clicking on any of the three buttons will perform the required deletions. Due to the sequential nature of operation in Expector (resulting from the application of Map Purity values referred to in Section 9.6) if a Map Purity file is deleted then its associated Label file and the Joint Probability file for that evidence layer must also be deleted. For that reason, there is no option available to remove only the Map Purity file.

9.11 Summary

The Expector software, comprising a standalone program with interfaces to GIS, provides an integrated suite of tools to implement the quantitative natural resource assessment method described in preceding chapters. The software provides user friendly forms-type tools for creating and editing a knowledge base. Tools are provided for performing a combination of the knowledge associated with individual data layers. In addition, GIS specific display utilities for examining and presenting the resulting map data layers are available, as are file management routines.

Although knowledge should be provided by an expert user, facilities are provided to seed the knowledge base with values derived from site sample data. If that site sample data is relatively rich, then Expector allows its use with little modification.

The next two chapters describe some of the applications to which Expector has been put during its development.

Chapter 10

DEVELOPMENTAL APPLICATIONS OF THE EXPECTOR METHOD

During the development of Expector, a number of test data sets were processed using the various algorithms and their encoded routines. This was both a test of the various interfaces and to ensure that the method was useable by land resource assessment professionals. Expector is a knowledge-based method, and the accuracy of any result will depend on the quality of the knowledge that forms part of the process. This testing was not designed to specifically test the accuracy of maps produced using the Expector method. However, some testing of output was performed to ensure that the results were at least consonant with the thinking of those professionals who provided the knowledge base.

10.1 Choice of test data sites

The choice of test data sites was governed largely by the availability of suitable data sets and knowledge bases. It was envisaged that the principal use of Expector would be in the mapping of individual soil properties rather than entities such as soil types. A similar approach was taken by Moore et al. (1993), who used linear models to relate terrain attributes to soil properties. Their trial site was at Sterling, Colorado, in the United States of America. It was decided to use, with permission, their data as a test set during development. Subsequently, a dataset for the East Yornaning catchment in Western Australia was acquired and used. Both studies are reported here.

10.2 Sterling, Colorado - inputs

10.2.1 Location and objectives

The Sterling site is located in north-eastern Colorado and is a long term study site for crop management in dryland agriculture (Moore et al., 1993). Water is a major determinant of growth and production; a characteristic which it shares with the majority of the Western Australian wheatbelt. Data from the site have been used in an example of the linear modelling of relationships between terrain attributes and soil physical properties (Moore et al., 1993). The Sterling data was also used in an

application of the PROSPECTOR method to the mapping of soil physical attributes (Cook et al., 1996).

10.2.2 Evidence datasets

The site is relatively small, covering only 5.4 hectares sloping from south to north with an overall elevation change of about 6 metres. The original survey team had collected a rich data set for this small area comprising 231 sites on a regular 15.24m (50 ft) grid At those locations, elevation, A horizon thickness, extractable Phosphorous, organic matter content (OM), sand, silt, and clay content were measured. Compound topographic index (Moore et al., 1991) had been calculated from the elevations. The data were made available as a spreadsheet listing of sample site co-ordinates and associated attributes.

An elevation surface was reconstructed from the point data using spline interpolation. Similarly, a surface was generated from the wetness index values of the original workers. Slope and aspect surfaces were calculated from the elevation surface using the appropriate functions in ARC/INFO. The measured OM values were converted to a surface, again using a spline interpolator.

Analysis of the original data indicated that there was a difference between soils having OM concentrations greater and less than 1.6 percent. A schema was constructed with a two class hypothesis attribute being the presence or absence of OM in concentrations of > 1.6 percent. Evidence variables selected were wetness index, slope angle and aspect.

Slope was classified into nine classes at 0.5 percent intervals from 0 to 4.5 percent; aspect was classified into the eight cardinal directions; and wetness index into seven equal classes between its maximum and minimum values. The schema is shown graphically in Figure 10.1.

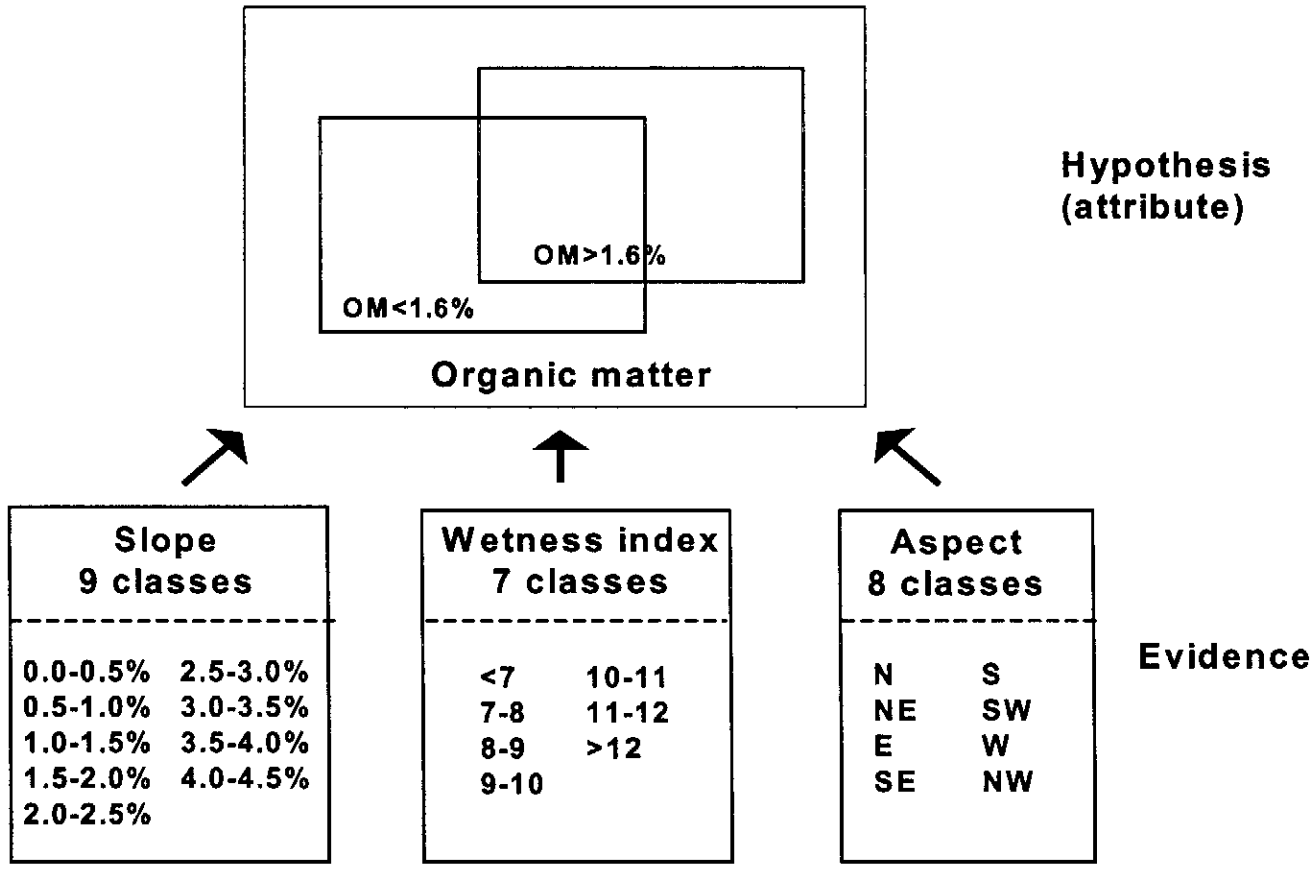

Figure 10.1 Schema for study at Sterling, Colorado

10.2.3 Knowledge base

A subset of 75 points was randomly extracted from the point data set. The prior probabilities of the hypothesis attribute (OM> 1.6 percent and its converse) were determined by reference to the measured values of OM at these points. These probabilities were 0.69 for Class 1 (OM<1.6 percent) and 0.31 for Class 2 (OM>1.6 percent).

The 75 points were classified according to their measured OM content and used to determine relationships between the OM classes and classes of the evidence variables. These relationships, expressed as Cross-tab files (c.f. Section 9.52), were used to seed the joint probability tables. In the absence of any local expert knowledge about the site, the values were only edited to ensure consistency with the prior probabilities - essentially a scaling operation to compensate for over and under-representation in the sampling.

Since this was essentially a proof of concept exercise, map purities were assumed to be 100 percent for all maps.

10.3 Sterling Colorado results

The Expector software and its ArcView interface were used to combine the individual evidence layer probability estimates into two surfaces representing the probability of occurrence of each of the OM classes. From these, a map of most probable OM class was developed. Figure 10.2 shows the two OM class probability maps draped over the digital elevation model.

The 153 sites in the original point data set which had not been used in the knowledge base were then classified into the OM classes and comparisons made at those points with the most probable OM class, as predicted using Expector. A comparison of those values is shown in Table 10.1.

The figures in Table 10.1 show that out of the 153 sites, organic matter class was correctly predicted at 127 sites, an overall accuracy of 83 percent. For individual classes 94 percent of Class 1 was correctly predicted whilst only 52 percent of Class 2 were correct.

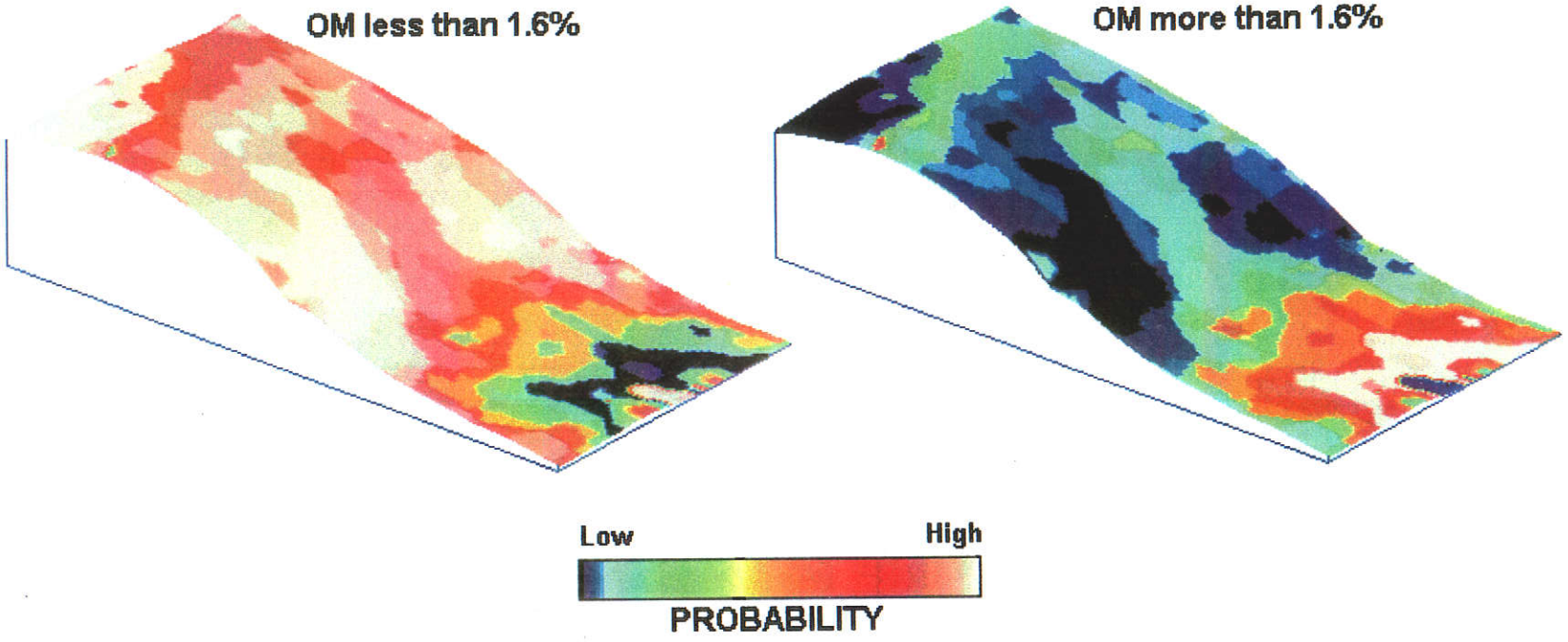

Figure 10.2 Organic matter class probabilities draped over a digital elevation model. Sterling Colorado

Field measurement	Number of sites	Correctly allocated	Percentage correct
Class1	111	105	94.5
Class2	42	22	52.3
Overall	153	127	83.0

Table 10.1 Observed and predicted OM classes at Sterling site.

Since the results for Class 2 (>1.6 percent OM) are worse than those for Class 1, that class was investigated. The probability of membership of that class at each of the test sample points was extracted from the Expector output map. Figure 10.3 shows a plot of those probabilities against the actual organic matter content. From this it can be seen that there is a general positive trend to the data. That is, the higher the probability of finding organic matter greater than 1.6 percent, the higher the actual OM value found in the field. There is, however, considerable overlap across the critical threshold of 1.6 percent OM.

The question of differences in accuracy between classes is further discussed in Section 10.7.

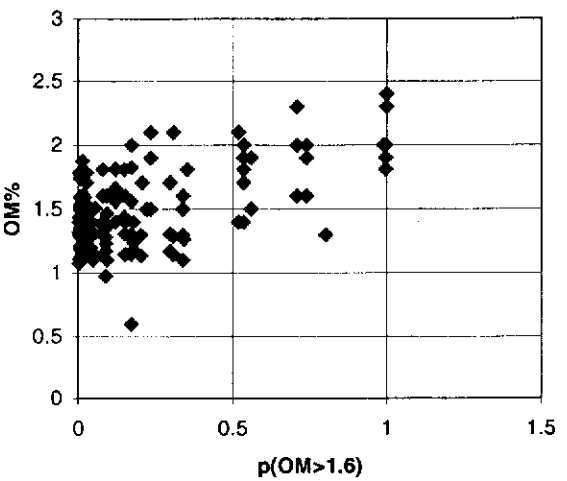

Figure 10.3 Plot of probability of membership of class OM>1.6 versus OM content.

10.4 East Yornaning, Western Australia

10.4.1 Site location and objective

The East Yornaning catchment is located in the south-west of Western Australia (Figure 10.4). The catchment has an area of approximately 200km^2 and is representative of the dissected lateritic landscape which occurs extensively within the region (Mulcahy, 1973).

A summary of representative soils of the region is given by McArthur (1991). Soil materials include highly weathered residual material, ferricrete, re-worked sediments, and freshly weathered regolith from granitic or doleritic outcrop. Soils over the area have been mapped (McArthur et al., 1977) predominantly as kurosols (Isbell, 1996) with tenosols on upper slopes and sodosols in the valley floors. The catchment has also been used to test the ability of airborne gamma radiometry to map soil types (Cook et al., 1996).

The objective of the Yornaning exercise was to test the knowledge based system by producing a map of selected soil attributes. Readily available data layers, mainly terrain attributes supported by the airborne gamma radiometry, were chosen as inputs. The attributes chosen for consideration were soil surface texture, depth to impermeable horizon, and gravel content in the top 50cm. The production of a map of soil surface texture, expressed as percentage clay content, is reported here. Available datasets for the catchment include a digital elevation model, air-photo interpretation of rock outcrops, and 1:250,000 regional geological mapping.

10.4.2 Terrain attributes

A digital elevation model of the catchment with a horizontal resolution of 25m was available. It had been generated from 1:25,000 colour photography using a stereoplotter. A number of derived attributes were generated from this data set using the standard tools in the ARC/INFO GRID module and some additional processing. These included catchment position, surface curvature, slope and position in catena.

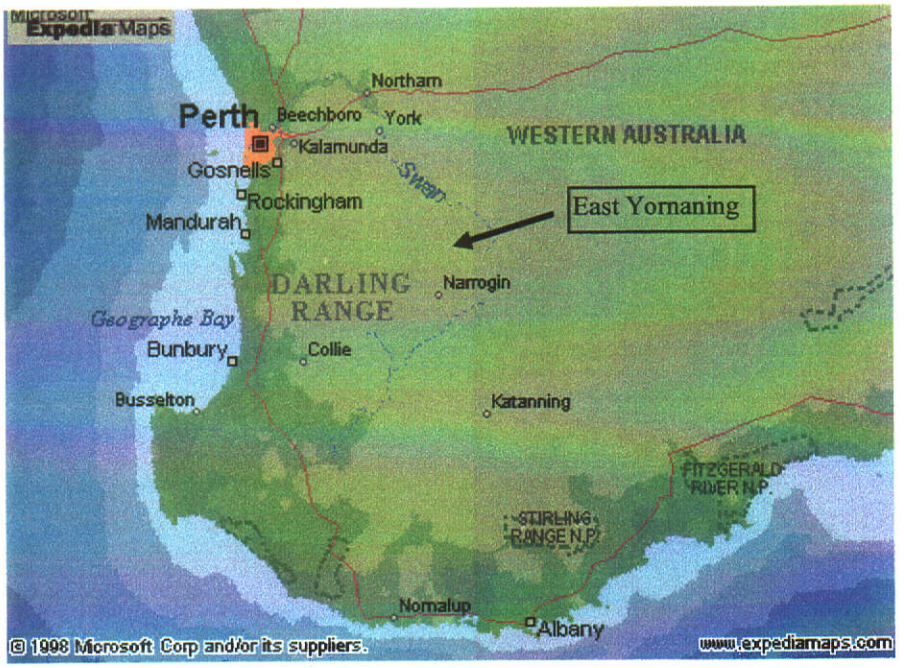

Figure 10.4 Location of East Yornaning, Western Australia

The catchment position dataset was created by dividing the elevation model surface into a series of sub-catchments using the FLOWDIECTION and BASIN routines in ARC/INFO. The surface hydrological network, obtained as a digital product from Agriculture WA, was examined and manually coded with stream order according to the Strahler (1952) classification. Each sub-catchment was then assigned a 'catchment order' equivalent to the order of stream running through it. In the East Yornaning catchment the main creek is a fifth order stream. At the outflow of the catchment this creek joins the Hotham River, a sixth order stream.

The surface curvature data set was generated initially using the CURVATURE command in ARC/INFO GRID. The results of this process are two data layers; one representing plan curvature, the other profile curvature. From these, a curvature class layer was synthesised. Classes were generated by taking all four possible combinations of positive and negative plan and profile curvatures.

Similarly, slope was calculated using the standard ARC/INFO algorithm and classified into nine classes. Classes represented an increment of one percent of slope, up to eight percent, with the last class containing all terrain with a slope in excess of eight percent.

Position in catena, termed for the purpose of this analysis as 'stream/ridge ratio', was calculated from the topographic surface and from line coverages describing the positions of the stream and ridge lines. Positional information about the stream lines was obtained from the surface hydrology coverage described above. Ridge lines were digitised from stereo air photos and checked by on-screen comparison with contours derived from the elevation model.

The PATHDISTANCE function in ARC/INFO was then used to create raster representations of distance to nearest relevant stream and nearest ridge. When distances to streams were assigned, the ridges were declared as barriers. That is, cells were assigned a distance to the stream *hydrologically* closest to them rather, than closest in simple surface distance terms. Similarly, the streams were used as barriers in the process that assigned distance to ridge. Once these two datasets had been created, a third, the ratio between them, was calculated. This was then divided

into three classes representing areas near ridges, areas near streams and mid-slope areas.

10.4.3 Other datasets

Other data set used in this work were geology, classified airborne gamma-radiometrics and distance from rock outcrops.

The geological data set was digitised from the compilation sheets for 1:250,000 Geological Survey of Western Australia map of Sheet SI50-3 Corrigin (Chin, 1986) There are only seven lithological types in the catchment, although considerably more occur in the area covered by the 1:250,000 map.

The classified airborne gamma-radiometric evidence layer was taken from a dataset flown by World Geoscience Corporation which had been used to assess the utility of such data for soil mapping (Cook et al., 1996). A multi-spectral classification of the data into landscape types resulted in an evidence layer with four classes. The classes represented in this data are granitic, sandplain, colluvial, and alluvial.

10.4.4 Development of schema for surface texture (clay content)

In consultation with a soil surveyor, an initial schema was developed based on ideas borrowed from the PROSPECTOR method. This called for a number of intermediate data layers such as 'favourable landscape position'. Such concepts are difficult to measure objectively. Since one of the overall aims of the development of the Expector method was that it should be simple for a field surveyor to use, a simpler construct was called for.

Discussions with soil surveyors suggested that 'soft' combined landscape parameters such as 'favourable landscape position' have a relationship to soil properties which contains interactions between various components of that parameter. In a probabilistic network, these interactions find expression in the process of data combination. It was, therefore, decided to create a schema that described a direct relationship between each of the seven evidence layers and the hypothesis. The schema is shown graphically in Figure 10.5.

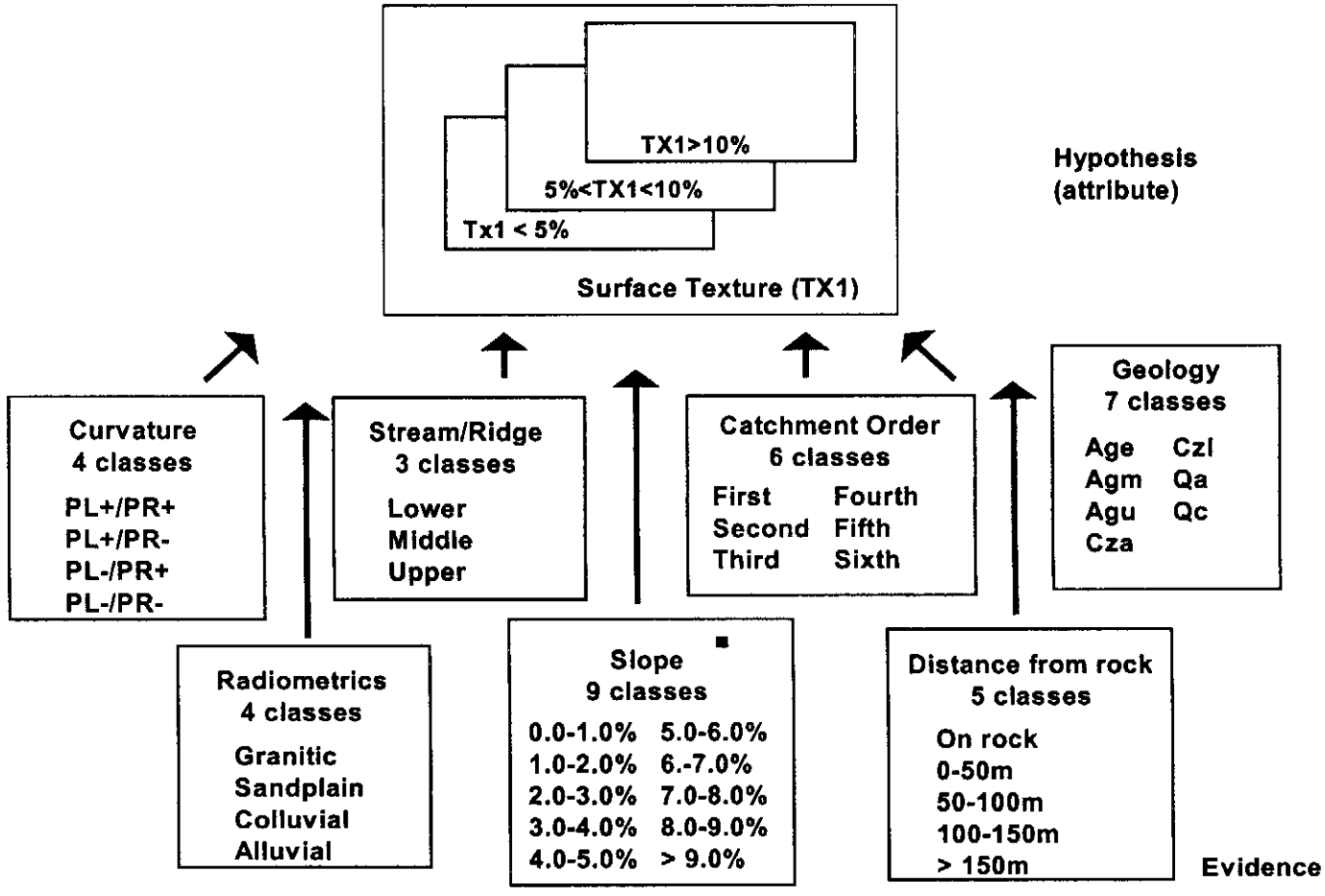

Figure 10.5 Schema for East Yornaning

The knowledge base for an Expector schema has four components. In this case, they are the prior probabilities of the individual clay classes and the various evidence classes, the map purity values, and the joint probabilities between evidence layer classes and individual clay classes. The prior probabilities were supplied from the available data, whilst the map purity and joint probability values were derived by expert consultation.

10.4.5 Assigning prior probabilities

In consultation with soil surveyors, it was initially decided to use four classes of surface texture as shown in the left-hand section of Table 10.2. The prior probabilities of occurrence of these were taken from an analysis of 189 site observations. On inspection of this dataset, it was found that only four of the sites fell into the range 20 to 100 percent clay. Since, in practical agronomic terms, the two lower clay content classes are the most important this last class was amalgamated with the next lowest one. Table 10.2 shows the classes and prior probabilities used in the analysis.

	Initial classes		Final classes	
	Range (percent clay)	Prior probability	Range (percent clay)	Prior probability
Class 1	0 - 5	0.14	0 - 5	0.14
Class 2	5 - 10	0.51	5 - 10	0.51
Class 3	10 - 20	0.33	10 - 100	0.35
Class 4	20 - 100	0.02		

Table 10.2 Classes and Prior Probabilities for East Yornaining

Prior probabilities for individual data layer classes were determined using the methods previously outlined from the data layers themselves.

10.4.6 Assigning map purities

The logic involved in setting map purities requires a consideration of both the provenance of the data and of its nature. The following briefly describes the process, for each of the data layers. The individual map purity values are listed in data panels 10.1 to 10.7 located in Appendix A.

The catchment order dataset was regarded as being reasonably accurate, although some 'slippage' across class boundaries was expected due to positional errors. Misclassification errors of more than one class were not expected.

Similarly, the curvature data layer, being derived from the same elevation model, was regarded as reasonably accurate. However, since curvature calculations take the second derivative of the surface, they are sensitive to slight errors in the elevation model. The curvature layer was, therefore, assessed as having an individual class purity of 70 percent, with the error evenly distributed across the other classes.

The geology data layer, being taken from a map published at a scale of 1:250,000, will have some inherent positional problems which need to be considered at the same time as likely misclassifications. The explanatory notes accompanying the paper copy of the map (Chin, 1986) proved invaluable here.

The granitic *Age* and *Agv* map units occur as a complex whilst *Agm* is lithologicaly similar. Therefore a reasonable degree of confusion is to be expected amongst them. The individual class purities for these three units were therefor set at 70 percent with the relative misclassification reflecting the degree of complexing.

The two alluvial units (*Cza,Qa*) are described by Chin as being uncertain in places and this, together with the potential positional uncertainty has resulted in each unit being assigned a purity of only 65 percent. Similarly, the possibility for confusion between *Qa* and the recent colluvium (*Qc*) as well as possibilities for confusion between *Qc* and the lateritic unit *Czl* are reflected in the purities assigned to those units. Due to its occurrence over areas of relict granite, the lateritic unit is allowed a degree of confusion with the granitic units

No formal assessment of the classification accuracy of the radiometric data was available, so classes were assigned an accuracy of 80 percent on the basis of expert opinion. Misclassification error was uniformly distributed across all classes.

The rock data layer was assumed for practical purposes to be accurate. The actual positions of the rock outcrops were well mapped from ortho-photos and the buffering error is expected to be negligible.

Being based on the elevation model, the slope map was regarded as being of high accuracy. The values are, in fact, local averages based on a three by three cell neighbourhood rather than point calculation. Although such datasets inevitably contain an error component (Dunn and Hickey, 1998), it was not quantified for these data. Each class was, therefore, assumed to be 90 percent pure with the error distributed symmetrically. The exceptions to this are at the ends of the distribution where error is all assigned to the adjacent class.

Using similar logic a value of 90 percent was assigned to the stream/ridge distance ratio map. This relatively high accuracy was based on its origins in the elevation model and the extreme un-likelihood, in a three class system, of mis-classifying a grid cell actually located near a ridge as being near a stream (or *vice versa*).

10.4.7 Assigning joint probabilities

Joint probability value tables were 'seeded' from the same dataset of site sample points from which the prior probabilities were set. Due to evident bias in the sampling, the numbers presented by the 'seeding' required considerable modification. They did, however, provide an indication of proportions within classes that had been over-sampled. Data panels 10.1 to 10.7, located in Appendix A, show the joint probabilities used in this analysis.

As an indication of the process involved in the setting of these joint probabilities, a potential logical conflict between probabilities is worthy of comment. The soil surveyor consulted as an expert indicted that the sandy textured class, with clay content less than 5 percent, would never occur on soils developed by weathering processes over the granitic geology units. At the same time he suggested that, on and near outcropping rock, the sandy texture would predominate. On the strength of the first statement alone, the joint probability between the granitic units and the low clay class would have been set at 0 percent. This would have prevented any other data layer from supporting such a class at those locations. In order to accommodate the

knowledge contained within the second statement, this joint probability was, therefore, modified to reflect the proportion of the granitic area in which rock outcrop occurs. These proportions were derived by cross tabulation in GIS of the geology and rock data layers.

10.4.8 Test data sets

A test data set should be totally independent of input data. However, in this case the site data set had only been used as a guide to setting the joint probabilities. Extensive modification had been made to these values in the light of expert input. This was considered to justify the use of the same sample points to test the accuracy of the output. The 189 points had been classified into the appropriate clay classes in order to set prior probabilities and were directly useable for comparison with the 'most probable class map' derived from the analysis.

An additional potential test data set was available in the form of a soil map for the project area. This had been compiled by the local farmers, led by an Agriculture WA project officer. The individual classes in this map were then ranked into the appropriate clay classes by the same expert soil surveyor who provide input to the knowledge base. Unfortunately, a number of classes in this soil map were so broad that they spanned two of the classes in the Expector analysis.

10.4.9 Output generation

On completion of the knowledge entry phase, the data combination algorithms in ArcView version of Expector were run. This resulted in the production of three maps, each one showing the probability of occurrence of one of the classes of surface texture. These maps are shown in Figure 10.6. From these, a fourth map showing the most probable class was generated using the **Pmax** function from the Arc View interface to Expector. The most probable class map is shown in Figure 10.7. The most probable class map was used as a basis for testing against both of the test datasets described in Section 10.2.8.

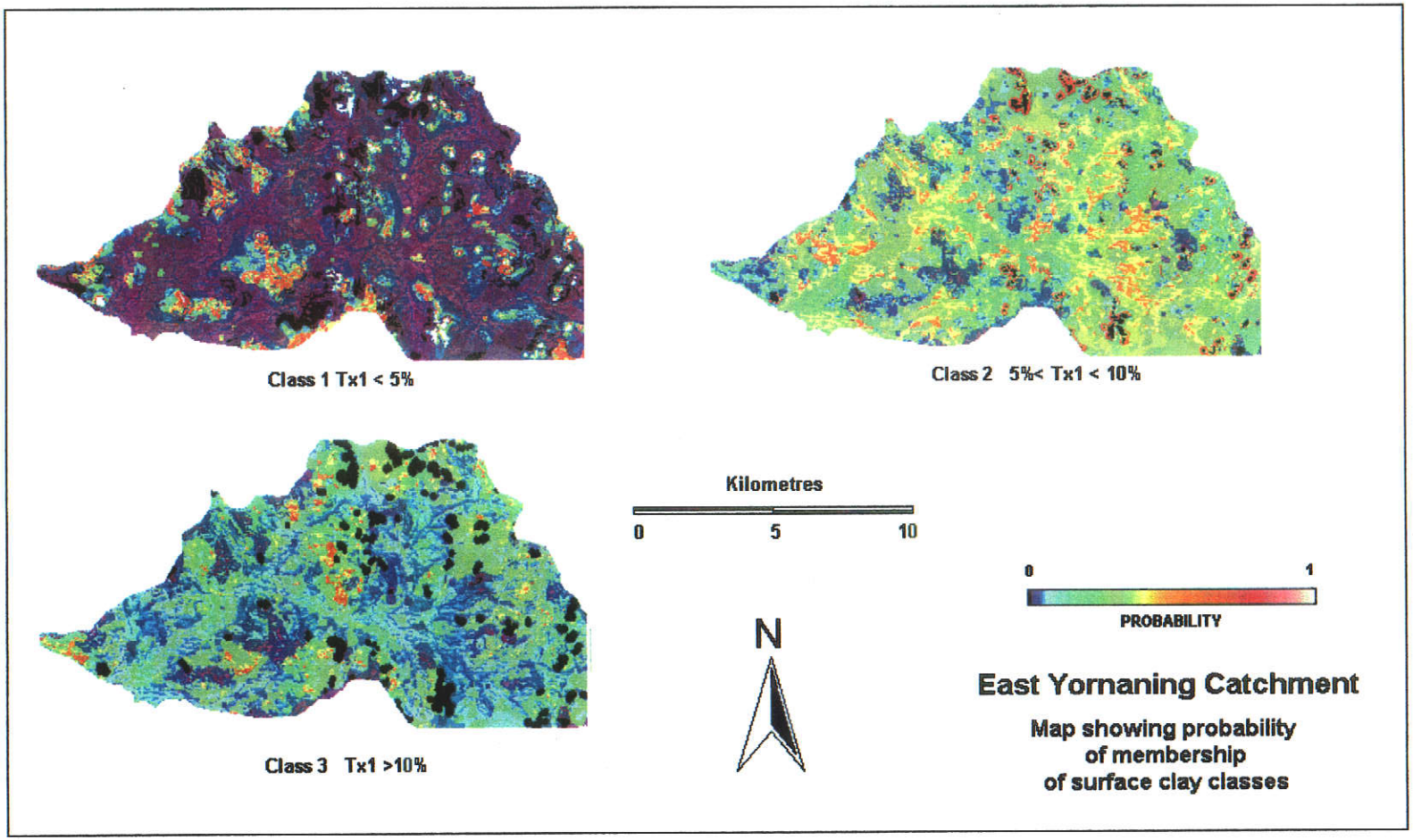

Figure 10.6 Maps of surface clay class membership, East Yornaning

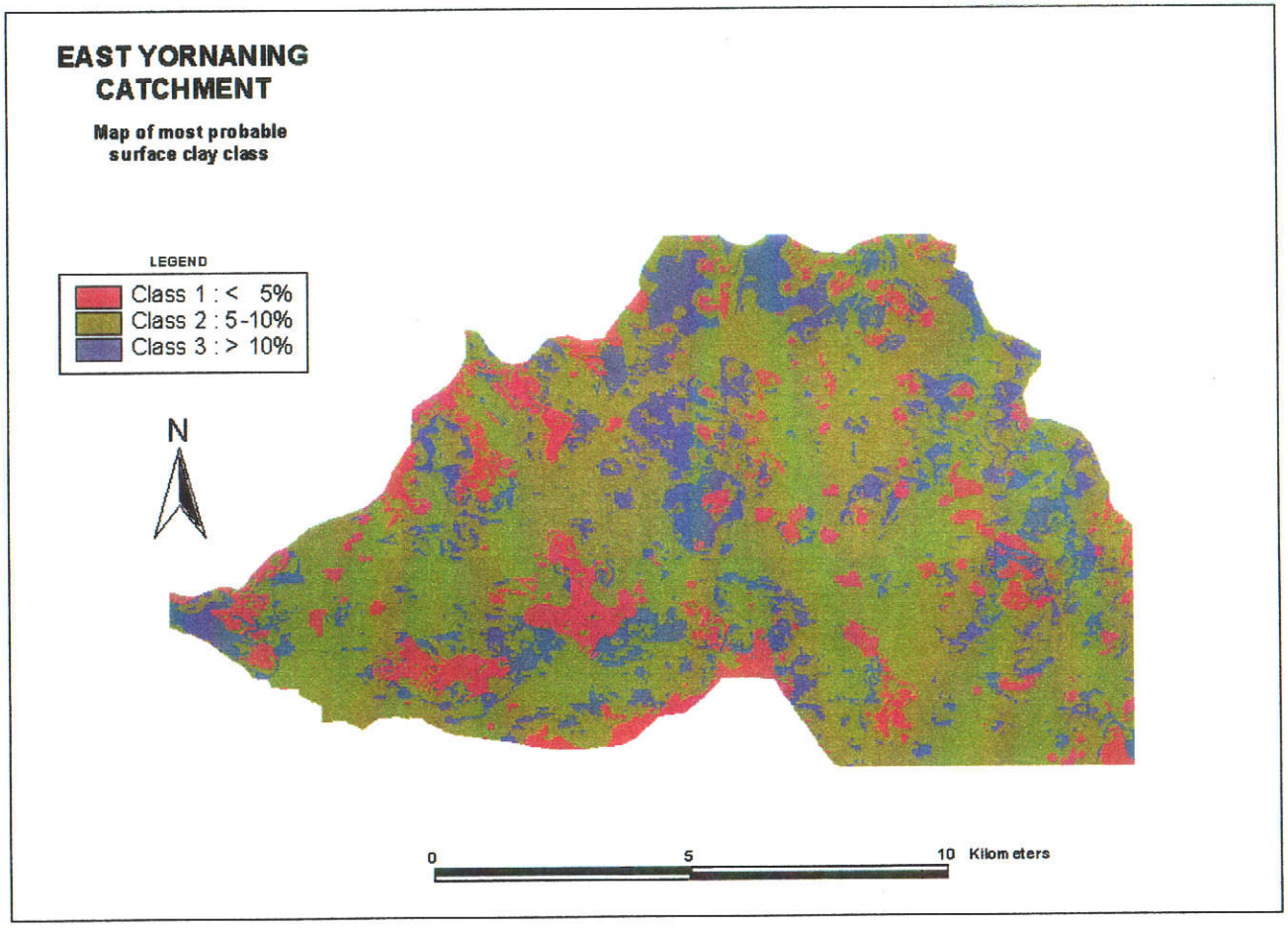

Figure 10.7 Most probable clay class map, East Yornaning

10.5 East Yornaning output maps - comparison to sample sites

10.5.1 Direct comparison at sample points

The class labels in the most probable class map for Yornaning were compared to the actual measured clay content at the 189 sample points. This comparison was performed using the same routines in the ArcView interface to Expector as had been used for the generation of Cross-tab tables in the knowledge building stage. Table 10.3 is a confusion matrix for this comparison. Table 10.4 shows the number and percentage of sites in each class, as measured in the field which are correctly allocated by the most probable class map.

Cell counts	Expector			
Field	1	2	3	Total
1	4	21	1	26
2	8	75	12	95
3	12	35	21	68
Total	24	131	33	189

Table 10.3 Confusion matrix for East Yornaning clay classes

Field measurement	Number of sites	Correctly allocated	Percentage correct
Class1	26	4	15.4
Class2	95	75	78.9
Class3	68	20	29.4
Overall	189	99	52.4

Table 10.4 Allocation table for East Yornaning clay classes

Table 10.4 shows that, overall, just over 52 percent of the cells were correctly allocated. The figure is considerably higher for Class 2, the most numerous class. Since the field measurements were taken at points whose location was surveyed in from aerial photographs, there is the possibility that some mis-location may have taken place. This may mean that the misallocation error is due to spatial inaccuracy rather than the inadequacy of the knowledge base. A neighbourhood approach to testing is, therefore, worth investigation.

10.5.2 Comparison in a local neighbourhood

An ArcView routine was written to extract the class number of all cells in a three cell by three cell neighbourhood, centred on that cell in which each site measurement is supposed to lie. Table 10.5 shows the results of that data extraction. From this it can be seen that for more than 62 percent of the sites there was at least one correctly allocated cell within the nine cell neighbourhood, that is within 50m. Slightly more than 53 percent of sites had a majority of cells within the neighbourhood correctly allocated.

Field measurement		Cells in nine cell neighbourhood which were correctly allocated					
Class	Total sites	One or more		Two or more		Five or more	
		Sites	%	Sites	%	Sites	%
Class1	26	5	19.2	5	19.2	4	15.4
Class2	95	84	88.4	82	86.3	77	81.0
Class3	68	31	45.6	27	39.7	20	29.4
overall	189	118	62.4	111	58.7	101	53.4

Table 10.5 Results of neighbourhood comparisons

10.5.3 Second most probable class

The most probable class map is essentially a device for presentation of the several (in this case three) individual probability maps. The individual probability maps contain information about the strength with which each cell is assigned to a particular class. The relative rankings of these give some indication of the ability of the knowledge base to partition the area between the classes. In some cases the second most probable class may not rank far behind the most probable. In order to determine the degree to which cells were misclassified by the method, an examination was made of the second most probable class.

The three individual class probability grids were exported to ARC/INFO and an Arc Macro Language (AML) routine written to generate a grid whose values corresponded to the class having the 'second chance' probability. This grid was then passed back to ArcView and values extracted from it indicating the second most

probable class at each of the site sample locations. The results are summarised in Table 10.6.

This shows that, overall, only about 30 percent of the sites which were incorrectly allocated by the most probable class map were correctly allocated by the 'second chance' map. It is significant that most of these sites are in Class 2. Combining these sites with those reported as correctly allocated in Table 10.3 gives an indication of those sites which were either correctly allocated or for which the correct class had the second highest probability. These results are summarised in Table 10.7.

Field measurement	Sites initially mis-allocated	Correctly allocated	
		Sites	%
1	22	4	18.2
2	20	19	95.0
3	35	0	0.0
Total	77	23	29.8

Table 10.6 Sites correctly allocated by 'second chance' map

		Correct class probability ranked first or second	
Field measurement	Number of sites	No. of sites correct	Percentage correct
Class1	26	8	30.3
Class2	95	94	98.9
Class3	68	20	29.4
overall	189	153	64.5

Table 10.7 Sites correctly allocated on first and second chance maps

Table 10.7 indicates that 64.5 percent of sites were either correctly allocated or had the correct class ranked second. However, since we are dealing with only a three class system this must be treated with some caution. There is a considerable disparity between the allocation accuracy for Class 2 and that for the other two classes. This will be further discussed in Section 10.7.

10.5.4 Overall accuracy of Yornaning map

To summarise, comparison of the Yornaning output maps with the site sample data suggests that soil surface texture has been correctly predicted at 52.4 percent of the sample sites. At 64.5 percent of the sites the correct surface texture class has been either correctly predicted or is the second most probable class. A relaxation to include the presence of at least one correctly classified neighbouring grid cell give a most probable class prediction accuracy of 63.5 percent. The innate variability of the map can be measured, using the neighbourhood analysis capabilities of ArcView, as the proportion of cells which have two or more classes in a nine cell neighbourhood. For the 'most probable class' map, 37 percent of all cells exhibit that degree of variability, indicating that the map units are far from homogenous.

Whilst an overall accuracy of 63.5 percent does not seem spectacular, it must be taken in the context of the accuracy of soil maps in general. Grealish et al. (1994) tested the ability of soil maps to represent various soil properties and, indeed, to represent soil classes. The results of that study showed that, whilst soil maps can predict soil order (Northcote, 1971) at a level of accuracy of around 65 percent, their ability to predict lower levels of classification in the taxonomy of soils rapidly decreases. Table 10.8 summarises the ability of the soil maps used in that study to represent different levels of classification.

Classification level	Example	Accuracy of prediction
1 Order	D (Duplex soils)	67.5%
2 Sub-division	Dr (Red clayey sub-soils)	52.5%
3 Section	Dr1 (Crusting A horizon)	29.0%
4 Class	Dr1.1 (No A2 horizon)	1%
5 Primary profile form	Dr1.11 (Acid reaction trend)	1%

Table 10.8 Accuracy of prediction of soil maps
(After Grealish et al. (1994)).

Grealish et al. (1994) also tested the ability of soil maps to represent individual soil properties. Those tests were carried out using the method of relative variance (Becket and Webster, 1971). On this basis, which calculates a score on a scale of 0 to 1 (with 1 indicating perfect representation) soil properties fell within the range of

0.1 to 0.32. That figure suggests that the best representation of soil properties by a traditional soil map would be an accuracy of a little over 30 percent.

10.6 East Yornaning output maps - comparison with soil map

10.6.1 Testing the soil map

Given the reported low accuracy of soil mapping, the farm soil map was tested against the reference site dataset in the same way as the Expector output map. Again, a nine-cell neighbourhood around each sample site was analysed to reduce the effect of short-range variation. Table 10.9 shows, for each of the sites, the class of the farm soil map occurring at that location. A very large proportion of sites (152 out of 189) fall either in unmapped areas or in the area where no distinction can be made in the farm map between Classes 2 and 3. As described above, this is due to large range of clay values which occurs in the soil types mapped as being present in these grid cells.

Field measurement	Class on farm soil map					
	Not mapped	Class 1	Class 2	Class 3	Mixed class	Total
Class 1	1	0	4	0	21	26
Class 2	2	2	15	4	72	95
Class 3	2	3	4	5	54	67
Total	5	5	23	9	147	189

Table 10.9 Comparison between site samples and farm soil map (all sites).

Table 10.10 examines in more detail the 37 sites where a useful comparison may be made. Overall, some 58 percent of sites have a majority of cells in the neighbourhood correctly allocated, whilst 54 percent show a correct allocation at the point. These figures are generally similar to those for the Expector output map. It is, therefore, of interest to compare the two maps.

Field measurement		Correct allocation within nine cell neighbourhood			
Class	Total	At the centre		Majority of neighbourhood	
	sites	Sites	%	Sites	%
Class 1	4	0	0.0	0	0.0
Class 2	21	15	71.4	15	75.0
Class 3	12	5	42.6	6	50.0
Total	37	20	54.1	21	58.3

Table 10.10 Comparison between site samples and farm soil map for matched sites

10.6.2 Comparison of Expector map with East Yornaning soil map

The digital copy of the farmers soil map and the most probable class map were compared using the ArcView 'Tabulate Areas' function. This compares the two maps on a cell by cell basis and produces tabular output that summarises the area distribution of the categories of one map across the categories of the other as a cross-correlation matrix.

Table 10.11 shows the cross correlation matrix between the farmers soil map and the most probable class map. Of particular interest here are the large number of cells (142677 out of 188765) which are either not mapped in the farm soil map or cannot be adequately assigned to one of the Expector classes.

Farm map Class	Expector class			Sum
	Class 1	Class 2	Class 3	
Not mapped	987	2676	1122	4785
Class 1	2228	6933	4139	13300
Class 2	768	16023	3180	19971
Class 3	1423	8000	3394	12817
Class 2 or 3	13188	95481	29223	137892
			Total cells	188765

Table 10.11 Comparison of Expector map and farm soil map

In general there seems to be little agreement between the two maps. However if we consider only those cells for which an unambiguous statement is made by both maps we have the situation as presented in Table 10.12.

Farm map Class	Total cells	Expector class			Cells in agreement	
		Class 1	Class 2	Class 3	Cells	%
Class 1	13300	2228	6933	4139	2228	16.8%
Class 2	19971	768	16023	3180	16023	80.2%
Class 3	12817	1423	8000	3394	3394	26.5%
Overall	46088	4419	30956	10713	21645	47.0%

Table 10.12 Comparison of Expector map and farm soil map for matching classes

Table 10.11 suggests that the agreement between the two maps is best in Class 2 and successively worse in Classes 3 and 1, respectively.

10.7 Summary

The results of two studies carried out during the development of the Expector method have been presented, together with an inspection of their accuracies. The more exhaustive of the two, at Yornaning, has also investigated the accuracy of a traditional soil map used as part of the validation process. This was performed by comparing the traditional map to a site sample database as well as to the Expector output.

In the Yornaning example, Expector has produced a clay class map which is of comparable accuracy to one developed by traditional mapping. It did so using much the same evidence and thought processes as were used in the development of the traditional map, but applied those in a quantitative and formalised method. By virtue of its quantitative nature, such an analysis is not only repeatable but also open to improvement by the inclusion of additional data. In addition, there is scope for refinement of the knowledge base in the light of the results of the first analysis. This topic will be discussed further in Chapter 12. The next chapter details some other applications of the Expector method which formed part of a demonstration of that method to land resource mapping professionals.

Chapter 11

FURTHER EXAMPLES OF THE EXPECTOR METHOD

As part of a programme to acquaint the natural resource mapping community of Australia with Expector, a number of demonstration projects were undertaken. These included the mapping of 'modal soil types' near Brookton in Western Australia, surface clay content near Bundaberg in Queensland, and land capability classes at Forth in Northern Tasmania. In all cases, the knowledge base was provided by experts from the soil mapping agency of the state involved. These experts have considerable experience of the areas being mapped and of the attributes represented by that mapping. A report of those three case studies forms the first part of this chapter. The chapter concludes with an example of the use of Expector to predict agricultural yield potential. This is of benefit in the context of precision agriculture when precise targeting of variable fertiliser rates is called for.

The first three of these case studies were carried out solely as demonstration projects, often with minimal resources and were not intended to prove the accuracy of Expector. Indeed, in none of the cases did available resources permit a formal evaluation to be carried out. The intention was to demonstrate the method as a means of ordering, quantifying and formalising the land mapping process. The agricultural yield prediction example includes a more formal assessment of its accuracy.

11.1 Brookton - Western Australia

The objective of this demonstration project was to show how attributes that can be readily mapped by aerial photo interpretation can be combined with products from a digital terrain model to predict the distribution of soils over an unmapped area. The expert knowledge base was provided by field officers of Agriculture WA, the agency responsible for soil mapping in Western Australia. The units to be mapped were soil units similar to those in a traditional soil-landscape survey. This was a departure from the intended principle use of the Expector method as a tool for mapping soil attributes. This departure was necessitated by the fact that the conceptual landscape

models developed by the field officers, and which form the knowledge base, were directed towards mapping soil units.

11.1.1 Site location

The trial area selected is about 30,000ha in area and is situated 20km. west of Brookton in the Eastern Darling Range (Fig11.1). The area comprises both ancient lateritic plateau and dissected landscape. For each of these areas, a unique set of rules can be developed which describes the spatial variability of the soils. The demonstration project was run only for the old lateritic surface.

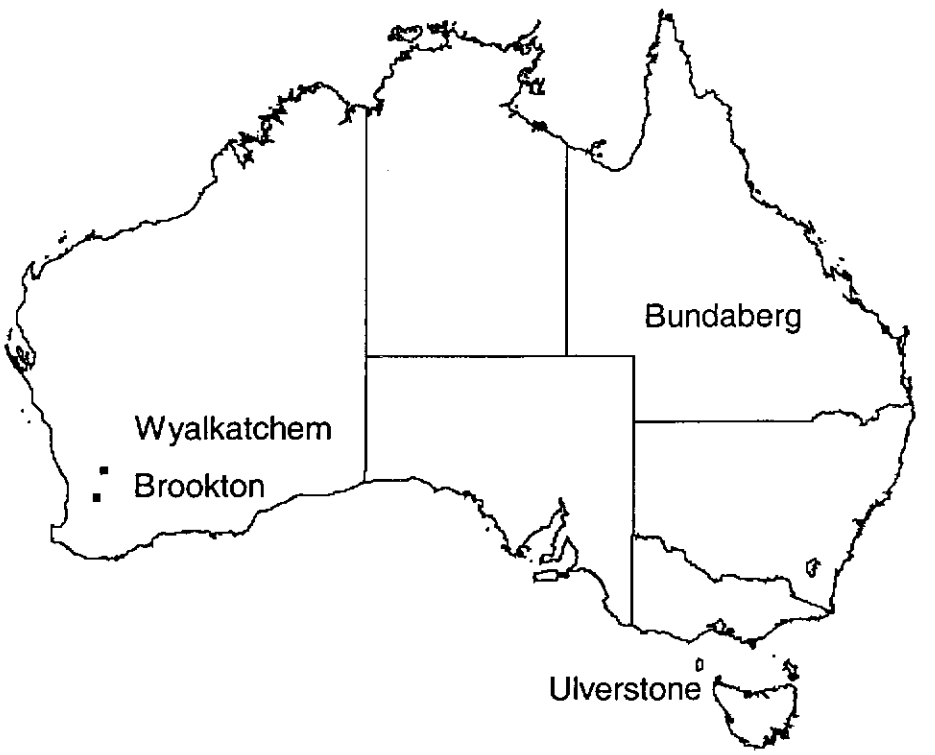

Figure 11.1 Location map for demonstration sites

11.1.2 Available datasets

Data sets considered by the experts for inclusion in the Expector demonstration included digital elevation model derivatives, prior geological mapping, and a number of features mapped from aerial photography.

The base digital elevation model was created by interpolation from 5m contour data supplied by the West Australian Department of Land Administration. These data were taken from 1:50,000 survey maps. A spline interpolation routine was used to resample the data onto a regular grid with a horizontal resolution of 25m. A section of this grid corresponding to the study area was extracted from the larger data set. Data layers representing a full set of terrain attributes including slope, plan curvature, and profile curvature were generated using the TAPES-G software (Gallant and Wilson, 1996).

Agriculture WA obtained colour aerial photography at a scale of 1:25,000 for use in the routine mapping of the area. Their staff digitised a number of features from that photography and delineated a boundary between the 'old' lateritic and 'new' dissected surfaces. In addition, data layers defining dolerite dykes, large areas of deep sand, and granitic rock outcrops were digitised from the photography. All boundaries on the aerial photographs were identified by field survey officers and drawn onto those photographs. The photographs were then registered to Australian Map Grid (AMG) using points that were referenced in the West Australian Department of Land Administration digital cadastral data base. The boundaries were then digitised as Intergraph Microstation design files.

These design files were imported into ARC/INFO and appropriate polygon identifiers attached. These vector datasets were rasterised at 25m resolution and co-registered with the previously discussed data.

11.1.3 Development of schema

The field officers acting as experts in this demonstration could devote only a limited amount of time to this project. It was, therefore, decided to develop a relatively simple schema to link three principal soil types with three pieces of evidence.

Based on their knowledge of the area, they identified three modal soil types as occurring with reasonable frequency in the old surface portion of the study area. These are white sands, yellow sands and ironstone gravels. The model the field officers have developed to predict the occurrence of these soil types is highly

dependent on landscape. Figure 11.2 shows the landscape model used and Figure 11.3 schematically describes the model.

11.1.4 Selection and preparation of evidence

Three evidence data sets were chosen for this study. They were slope, topographic position, and the presence of sand as mapped from air photos. All evidence layer preparations were carried out in ARC/INFO, although the Expector analysis was performed using the ArcView interface. At the grid dataset level, the two systems are entirely compatible.

The landscape model developed by the field staff suggested that five classes were appropriate for the slope evidence layer. The ranges of these classes are shown in Figure 11.3.

The conceptual model relied in part on the concept of position in the landscape or catena. In undulating terrain like this study area, compound topographic index (wetness index) (Moore et al., 1991) provides a surrogate for this entity. A compound topographic index data layer was generated from the slope map and classified into three classes representing upland, mid-slope, and lower areas.

The sand evidence layer was a simple presence and absence map derived from the polygons digitised from the 1:25,000 air-photos.

11.1.5 Prior probabilities and map purities

Initial counts of class abundance for each evidence layer were extracted using the ArcView interface and used to assign class probabilities using the processes described in Sections 9.5.1 and 9.6.2.

The experts assigned map purities to each of the chosen evidence data layers. The map purity data is listed in full in data panels 11.1 to 11.3 in Appendix A. Those layers derived from the digital elevation model were assumed to be 100 percent correct. The sand layer was believed to contain some inaccuracies due to misinterpretation of textures on the aerial photography.

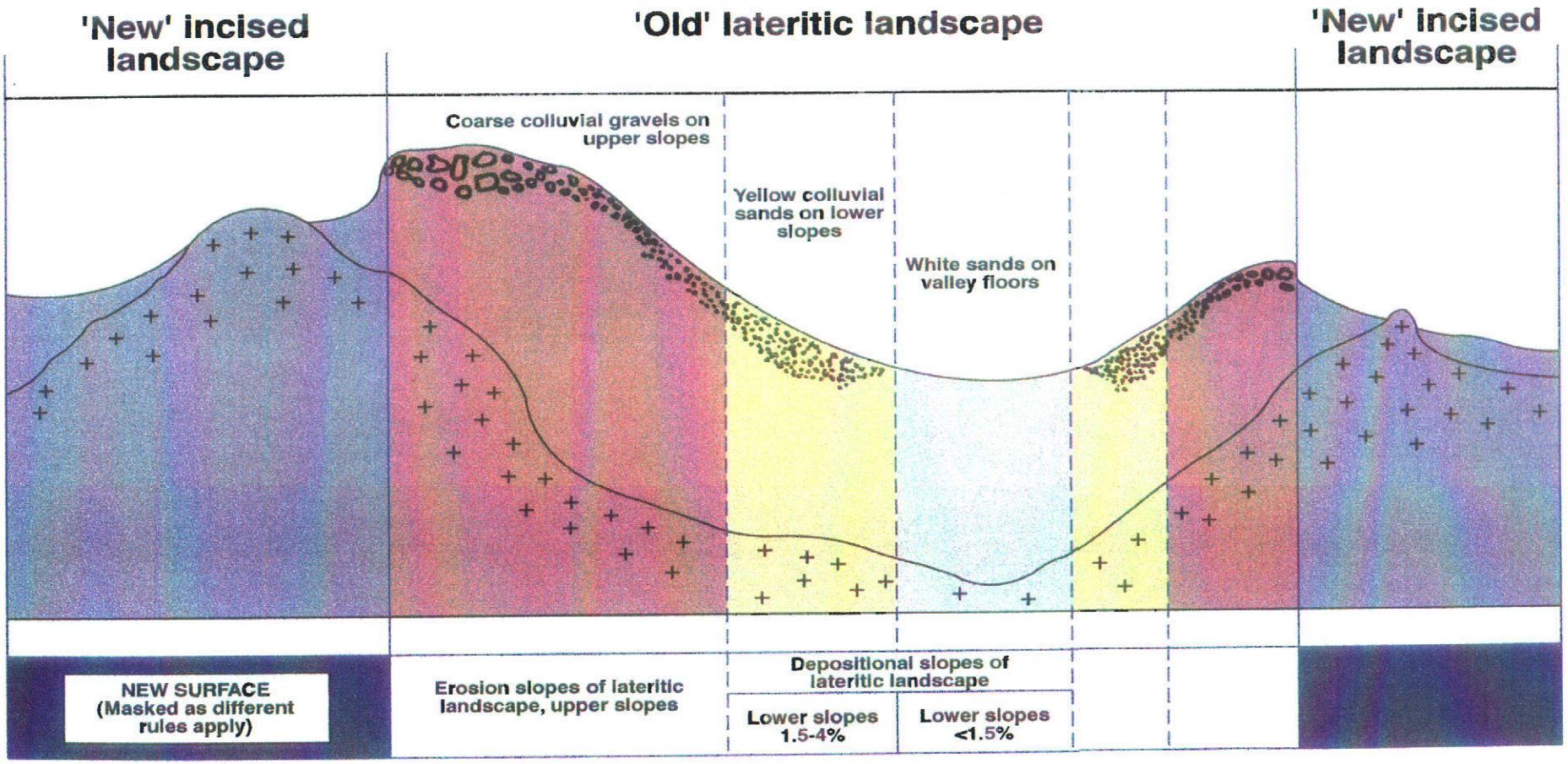

Figure 11.2 Conceptual model of landscape at Brookton (from Verboom et al. (1997))

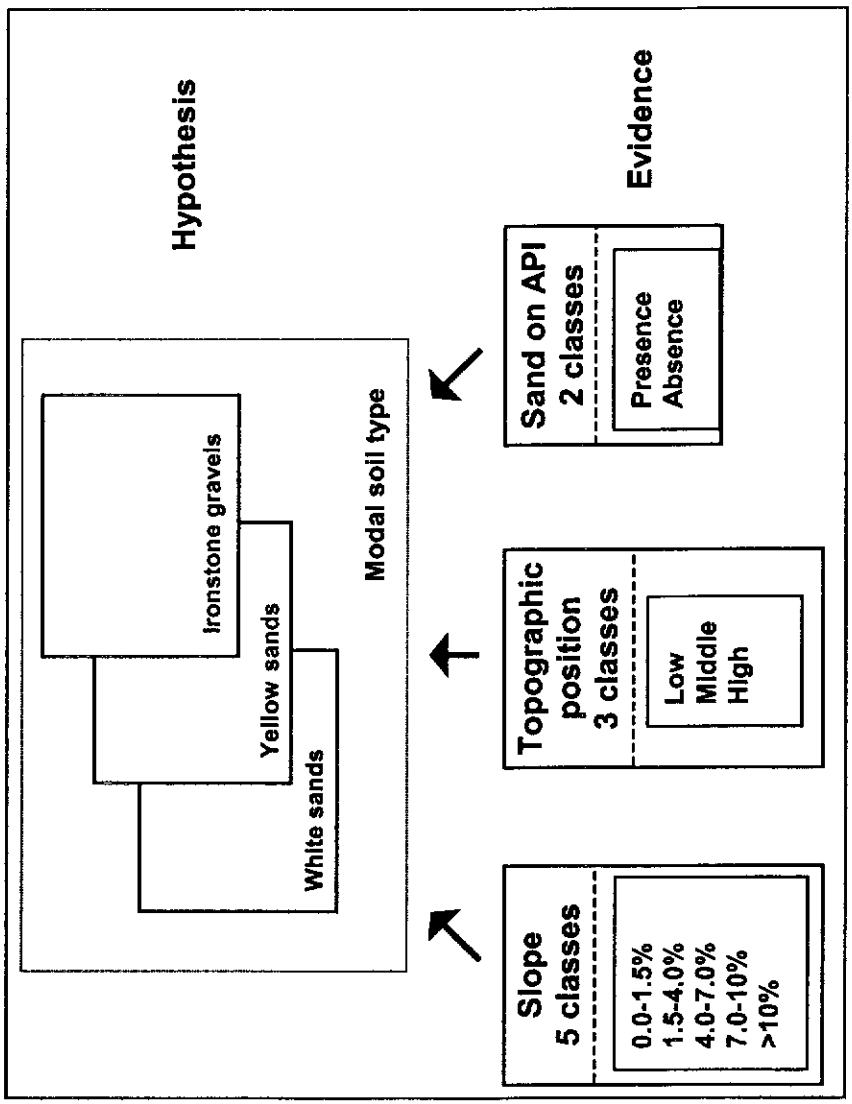

Figure 11.3 Schema for Brookton study

It was considered that the probability of areas of sand being correctly mapped was relatively high, with a value of 0.96 being assigned. Since there might well exist (within the larger area identified as not being sand) some small unmapped pockets of sand, the non sand class was assigned a lower purity of 0.85.

11.1.6 Using site data in the knowledge base

A field data set comprising 67 sample sites throughout the study area had been acquired by the field officers as part of their routine survey of the area. These sites were classified on the basis of observations in field notes into the three modal soil types. The proportion of each type occurring in the samples set was then used as the prior probability for that 'modal soil type'.

The Expector ArcView interface tools were used to extract Cross-Tab files using the method described in Sections 9.5.2 and 9.6.4. These provided seed values to the Expector Joint Probability Editor. The editing process reshaped the coincidence tables to more truly reflect the soil surveyors' belief in the relationship between the evidence data sets and the soil types. Since they had been personally involved in the sampling, they had a good grasp of possible biases in the system. Joint probability data are listed in full in data panels 11.1 to 11.3 in Appendix A.

The joint probabilities used for the slope layer express the strength of a number of linkages in the conceptual model. For example, ironstone gravels have a high probability of occurrence on slopes above 4 percent, whereas sands predominate on lower slopes. Moving to very low slopes, white sands are believed to be predominant.

Joint probabilities for the air-photo interpreted data layer express the belief in a very close correspondence between the presence of sand in this data layer and the two 'sand' soil classes.

Again, a few clear relationships are also seen in the joint probabilities between topographic position (as expressed by compound topographic index) and the soil classes. The field officers clearly believe that the ironstone gravels appear almost exclusively high in the landscape, whilst the white sands appear almost exclusively

low in the landscape. Reference to the graphical representation of the landscape model in Figure 11.2 will confirm this. It should be noted that this 'rule' will compete, in a probabilistic sense, with that which assigns white sands strongly to low slope areas.

Areas of low slope that are low in the landscape, such as valley bottoms will tend to be assigned as white sand, whereas areas of low slope higher in the landscape will have white sand favoured due to slope but gravel favoured due to position in landscape. This will combine to give a close tie between white sand and gravel. In this case, the third factor (based on the air-photo interpretation) will act as a 'tie breaker' since it provides, where present, very strong support for white sands.

11.1.7 Results

This case study was intended primarily for demonstration purposes and to familiarise the land resource surveyors with the Expector method. Although no quantitative measures of accuracy were taken the field staff involved believed that the resulting maps were generally satisfactory. Output comprised maps of the probability of occurrence of each of the three soil types and a map showing the most probable soil type. Figure 11.4 shows the map of the most probable soil type for the southern half of the study area.

11.2 Bundaberg - Queensland

A demonstration project was carried out in the Bundaberg region of Queensland (Figure 11.1) with expert knowledge being provided by staff from the Queensland Department of Natural Resources (QDNR).

Bundaberg is in a sugar cane growing region and there is demand for land resource information for local and regional planning at a semi-detailed scale of 1:50,000. The land resource information is usually presented in terms of hazards or land qualities which are a synthesis of several, more basic, resource attributes. One of those basic attributes is surface clay content which was chosen as the demonstration attribute.

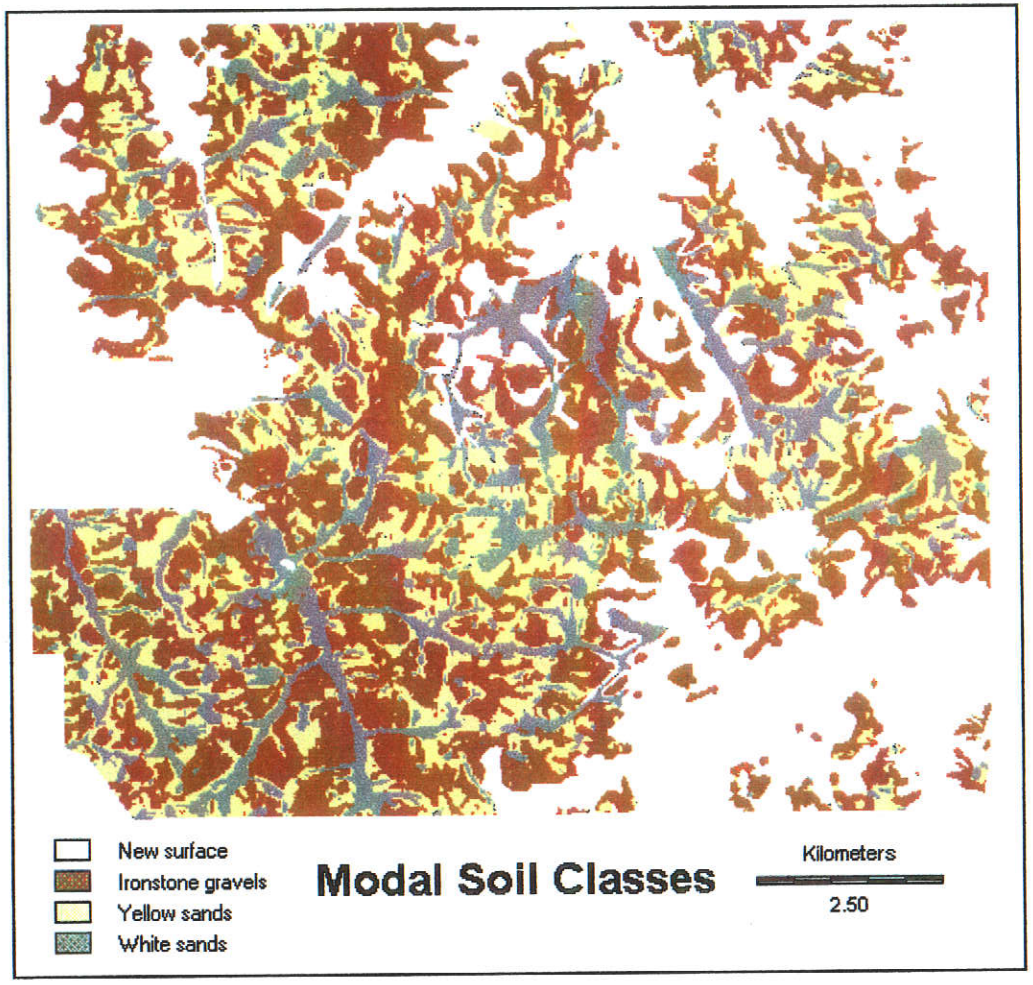

Figure 11.4 Most probable class map, Brookton

11.2.1 Objective and location of study

Local staff were interested in assessing the usefulness of the Expector method to spatially extend their conceptual knowledge. Conceptual knowledge had been developed in an area where the geology, geomorphology and soils are well understood and have been mapped. The study area has similar geology and soils, but the soils and soil properties had not yet been mapped.

The trial area selected was approximately 30,000ha which forms part of the Childers 1:100,000 map sheet to the south-west of Bundaberg. The objective was to predict the surface texture as membership of three broad groups. Those groups were determined, with reference to local conditions, as being 0-20 percent, 20-35 percent, and >35 percent clay, respectively.

11.2.2 Available data sets and schema development

A digital elevation model of the area had been generated by QDNR. From this, slope and compound topographic (wetness) index layers were derived. A geology map was also available in digital form. The slope map and wetness index maps were divided into classes which accorded with the surveyors conceptual model of that particular landscape. Figure 11.5 shows a schematic diagram of the model used.

Also available was a data set of 242 sample sites at which surface texture had been measured. These were used to determine prior probabilities for the three texture classes. Query tools in the Expector ArcView interface were used to generate cross-tabulations between the evidence layer classes and the hypothesis classes. These cross-tabulations were then edited using the Expector Joint Probability Editor to remove bias and incorporate the surveyors knowledge. Figure 11.6 shows the geology evidence layer overlain by the sample sites.

Map Purity and Joint Probability data are shown in full in data panels 11.4 to 11.6 in Appendix A. The three input data layers together with the three probability maps (one for each class of surface texture) are shown in Figure 11.7. As with the previous study, a map showing the most probable class is included.

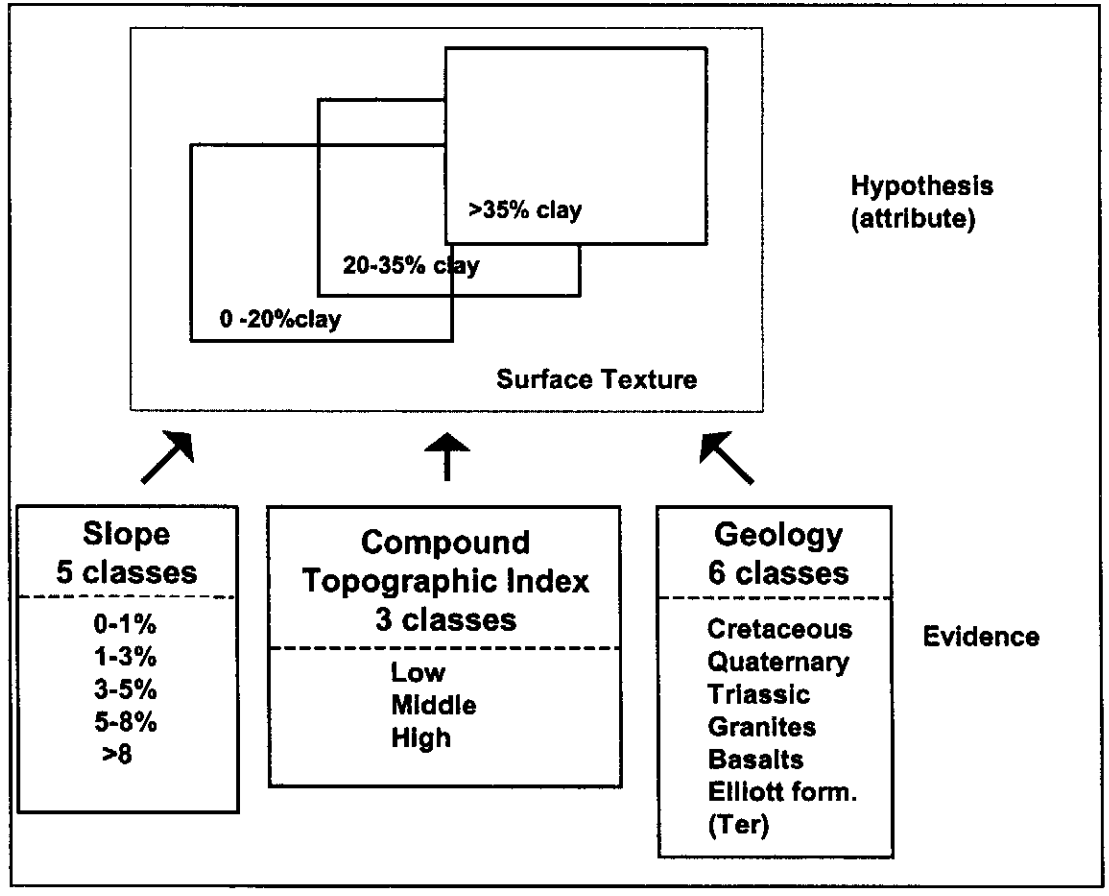

Figure 11.5 Schema for Bundaberg (Childers) study

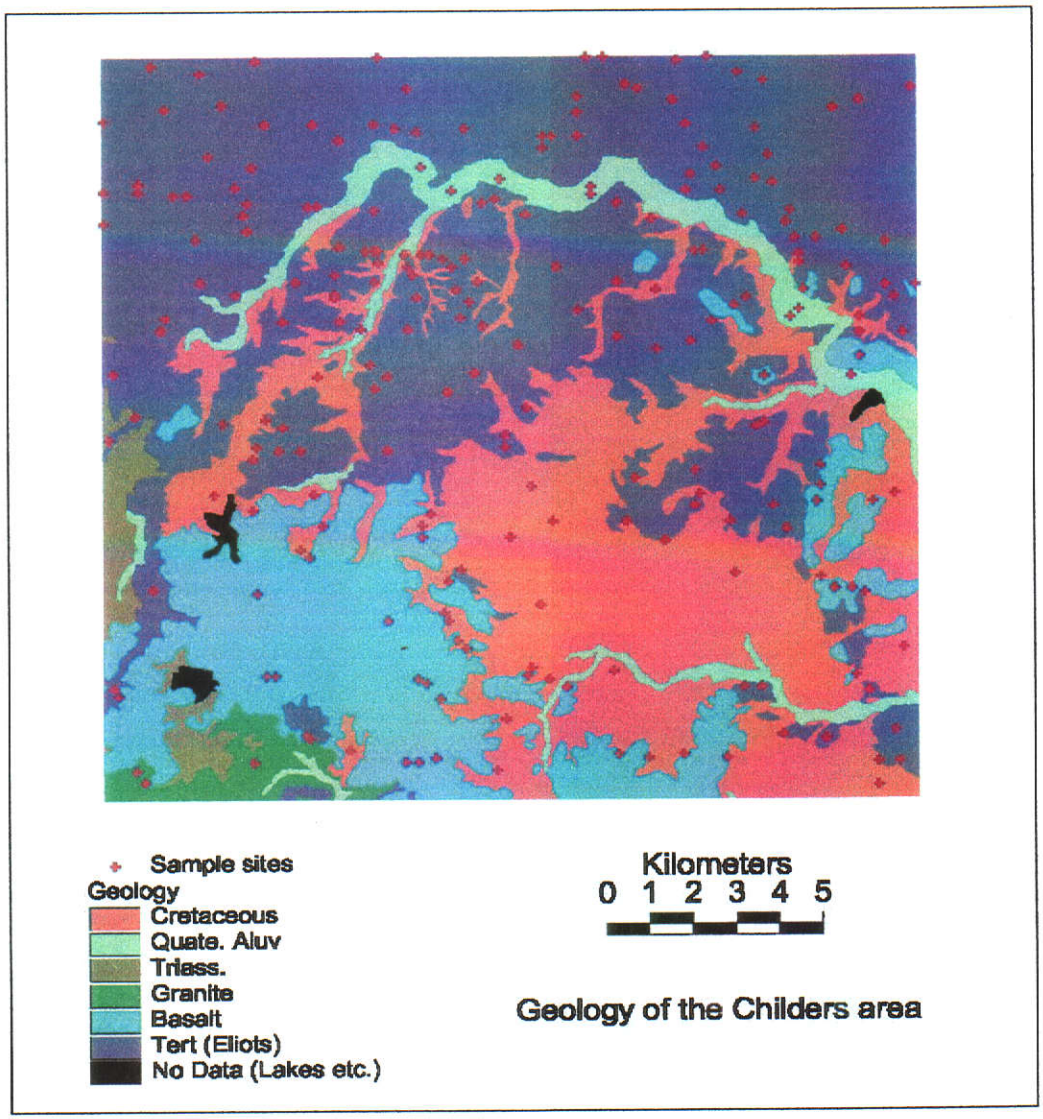

Figure 11.6 Bundaberg: site data overlying geology

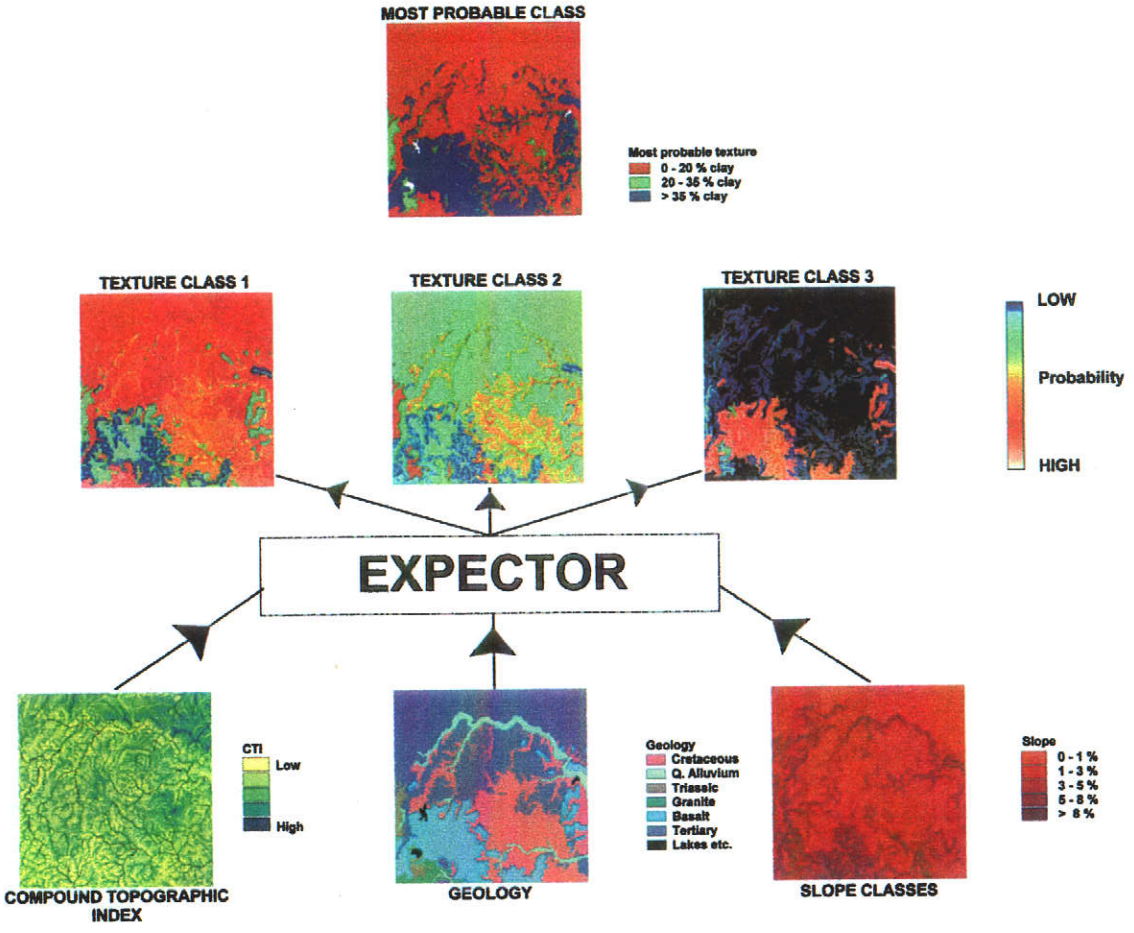

Figure 11.7 Graphical illustration of the Bundaberg demonstration

11.2.3 Results

The Bundaberg demonstration project was put together rapidly in order to demonstrate the concept of the Expector method. Discussion with field staff suggested that the data chosen as evidence was less than optimal for separating out the surface clay classes. An examination of the individual probability maps indicates that they are very heavily controlled by geology. In most areas, a clear prediction of surface texture has been made. However, in the area of Cretaceous deposits in the eastern part of the study area, the first two texture classes have roughly equal probability. This is described by field staff as being a difficult area to map, with both textures occurring within that unit. This is reflected in the equal posterior probabilities of the classes. An additional data layer is required as a 'tie-breaker' to determine which of the two texture classes is prevalent at any particular location. At the time of the demonstration, no suitable data sets were available.

11.3 Forth -Tasmania

A case study that looked at the use of Expector for land capability classification was carried out in conjunction with personnel from the Tasmanian Department of Primary Industry and Fisheries (DPIF), Launceston.

11.3.1 Location and objectives

In Tasmania, a classification method based on that used by the United States Department of Agriculture (USDA) is used to map land capability. The classification comprises seven classes, ranked in order of increasing degree of limitation to use and in decreasing order with respect to versatility. Class 1 has virtually no limitations to intensive cropping, while Class 7 is unsuitable for agriculture.

At the time of the demonstration, DPIF personnel were engaged on the production of a land capability map for the Forth 1:100,000 map sheet, near Ulverstone on the north coast (Figure 11.1). This survey was to be conducted using conventional methods (API, fieldwork, etc.). The DPIF team were interested in the ability of the Expector method to produce a reconnaissance map of land capability, rather than soil attributes.

11.3.2 Datasets and schema

Over the whole study area, the main limitations to land use are (in order of importance): topography, climate and soils. Other limitations of importance in small areas include flood risk and rock outcrops. Slope is seen as one of the key topographic variables since it encompasses ease of cultivation as well as susceptibility to mass movement, erosion risk, etc. A number of climatic factors such as rainfall, frost risk and temperature are also important, as are soil related variables like depth, waterlogging, wind erosion susceptibility, and fertility.

Slope, altitude, and geology were chosen from the readily available data sets. Altitude acts as a surrogate for radiation and frost risk, whilst slope has a direct bearing on the workability of the land. The geology of the area, which includes Tertiary basalt, Permian sediments, Pre-Cambrian siltstone, and Quaternary alluvium, has considerable influence on Land Capability. Many of the soils, in particular those formed on Tertiary basalt, have good physical and chemical properties and are highly suitable for agriculture.

11.3.3 Results

Figure 11.8 is a map of the most probable class for the study area. The DPIF staff regarded the results as promising. It was, however, noted that a strong relationship between basalt, slope, and land capability classes 1-3, was more dissipated than expected.

The reason for this dissipation seems to be that the area was treated as a whole. The surveyors' mental models contain nuances of variable behaviour throughout the area that were not represented in the probability assignments. The solution to this problem lies either in partitioning the area into broad sections in which different rules apply or in the addition of an extra layer, perhaps based on a geomorphic or geological divide, for which a moderating set of joint probabilities is developed.

11.4 Agricultural yield prediction.

The Expector method lends itself to any situation in which spatial data layers are combined to make a prediction about some other entity that occupies the same space.

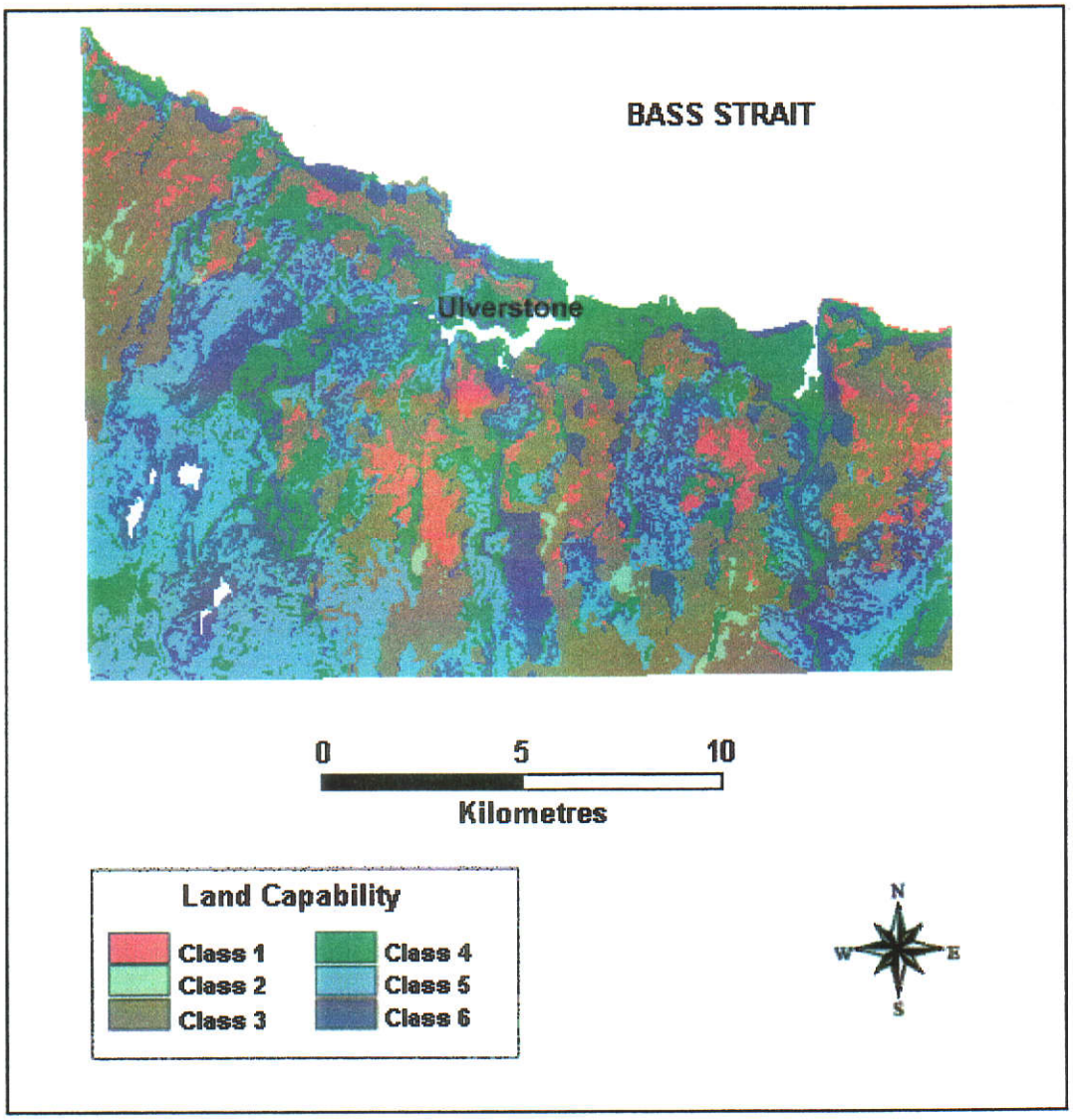

Figure 11.8 Most probable class map for Forth, Tasmania

The only prerequisites are that there be some logical connection between the predictive and predicted layers and that some body of knowledge exist which can explain that connection.

The emerging field of Precision Agriculture is concerned with managing the variability inherent in cropping systems. This variability is caused by many factors, amongst which are spatial variation in the fundamental properties of the soil resource. Precision agriculture deals with deficiencies in this resource by spatially variable application of nutrients and other ameliorants. One strategy used in these circumstances is to vary fertiliser in response to expected grain yield with, for example, extra fertiliser concentrated on areas which are expected to yield well.

Expector was used in a study to determine likely grain yield in an 80ha wheat paddock near Wyalkatchem, Western Australia (Figure11.1). This work was carried out as part of ongoing research into Precision Agriculture by CSIRO Land and Water.

11.4.1 Development of schema

A considerable number of things influence the likely yield from any particular area, and these operate at different spatial scales. The overriding factor in rain-fed agriculture such as that practised in Western Australia is climatic. Whilst rainfall does not vary significantly across a paddock, the way in which the soil stores and handles the available moisture does vary. Such variation is largely the result of soil physical characteristics.

For the paddock in question, a map showing soil types had been prepared, using the conventional mapping techniques of field observation and delineation of boundaries on air photography. This map is considerably more detailed than those usually available and identifies eight different soil types within the 80ha.

An intensive grid sampling campaign had resulted in the collection of soil physical and chemical data at over 40 locations located on a nominal 100m grid. Soil sampling and analysis is relatively expensive, and it is questionable whether such intensive sampling is economically viable for broadacre crop management. The soil

map does offer a cheaper, though less specific, source of soil information. However, in this particular paddock it was known that acidity was a major limitation to crop growth. In the light of these considerations, a decision was taken to use the soil map as an evidence layer representing soil physical properties in general and to incorporate the pH data from soil tests as an additional layer.

Another good predictor of one season's yield is the yield for the previous season. In this case, crop yield data had been collected for the 1995 harvest using a grain yield monitor and a map showing the spatial distribution of yield was available. The intention was to predict the likely yield in the 1996 season using as evidence the soil map, the pH data, and the prior seasons yield map. The attribute to be mapped was defined as 'yield in excess of 1.5 tonnes per hectare (tha^{-1}),' resulting in a two class system comprising this class and its converse. The value of $1.5tha^{-1}$ was based on the approximate break-even grain yield for that farm.

11.4.2 Data classification and determination of prior probabilities.

The soil map is already categorised, with all classes being distinct and requiring no amalgamation. The yield map was classified into six classes, five covering an interval of $0.5tha^{-1}$ each with the sixth class encompassing those few areas in which yield exceeded $2.5tha^{-1}$.

A pH surface was interpolated from the point observations using the spline interpolation function in ArcView. This surface, pH values for which ranged from 4.2 to 4.7, was then classified into five classes each covering a pH interval of 0.1.

Prior probabilities for each of the evidence classes were determined from the areas occupied by each class using the methods previously described. Prior probabilities for hypothesis attribute were set by an expert, after consideration of the previous years yield figures, as a 35 percent probability of achieving better than $1.5tha^{-1}$.

11.4.3 Knowledge base - map purities

Map purities were determined by expert assessment. In the case of the surface derived from pH observations, the quality of the data was considered to be high. All

classes were assigned a purity of 90 percent with the misclassification error distributed to numerically adjacent classes.

The purity of the soil grid was considered to be a little lower. An initial allocation of 80 percent accuracy with the misclassification error evenly distributed across all other classes was adjusted to 79 percent, thereby allowing simple distribution of the remaining 21 percent across the seven other classes.

Since it was measured with a grain yield monitor having a spatial resolution of the same order as the grid used for the analysis, the previous years yield data was considered to be 100 percent accurate. All the map purity values are shown in data panels 11.7 to 11.9 in Appendix A.

11.4.4 Knowledge base - joint probabilities

The joint probability values were determined using the usual combination of seed values extracted for the data, and expert opinion. The joint probabilities are presented in data panels 11.7 to11.9 in Appendix A.

For the pH layer, it is clear that the more acid the soil the greater the chance of a depressed yield. Even the classes with a higher pH are still very acidic and for none is there a significant chance of exceeding the expectations set by the priors.

In the case of the soil data layer, some soils, such as Class 4 (Shallow pale sand) are not expected to cause any divergence from prior expectations. Others such as Class 6 (Deep sandy duplex) and Class 8 (Shallow loamy duplex) cause respectively a decrease and increase in the chance of achieving a better than break-even yield.

The joint probability table for the yield layer shows some apparent contradictions that would not have appeared if the exercise had been data driven. It takes into account the known fact that, overall, wheat yields in a paddock going into a second consecutive year of wheat crop in a rotation show a decline. This is exemplified by the suggestion that, whilst areas which failed to reach the break even point in the previous year are expected to continue to make a loss, areas which just broke even are shown as being more likely to make a loss in the second year. This argument

assumes that no fertiliser is applied and is perfectly legitimate in a scenario that is designed to aid the placement of such fertiliser.

11.4.5 Results

Figure 11.9 shows the input data layers, the probability surfaces for the two states of the hypothesis, and a most probable class map. The subsequent growing season the paddock was, for all practical purposes, treated to uniform application of fertiliser. It is, therefore, possible to compare the results of the prediction with the actual crop in the subsequent year. Table 11.1 shows the basic statistics for the actual yields within the areas predicted as being high or low yielding.

Predicted Class	Actual second year yield (tha^{-1})			
	Min	Max	Mean	SD
1 (<1.5tha^{-1})	0.00	2.48	0.99	0.39
2 (<1.5tha^{-1})	0.29	2.55	1.22	0.38

Table 11.1 Yield statistics for predicted areas

Some separation between the two classes can be seen in these figures. The differences between the two means are of the same order as the difference between the minima. The maxima, however, are very close. There was an overall depression of yield across the paddock, due to factors that were not considered in the schema. These factors included weeds and 'haying off' due to lack of moisture at the end of the growing season. The overall result of these effects was that both the 'high' and 'low' yielding areas suffered from depressed yield, with the mean of both being less than 1.5tha^{-1}.

11.5 Comparison of results and discussion

None of the studies reported here have been subjected to a rigorous formal evaluation. However, in the case of the three demonstration studies, the land resource assessment staff involved considered that the maps produced were a reasonable representation of their mental models, at least within the limitations of the data used. Some quantitative analysis is provided for the grain yield prediction

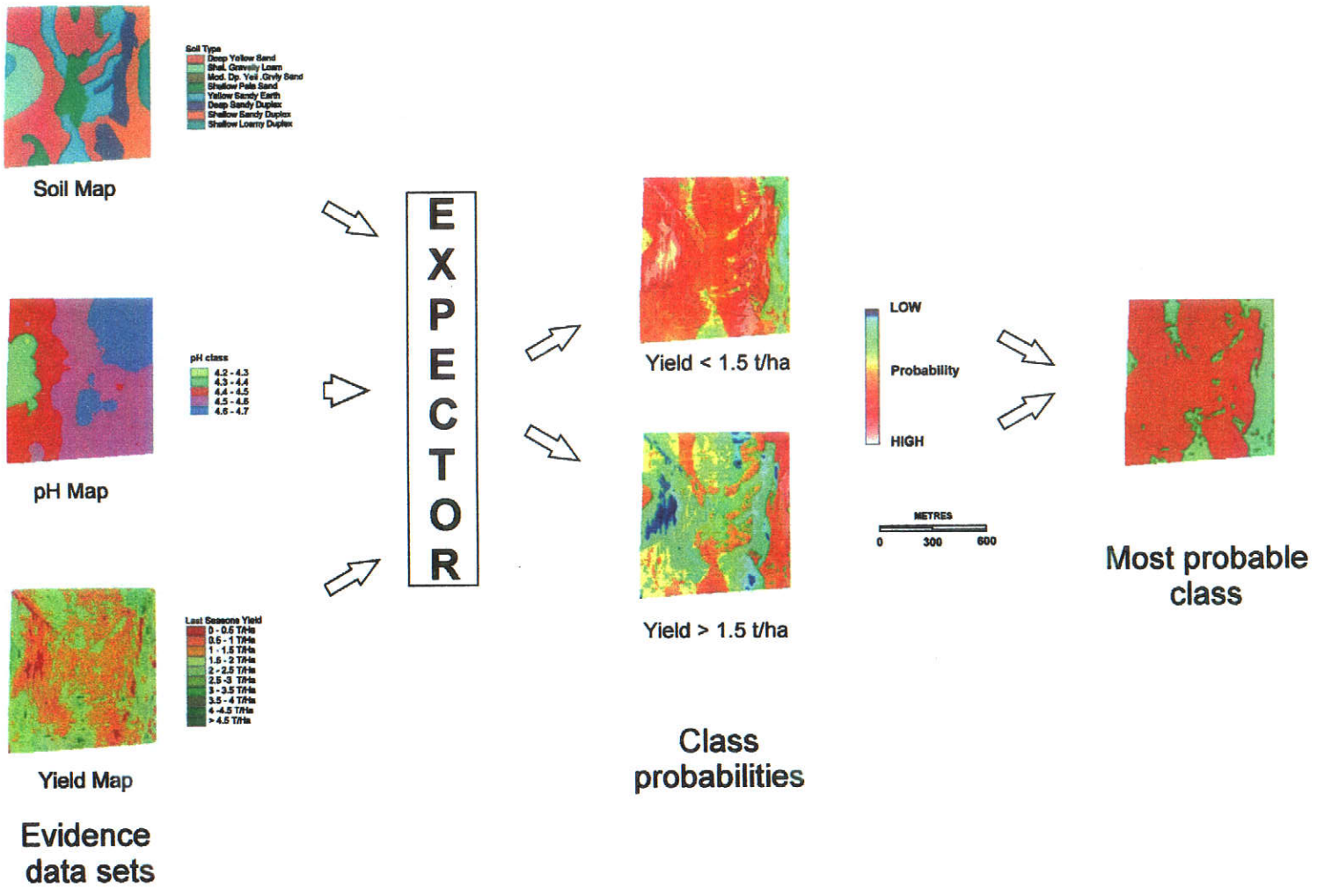

Figure 11.9 Graphical description of yield predication analysis

example. That analysis highlighted the fact that the overall yield of the paddock was depressed by factors which were not included in the model.

A number of operational considerations were highlighted during these demonstrations. These include the question of sufficiency of the data used. In the Bundaberg example, it was clear that an extra data layer was required as a 'tie breaker', whilst in the Tasmanian example there was a need to either partition the landscape and create two models or to add an additional, mediating, layer.

11.6 Summary

Three examples have been presented of the application of Expector to natural resource mapping and one of its use as a tool in the assessment of likely grain yield for fertiliser recommendation purposes. Although most were not quantitatively assessed, these examples give some insight into the processes involved in creating Expector schemae and have highlighted some points for further investigation. The next chapter discusses some of these points and other operational considerations that came to light during the development of the Expector method.

Chapter 12

THE IMPACT OF EXPECTOR ON THE SOIL MAPPING METHOD

The Expector method was developed to quantify and formalise the existing natural resource mapping process. Several examples of applications of the method have been provided in preceding chapters. The impact of introducing a quantitative method may be examined from two standpoints. Firstly, how well does the method achieve its original intentions and, secondly, what effect does its use have on the way a resource survey is conducted. This chapter begins with a discussion of the effects of adoption. It then examines how effectively the method has performed in those examples and discusses some of the sources of error. The opportunity to eliminate some of these errors by refinement of the knowledge base is also discussed. Since some errors can be attributed to sampling bias, the differences between free and grid sampling techniques are examined. The use of Bayesian networks imposes a constraint of conditional independence on input data selection which is not present with mental models. The chapter concludes with a discussion of conditional independence and suggests some tests to determine the eligibility of input data layers.

12.1 The effects of adoption

The adoption of a new method of surveying can have a number of effects. These include the effect on model construction, fieldwork, outputs and the validation of that output. These are discussed briefly here in reverse order and amplified later.

12.1.1 The effect on outputs

The outputs of the process are noticeably changed. Considering a soil survey example, instead of a map representing soil classes, the primary output is in the form of several GIS data layers showing the probability of occurrence of individual soil attributes. This is more flexible than a hard-edged class map and can still be reduced to a classified map if required. This may be either a most probable class map for the attribute in question or a map synthesised in a GIS query from several attribute probability maps. It is also possible to use Expector output in another Expector analysis such as the agricultural yield potential example documented in Chapter 11.

In essence, the output is more dynamic and fluid than a conventional printed map. This greater flexibility has been achieved without sacrificing the ability to produce the conventional product if required.

An important part of the traditional output is the memoir or report that accompanies the map and which includes much information about the spatial variability and composition of individual map units. A set of output data GIS layers incorporates an inherent description of that variability. The result of any query can still be backed up by reference to the individual probability maps. In addition, the knowledge of relationships used in the mapping process is also available in either map or tabular form. For any given input evidence data layer a set of GIS 'virtual data layers' exists showing the probability of occurrence of the mapped attribute based solely on that piece of evidence. These, together with the files used in input data reclassification and the schema files for each attribute mapped, provide readily accessible documentation of the model building process.

12.1.2 Validation

The traditional map of soil types can be validated by visiting a number of sites and comparing the mapped soil type with that observed on the ground. Knowing the parameters that bound the soil class in attribute space, it is easy to determine whether the material at a sample point is appropriately mapped. At a simple level a positive or negative result can be recorded at each sample point and statistics prepared showing the accuracy of the map in general. In practice, the closeness of correspondence of the observed soil to that mapped will be noted and used as part of an iterative process of map improvement.

Maps showing the probability of occurrence of either individual attributes or of soil types are harder to validate. At the extremes of the distribution the case is simple. If an attribute class is mapped as having a probability of 1 at any point then, if the map is correct, we should find it there. The problem is more difficult if the probability of occurrence is mapped as 0.8 or 0.4 or any other indefinite figure. It is reasonable to suppose, however, that if we visit a number of points at which the probability is 0.8 we would expect to find the attribute more often than not. If sufficient samples were taken we would expect the proportion of correct sample to approach 80 percent.

Unfortunately validation schemes based on this principle require an enormous sampling effort. It has been estimated that in order to validate the probability maps for the Yornaning example, a minimum of 1000 sample sites would be required (Fox, 1996). Clearly this precludes the use of limited data sets such as those available for the examples presented here. In addition, such a scheme would make a statement about the accuracy of the map as a whole, but still leave open to question the accuracy of a point where an attribute is, say, 80 percent probable, but is not found.

Alternative validation strategies are therefore required, and two have been used in the examples presented in Chapter 10. One is generally useable, whereas the other can only be used if the attribute being mapped is measured on a numerical scale, rather than representing a presence - absence condition.

Since the attribute being mapped has been divided into classes, field observations may be compared to a most probable class map. This essentially reduces the validation process to the one described above for a traditional soil type map and has been applied to the examples in Chapter 10. In the case of attributes with a numerical scale such as clay content or organic matter, it also is possible to inspect the relationship between the actual values of the attribute and the various probabilities of class membership. This method was used in the validation of the Sterling data (Figure 10.3). In that case a positive relationship was found between the organic matter content at sample sites and the probability that organic matter exceeded 1.6 percent.

12.1.3 Effect on model construction and fieldwork

The surveyor must make some subtle changes to both their conceptual model construction and, to a lesser degree, their fieldwork. A conceptual model must be developed for each attribute being mapped rather than for a broad association of attributes such as soil type. To a large extent, this is what the surveyor traditionally does, although representation as soil types forces them to compartmentalise the models at an early stage of development.

Although the models developed using Expector are formally expressed in numerical terms they are not inflexible. A surveyor using the traditional method will refine a

mental model in response to additional information and validation of the map product. A surveyor using Expector can perform a similar iterative process.

Except where the method is to be used purely for the production of reconnaissance maps, fieldwork still forms an essential part of the model development. The demonstration example from Tasmania reported in Chapter 11 was carried out without any fieldwork by transferring the knowledge developed from work in an adjacent area. It should, however, be noted that this practice is not without its dangers.

The results from the Tasmanian example suggest that an additional evidence layer may be needed to subdivide the study area. This raises the question of the extendibility of models and of "model-drift". A model may become inappropriate, either as the result of crossing some geomorphic divide, or as a result of continuous change such as a steady increase in altitude with distance from the sea. There is an opportunity for further research into the recognition of the bounds of models, and into means whereby they may be made adaptive.

Two alterations to field work practices are suggested by the experience reported in the preceding two chapters. The first is to the design of sampling schemes and is dealt with in more detail below; the other is to the nature of the data collected at sample sites. Expector is capable of using input data layers of varying accuracy and has a facility for developing tables of input map purity. If the data layers to be used are defined prior to the fieldwork stage, then appropriate measurements can be made at sample sites to assist with the determination of those map purity figures.

Perhaps the greatest effect on model building is the selection of evidence data sets. Using a traditional mental model the surveyor is freed from the constraints of conditional independence of evidence imposed by the use of Bayesian nets. The enormous flexibility of mental models allows breaches of independence to be dealt with in a informal way. A surveyor using Expector must keep the question of conditional independence in mind when selecting data sets and may need to combine some datasets into indices, as suggested in Section 8.2.5. A discussion of conditional independence and suitable tests is provided later in this Chapter

12.2 General discussion of results

The two developmental examples discussed in Chapter 10 were subject to more rigorous testing of accuracy than the demonstration examples in Chapter 11. They are re-examined here from the standpoints of overall accuracy, relative accuracy between classes, and sources of error.

12.2.1 Absolute accuracy

At Sterling, the Expector maps generally provide a good representation of the occurrence of the organic matter classes. The overall accuracy of prediction is high (83 percent), although for the least well predicted class the accuracy drops to 52 percent. At East Yornaning, the most probable class map provides a poorer representation of the actual clay class occurring on the ground. The overall classification accuracy is slightly over 50 percent. Although low, this later figure can be considered in the context of the accuracy of the traditional method. Ragg and Henderson (1980) (cited in Dent and Young, 1981, p. 95) showed that, using the current broadly defined soil series, map purity is usually only 50-60 percent.

12.2.2 Relative class accuracy

In both the Sterling and Yornaning examples, the relative accuracies of the classes were in the same general ratio as their prior probabilities. In the ludicrous situation of a complete absence of evidence data layers, Expector would produce class probability maps with all cells set to the value of the class prior probability. This means that a most probable class map for Sterling would show Class 1 throughout and a similar map for Yornaning would show Class 2 throughout. This, although patently wrong, would result in an apparently 'correct' classification of 72 percent at Sterling and of 50 percent at Yornaning. In both these cases, however, only sites in the predominant class would be correctly classified. All others would be wrong. The Expector results presented here have improved upon that in making a number of correct predictions for those classes that are, *a priori*, less probable. It is reasonable to speculate that with additional, powerful, evidence layers this could be greatly improved.

12.2.3 Sources of error

With a method such as Expector, lack of output precision may be accounted for in a number of ways. It may be due to inaccuracy of input data, inadequacy of the input data to describe the predicted attributes or to inaccuracy in the specification of relationships between input and output layers (ie imperfect knowledge).

If we consider the examples from Chapter 10, there is a disparity between the relatively high overall accuracy of the Sterling result (83 percent) and the less conclusive figures from Yornaning (52 percent). This must be viewed in the light of the differing densities of knowledge base. In the Sterling example, the knowledge base was provided by a regular grid of 75 observations over an area of 5.4ha. - one observation for every $720m^2$ At Yornaning, the dataset used to seed the knowledge base was based on samples taken at an average density of one observation for each $690,000m^2$, but not on a regular grid The Yornaning knowledge base was, however, augmented by the observations and experience of a soil surveyor.

This difference in density of the knowledge base also contributes to the input data accuracy since it is used to determine the prior probability of the hypothesis classes. Those for Sterling were based on a potentially more representative sample than those for Yornaning. The specification of the relationships was also based on that rich dataset.

The disparity in accuracy also suggests that the evidence layers used at Sterling were better able to predict the mapped attribute than those used at Yornaning. Both the choice of evidence and the way in which the interactions of that evidence are considered will depend largely on the experience of the individuals involved. The Expector method differs from that currently used by land resource professionals and, in order to gain maximum benefit from it, they must develop experience in its operation.

12.2.4 Opportunities for refinement of knowledge base

Some of the disparity in accuracy referred to above can be overcome by refinement of the knowledge base. If a first pass run of the Expector method, using available data, fails to produce a satisfactory map of the attribute required, then two principal

options are open to the analyst. Either the relationships between the evidence and the hypothesis attribute must be modified in the light of experience or new evidence data (which has the ability to make the necessary discriminations) must be recruited. The required course of action may be indicted by the output from the first pass.

On reviewing the results of an Expector run, the analyst may discover that the probabilities they have set do not adequately express the model they believed they were expressing. If that is the case, it is necessary to rethink the logic behind the model. This may simply require the adjustment of joint probabilities in order to change the weighting in favour of a particular attribute class or it may require a redefinition of the model structure. This may involve the combining of data layers to provide a composite layer, such as a terrain index, which better expresses the fundamental relationships.

In its current form Expector records the rules relating evidence and hypothesis solely as numerical tables of probability. It is the responsibility of the user to record the reasoning behind the setting of those probabilities. It would be possible for future versions of the software to incorporate a facility which recorded rules in plain language.

It is possible to use Expector to carry out a preliminary combination phase. However, this introduces problems associated with the setting of prior probabilities for concepts which cannot readily be measured. It was precisely this difficulty which lead to the adoption of the flat schema structure used by Expector. It is, therefore, preferable to use other data combination methods to generate indices, which are then used as input variables for Expector.

In the example from near Bundaberg, described in Chapter 11, there was a large area for which the posterior probability of two classes was similar, both being just under 0.5. Clearly, the third class was not likely to occur in that area, but the evidence used so far was unable to determine which of those two classes was predominant. In a case such as this the analyst/surveyor must consider whether there is another evidence data set which can act as a tie-breaker.

12.3 Fieldwork and sampling strategies

The density and pattern of sampling used in a survey have a bearing on its accuracy, both using the traditional method and using Expector. Given the number of ways in which it can impact on the results of the survey, it is worth considering if modification needs to be made to sampling techniques. With the traditional method, the choice of sampling strategy depends on a number of factors. Among them are considerations of the expected variability of the resource being mapped, accessibility, and cost. Broadly speaking, two methods are used: grid survey and free survey.

With grid survey, sampling proceeds on a regular Cartesian grid whose spacing may be determined by the intended scale of publication of the map. Free survey allows the collection of samples at points identified by the surveyor either in the field or from air-photos. These are often in areas where there is some doubt or where a particular boundary requires investigation. They may also be taken to establish the characteristics of a particular distinctive zone. It is not uncommon for such sampling to be carried out at a nominal spatial density similar to that used for grid sampling. It is also common for sampling in free survey to run close to roads and other accessways. This provides an economic benefit in that sampling proceeds faster and also enables sample points to be readily located on base maps simply by measuring, often with a vehicle trip counter, the distance along linear features. With the advent of affordable lightweight differential GPS equipment, the second of these reasons is now less important,

Whilst both free survey and grid sampling may ultimately gather the same number of points in a study area, the different collection patterns have an effect on the utility of the sampling within the Expector method. Sampling is used in Expector to seed joint probability distributions and to set prior probabilities.

12.3.1 The effect on prior probabilities

The effect of sample bias on prior probabilities can be illustrated with reference to the Yornaning example. We will consider a hypothetical attribute with three states which we wish to map using Expector and assume that we have a completely accurate map of its spatial distribution. This example has been synthesised from the

geology map. This map, which in the normal course of events it would be impossible to produce, purports to represent the true state of an attribute. Figure 12.1 a) shows the actual site sample used in a survey of this catchment. They take toe form of nested transects along roads and tracks. The points are overlaid on a map of this hypothetical land attribute with three classes. Figure 12.1 b) shows a regular grid of points overlaid on the same map.

The actual sample points used in the survey are in a nested scheme along transects which are governed by the road network in the area. There is, on average, one sample point for every 0.69km^2. The regular grid has been generated with a cell size of 830m giving the same overall density of points. Table 12.1 shows the proportion of each of the three classes, achieved by a direct cell count and from each of the two sampling schemes.

Class	Method of estimation		
	Direct measure	Transect samples	Regular samples
1	0.299	0.212	0.316
2	0.477	0.651	0.449
3	0.224	0.132	0.235

Table 12.1 Prior probabilities for hypothetical land attribute

The direct cell count method would not be available if this truly were an attribute being mapping with Expector, but is provided here as a measure of the 'true' prior probability. The best estimate is provided by regular grid sampling. Exactly the same number of points are used in both schemes so laboratory analysis costs are identical. As noted above, using modern positioning techniques the samples can be located readily, however, some areas may not be readily accessible because they are remote from the road network. The surveyor must make a trade-off between accuracy and economics.

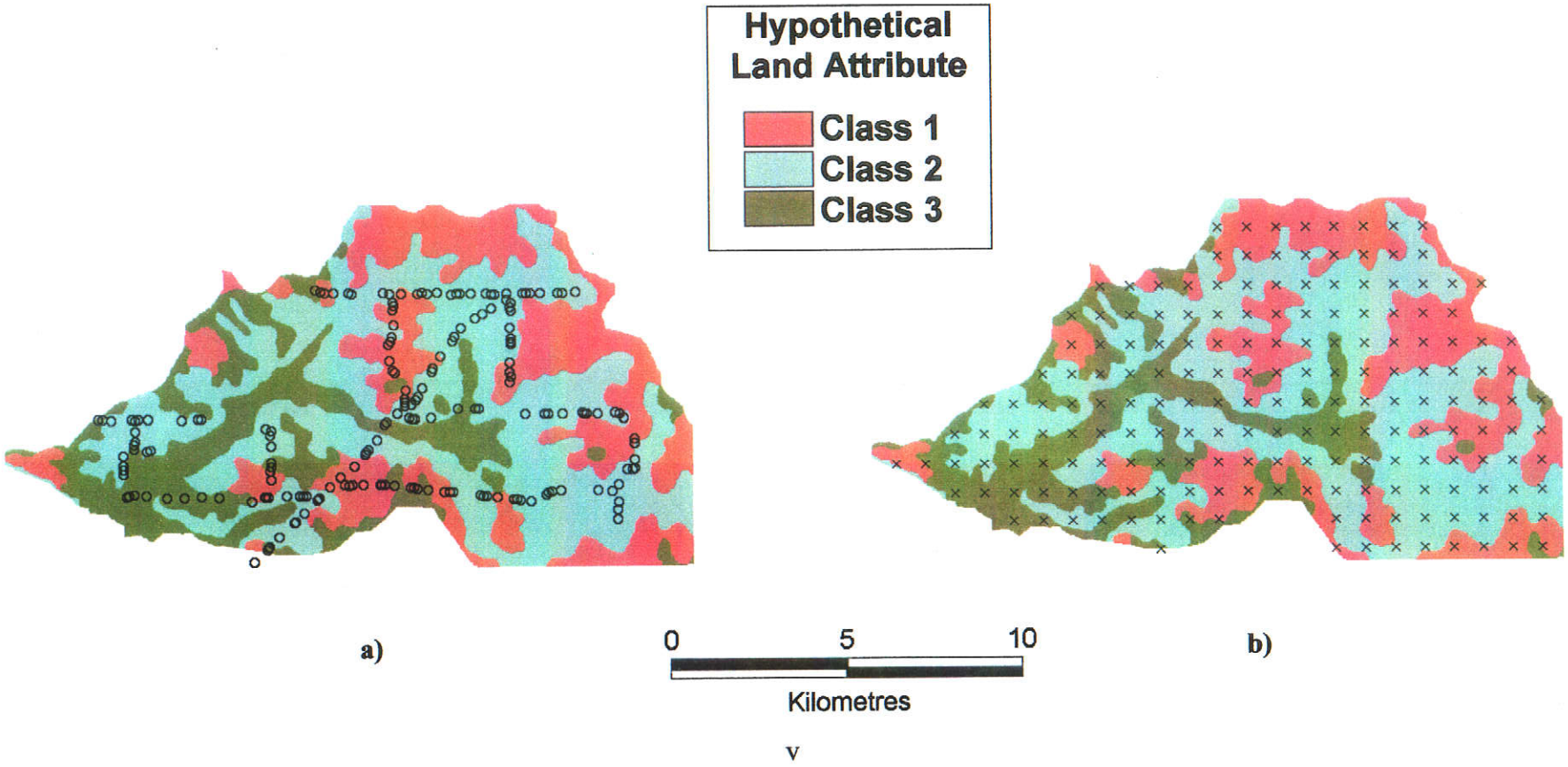

Figure 12.1 Sampling schemes at Yornaning, overlaid on a hypothetical land attribute
a) Transects along roads, b) Regular grid

12.3.2 The effect on joint probability estimates

We now turn to sampling used to determining the joint probability distributions between an attribute to be mapped and an evidence layers. In this case, there may be a considerable number of small areas. Figure 12.2 shows a map that indicates all nine possible combinations between an evidence layer (stream/ridge ratio) and our hypothetical land attribute. Both sampling schemes have been overlaid.

Table 12.2 show the joint probabilities derived by direct cell count, estimation from the linear transects and estimation from a regular grid. Once again the regular grid approaches the true situation the closest, although it is still not absolutely correct. It remains to the surveyor to use their skill and judgement to resolve any lingering bias. Unfortunately, they may never know if they are completely right.

Land attribute class	Stream-ridge class	Method of estimation		
		Direct measure	Linear transect	Regular grid
1	1	0.08	0.06	0.11
1	2	0.28	0.42	0.32
1	3	0.12	0.04	0.14
2	1	0.10	0.12	0.10
2	2	0.13	0.16	0.07
2	3	0.04	0.03	0.04
3	1	0.12	0.04	0.11
3	2	0.07	0.07	0.07
3	3	0.06	0.05	0.04

Table 12.2 Joint probabilities for land attribute and stream/ridge ratio

Class combinations resulting in small areas which are less than the nominal area assigned to each sample point are at risk of not being represented in a cross tabulation based on a sampling scheme. This is particularly true if a class combination has a number of small occurrences scattered throughout the map area.

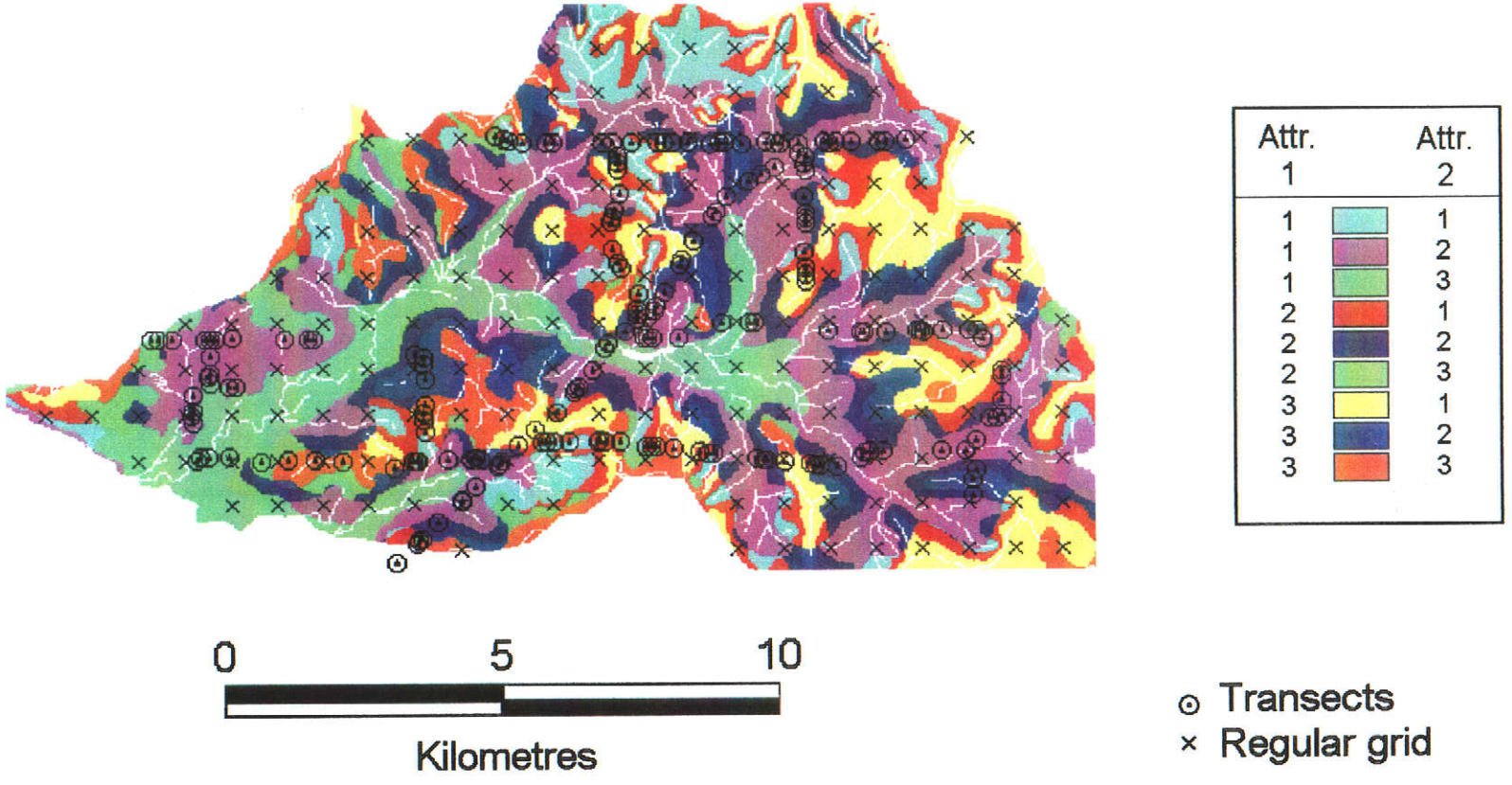

Figure 12.2 Sampling schemes laid over two coincident attributes

12.3.3 Sampling for validation

The following discussion applies not only to sampling for validation of a most probable class map but also to sampling for determination of the purity of input data layers. As a general principle, the sampling strategy should be arranged so that the number of points reasonably represents the areal extent of each class. If the sampling is designed to determine prior probabilities, and there is no prior knowledge of the class boundaries of the attribute being mapped, a regular grid sample scheme would be expected to give the best results.

In any sampling scheme, it is possible that small areas will be missed with the result that some small classes may be under-represented. If the attribute to be mapped approximates to a continuous surface, the effect of this will be less than if the attribute contains a number of discontinuities.

12.4 Model construction - conditional independence

As noted above, the data layers used in an Expector model are subject to a mathematical constraint on their independence that is not present with a mental conceptual model. In the derivation of the mathematical equations used in the Expector method, an assumption is made that the individual data layers are conditionally independent. This assumption allows a considerable degree of simplification in the calculus. In practice, this assumption is at risk of being violated and it is worth considering what may be done to minimise the effects.

12.4.1 The importance of conditional independence

The conditional independence assumption is used to simplify a complex joint probability relationship into the product of a number of simpler relationships. The assumption from Equation 8.8 is restated here for the case of two pieces of evidence.

$$\mathbf{P(E_1, E_2 \mid H_1) = P(E_1 \mid H) \,.\, P(E_2 \mid H)} \qquad (12.1)$$

In order to combine **N** data sets, values are required for $\mathbf{P(E_1,E_2,.....,E_N|H)}$, - the conditional distribution of all datasets taken together given the hypothesis. That is the conditional probability for all possible combinations of evidence states for all possible hypothesis states.

The size of this distribution is a function of the number of evidence layer, the number of states in the layers, and the number of hypothesis states. For most practical purposes, this is a multi-dimensional distribution for which it is impossible to specify all members. This is particularly the case where the probability distributions are being specified on the basis of expert knowledge.

In a case with five evidence layers, each of only two states, and a hypothesis with two states, the full distribution contains 480 terms which the user would have to estimate. Even if guided by sample data this still becomes an onerous task. Under the assumption of conditional independence this simplifies to five, separate, joint distributions each of only four terms - an altogether more tractable situation.

12.4.2 Conditional independence in the context of Expector

A mapping professional using the method would do well to ensure that they are clear within their own mind as to the relationships between the various evidence layers. It is the submission of the author that violation of *conditional* independence in the true sense is perhaps secondary to violation of an assumption of *functional* independence.

Two data sets may have a degree of dependence on each other but have markedly different effects on the hypotheses attribute. For example two data layers, say slope and wetness index, derived from the same digital elevation model may exhibit a degree of dependence. However, if their relationship to the attribute being mapped is for different *causal* reasons then there are logical grounds for the inclusion of both of them in a schema. That would be the case if slope contribute the attribute due to physical effects and wetness index due to a chemical effect. However, if both contributed through a similar process, it would be better to omit one of them. In other cases, it might be preferable to combine the two datasets, using non-probabilistic methods, into a single dataset which is then independent of the others in the analysis.

12.4.3 Tests for conditional independence

Whilst common-sense is a great aid to the analyst in determining to what extent the data-sets are dependant, formal tests do exist. In essence, Expector uses either continuous data which has been categorised or data which is by its very nature

categorical. The degree of dependence between continuous datasets may be tested using simple regression methods, whilst comparisons between categorical data can be achieved in a number of ways. Bonham-Carter, (1994, p. 242) suggests methods for pairwise comparison of datasets. These include chi-squared tests, entropy measures, and the kappa index of agreement.

Of these three methods, the kappa index is the only one readily available in a commercial desktop GIS, being implemented in IDRISI (Eastman, 1997). A disadvantage of the kappa index is that it is only applicable where the two data sets have the same number of classes. Chi-squared tests have the limitation that whilst the test statistic has a value of zero when there is no association between the datasets, it has a variable upper limit. Entropy measures have the advantage that the resulting measure of dependence (the joint information uncertainty) has a lower limit of zero and an upper limit of one. This renders them more accessible as a user-friendly test for independence.

As part of the development work on Expector, an ArcView script was written which calculates various entropy measures for pairs of categorical grid datasets. This tool was used to calculate the joint information uncertainty of all pair-wise comparisons of the data used in the Yornaning example. The results are presented in Table 12.3

	Strdist	Slope	Rock	Radiom	Geol	Curv	Catpos
Strdist	1	0.17	0.02	0.02	0.11	0.01	0.02
Slope	-	1	0.03	0.02	0.14	0.01	0.04
Rock	-	-	1	0.02	0.06	0.00	0.02
Radiom	-	-	-	1	0.04	0.00	0.01
Geol	-	-	-	-	1	0.01	0.09
Curv	-	-	-	-	-	1	0.00
Catpos	-	-	-	-	-	-	1

Table 12.3 Joint information uncertainties for Yornaning data

According to this table, the only dataset pairs which exhibit any noticeable degree of correlation are slope - stream distance and slope - geology. In the case of the geology map, the valley bottoms and areas of uniform low slope are mapped as alluvium. This will naturally cause a degree of correlation between the two maps. Other geological units will be less tied to slope units.

The case of the correlation between stream distance and slope is more complex. In a landscape comprising a valley incised into a peneplain, it is reasonable to expect there to be some degree of correlation between these two datsets. The areas near the stream will be flat with slopes increasing as one moves away from the stream. However, the interfluves also exhibit large areas with low slope. It was precisely for this reason that the stream:ridge ratio dataset was added to the schema, effectively to act as a tie-breaker. Under these circumstances, disregarding the slight correlation between the two datasets would seem to be justified. Whilst the two data layers do have a degree of statistical dependence, their logical effect on the attribute being mapped is independent.

It should also be noted that simply because two spatial datasets occupy the same physical space, there will inevitably be some slight correlation between them. The data analyst or surveyor should carefully consider cases where statistical correlation is found, some may be spurious, or be present simply because both the data layers are related to a common attribute - the one being mapped.

12.5 Summary

The most noticeable effect of the adoption of the Expector method is a change in the output of the Natural Resource Mapping process. A set of GIS data layers representing probability of occurrence of individual soil properties is more flexible and dynamic than a choropleth map. Although the validation of probability maps is problematic, gains have been made in the ability to represent that state of knowledge with no apparent loss of accuracy.

Adoption of the method requires some changes to both model construction and fieldwork. These can generally be accomplished with little extra effort, although the pattern of sampling may require alteration. Experimental work indicates that grid sampling provides a better guide to the setting of probabilities.

An examination of the errors encountered in previous examples of the methods output suggests that some of them are amenable to refinement of the knowledge base used in the individual models. This can be accomplished either by adding additional

evidence layers or by a sharpening of the relationships described by probability tables.

The effect of the conditional independence constraint on model building has been discussed and a practical test outlined which enable the degree of independence between evidence data set to be determined. The application of that test to the data used in the Yornaning experiment highlighted the need for the analyst to examine the difference between statistical independence, and causal or logical independence.

Chapter 13

SUMMARY AND CONCLUSIONS

This thesis has described the development of a method by which knowledge, particularly that pertaining to natural resource mapping, may be effectively represented in a Geographic Information System framework. Software has been written to demonstrate the concepts involved and has been tested on a number of datasets. In general, the resulting maps present fundamental attributes more flexibly than traditional methods and are a good representation of the mental models of the experts whose knowledge they encapsulate.

13.1 The need for knowledge representation in GIS

There is an increasing demand for natural resource information, imposed by both an increasing population and a growing awareness of the fragility of the natural resource base. This demand is not so much for *more* information, but for *better*, *'smarter,'* and more *flexible* information.

Traditional natural resource and soil mapping methods have reached a high level of sophistication using conceptual and statistical models to represent complex landscapes. Both types of models have drawbacks but conceptual models, although limited by virtue of being abstractions, are both flexible and appealing to the natural resource surveyor. They are capable of representing complex entities and relationships over large areas and of incorporating sparse or uncertain data.

Unfortunately, information is lost when those models are represented as traditional choropleth maps. Much of this lost information concerns fundamental soil attributes, matters of interest to the plants that form the basis of the agricultural system. It is suggested that a multi-layer representation of the likelihood of occurrence of particular soil attributes may preserve that information. The use of GIS offers the opportunity to represent spatially variable soil attributes in a flexible fashion. It is suggested that linking a knowledge representation tool to a GIS provides an increase in the utility of soil information.

13.2 Expert systems, knowledge and GIS

Knowledge is represented in the human brain in informal, loosely defined, and 'fuzzy' ways. In order to mimic this in a computer, a means is required to formalise knowledge in a way that enables it to be manipulated mathematically. Expert systems, a branch of Artificial Intelligence (AI), seek to capture the reasoning behind decisions made by experts in particular domains of knowledge

Expert system methods typically use either logic or probability as representational frameworks. Schemes using probability are more intuitively suited to representing the kind of 'imprecise' and flexible knowledge inherent in conceptual models of landscapes. Bayes' rule, a method of probabilistic calculation defined some 240 years ago, has recently found favour as a means of manipulating and combining representations of knowledge.

A number of expert systems which use probability as a means of knowledge representation have been designed for use in the field of medical diagnosis. There are parallels between medical diagnosis and natural resource mapping – both use symptoms and surface expressions as evidence from which to draw inferences.

Not all such systems are in the field of medicine. PROSPECTOR, an expert mineral prospecting consultant, was an early venture of such systems into mapping. Unfortunately, there are a number of shortcomings in its representational calculus. These may largely be overcome by adhering to the general principles of a class probabilistic of systems known as Bayesian networks, or Causal Probabilistic Networks (CPN). Again, there are examples of such systems in the field of medical diagnosis and some have been examined and described in earlier chapters.

Whilst such tools may be useful for representing knowledge, a spatial component is required in order to apply them to natural resources mapping. That is provided by GIS which, as a general technology, draws on long established concepts of cartographic modelling. Many of these have their roots in land resource assessment. The technology of GIS is diverse, with a number of data representation schemes and proprietary systems in existence. Central to all of them, however, is a system for

storing, analysing and integrating spatial data; then displaying it as enhanced information products.

13.3 Using knowledge and GIS to quantify soil mapping

An examination of the soil mapping method shows it to comprise two principal components; model building and data combination. It is suggested that it can be quantified and formalised by combining an expert system and a GIS. A custom written expert system tool can handle model building, with a GIS being responsible for data preparation and combination. The basic structure of a causal probability networks provide a mechanism to manipulate the multiple evidence threads that either support or contradict propositions about the existence of particular attributes at various levels or states. Algorithms and methods for obtaining the parameters and probability distributions used by a CPN have been described and a practical method of handling inexact data devised.

The process of taking a map as a piece of evidence begins to bring the expert system to life as a mapping tool and requires that it have effective interfaces to GIS. Through those interfaces, access is provided to tools for aggregating the evidence provided by several inputs. The algorithm described for that process departs from standard CPN calculus in that it uses input distributions, provided by the expert system, which are intuitively more meaningful to a natural resource surveyor.

13.4 A software implementation

The expert system tool, named Expector, was written in Microsoft Visual Basic for use under a 32 bit PC operating system. This language and operating system combination was chosen after consideration of patterns of GIS usage within the agencies charged with mapping the soils of Australia. Interfacing routines to GIS have been constructed in the native scripting languages of those systems. The systems covered here are ARC/INFO and ArcView. An interface to Microstation GIS has been constructed by others. At the most basic level, the general algorithms used for interfacing and data combination are assembled from 'primitives' which are available in any GIS which possesses map-algebra capabilities. They are, therefore, capable of being implemented in almost any GIS.

Expector provides user-friendly forms-type tools for creating and editing a knowledge base, and for the combination of the knowledge associated with individual data layers. GIS specific display utilities are also available for examining and presenting the outputs.

Expector may be regarded as an overall process that maps the probability of occurrence of a number of states of some attribute, known as the hypothesis. The mapping is based on evidence provided by extensive datasets covering the area of interest. The knowledge used in the process may be drawn from site sample data, from an expert or from a combination of both. Expector enables the knowledge base to be seeded with values derived from site sample data. The expert can then edit or even discard those values according to their opinion of the richness of the sample data and of any bias in its derivation.

13.5 Applications and demonstrations of the method

Two studies were carried out during the development of the Expector method. One (essentially as a proof of concept) used a rich sample dataset from Sterling, Colorado to provide knowledge about a small area. Another, at East Yornaning, Western Australia, used soil surveyor knowledge to produce a map of soil surface texture over a moderately large catchment. The surveyor's knowledge was assisted by a relatively sparse sample dataset. The East Yornaning example achieved a level of accuracy in its representation of soil texture equivalent to that of a traditional soil map.

This was achieved using much the same evidence and thought processes as were used in the development of the traditional map. However, since they were applied using a quantitative and formalised method, the analysis is not only repeatable and transferable, but is also readily open to improvement. That improvement can come through refinement of the knowledge base in the light of the results.

Three projects were undertaken to demonstrate Expector to the natural resource mapping community of Australia. These were at Brookton in Western Australia, Bundaberg in Queensland, and at Forth in Northern Tasmania. The attributes mapped were 'modal soil type', surface clay content, and land capability class

respectively. It is interesting to note that, although Expector is designed to map individual attributes, in two of these cases the mental models of the land resource professionals recruited as experts led to traditional class based concepts.

Expector also has potential for use with other problems involving the combination of diverse evidence streams. This was demonstrated by its use to predict agricultural yield potential in the context of precision agriculture. Such information is required for the precise targeting of fertiliser rates.

The three case studies in land resource assessment were not intended to prove the accuracy of Expector but to demonstrate the method to potential users. In two cases however, the demonstration highlighted interesting operational considerations that could only have been addressed by the collection of additional data. Since there were insufficient resources to allow this, no formal evaluation of the demonstrations was carried out. A more formal assessment was made of agricultural yield prediction example, which proved capable of distinguishing between high and low yielding parts of a paddock. This analysis indicated the presence of unforeseen yield reducing effects which had not formed part of the original knowledge base.

13.6 Operational considerations

In order to investigate the effect of different fieldwork sampling strategies on the setting of probabilities, an experiment was conducted using data from East Yornaning. The results of this suggested that a regular grid sampling scheme will provide closer estimates of prior and joint probabilities than a transect scheme. It was also noted that the setting of joint probabilities from sampling is particularly difficult due to the relatively small size of joint areas. The use of knowledge to overcome such deficiencies is one of the key features of Expector.

A major operational consideration of the use of Bayesian networks such as Expector, is that they operate under an assumption that the input data layers are conditionally independent. The data used at Yornaning were tested, using entropy measures, for conditional independence and little dependence was found. The only cases where it was present to any degree were readily explained by an examination of the underlying causation. This serves to highlight the need for the analyst to have regard

for the differences between statistical independence, and causal or logical independence.

13.7 Conclusions

A method has been presented by which the knowledge inherent in natural resource mapping may be represented in a GIS framework. It is designed primarily to provide an output in the form of GIS data layers showing the probability of occurrence of various states of selected attributes. The method is available in a user-friendly software implementation for use on Personal Computers and capable of working in conjunction with popular proprietary GIS. It uses a probabilistic (Bayesian) network as its central calculus and is capable of handling data with inherent imprecision. It uses a knowledge base and extensive datasets to spatially extend that knowledge across the landscape. That knowledge base may be derived either from sampling or from expert opinion.

In developmental tests, the method proved capable of matching the ability of a traditional soil map to predict basic soil attributes. Subsequent demonstrations of the method highlighted an interesting range of operational considerations. In essence, these show that the method, being knowledge based, is only as good as the knowledge used. That knowledge must be applied to select the correct input data, and to correctly specify the relationships between those data and the attributes being mapped.

REFERENCES

Adams, J. B. (1987) Probabilistic Reasoning and Certainty Factors, in: *Rule Based Expert Systems*, Buchanan, B. G. and Shortliffe, E. H. (eds.), Adison-Wesley, Reading, Massachusetts, pp. 263-271.

Agricola, G. (1950) *De re metalica* (Translated by H.C and L.H. Hoover), Dover Publishing, New York, New York, 638 pp, (Reprint).

Agterberg, F. P. (1989) Computer Programs for Mineral Exploration, *Science*, Vol. 245, pp. 76-81.

Antenucci, J. C., Brown, K., Croswell, P. L., Kevany, M. J. and Archer, H. (1991) *Geographic Information Systems; A guide to the technology*, Van Nostrand Reinhold, New York, New York, 301 pp.

Aspinall, R. (1992) An inductive modelling procedure based on Bayes' theorem for analysis of pattern in spatial data, *International Journal of Geographical Information Systems*, Vol. 6, No. 2, pp 105-121.

Barker, V. E. and O'Connor, D. E. (1989) Expert Systems for Configuration at Digital: XCON and Beyond, *Communication of the ACM*, Vol. 32, No. 3, pp. 298-310.

Bayes, T. (1763) An Essay Towards Solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society*, Vol. 53, pp. 370-418.

Bayes, T. (1958) An Essay Towards Solving a Problem in the Doctrine of Chances, *Biometrika*, Vol. 45, pp. 296-315, (Reprint).

Becket, P. H. T. and Webster, R. (1971) Soil Variability: A Review, *Soils and Fertilisers*, Vol. 34, pp. 1-13.

Bellhouse, D. (1993) The role of Roguery in the History of probability, *Statistical Science*, Vol. 8, No. 3, pp. 410-420.

Bender, E. A. (1996) *Mathematical methods in artificial intelligence*, IEEE Computer Society Press, Los Alamitos, California, 636 pp.

Bennet, J. S. and Engelmore, R. S. (1987) Experiences using EMYCIN, in: *Rule Based Expert Systems*, Buchanan, B. G. and Shortliffe, E. H. (eds.), Addison-Wesley, Reading, Massachusetts, pp. 314-328.

Berry, J. K. (1993) Cartographic Modelling: the analytical capabilities of GIS, in: *Environmental Modelling with GIS*, Goodchild, M. F., Parkes, B. O. and Steyaert, L. T. (eds.), Oxford University Press, Oxford, U.K., pp. 58-74.

Bonham-Carter (1994) *Geographic Information Systems for Geoscientists: Modelling with GIS*, Pergamon/Elsevier, Kidlington, U.K., 398 pp.

Bonham-Carter, G. F. (1991) Integration of Geoscientific Data using GIS, in: *Geographical Information Systems; V. 1: Principles*, Maquire, D. J., Goodchild, M. F. and Rhind, D. W. (eds.), Longman, London, U.K., pp. 171-184.

Bouma, J. (1989) Using soil survey data for quantitative land evaluation, *Advances in Soil Science*, Vol. 9, pp. 177-213.

Buchanan, B. G. and Shortliffe, E. H., (eds.) (1987) *Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, Reading, Massachusetts, 748 pp.

Burrough, P. A. (1986) *Principles of GIS for land resource assessment*, Oxford University Press, Oxford, U.K.,194 pp.

Burrough, P. A. (1992) Development of intelligent geographic information systems, *International Journal of Geographic Information Systems*, Vol. 6, No. 1, pp. 1-11.

Burrough, P. A., van Gaans, P. F. M. and Hootsmans, R. (1997) Continuous classification in soil survey: spatial correlation, confusion and boundaries, *Geoderma*, Vol. 77, pp. 115-135.

Chin, R. J. (1986) Corrigin, Western Australia (Sheet SI 50-3), *1:250,000 Geological series - explanatory notes*, Geological Survey of Western Australia, Perth, Western Australia, 21 pp. plus map.

Cohen, P. R. (1985) *Heuristic Reasoning about uncertainty: An Artificial Intelligence Approach*, Pitman, London, 204 pp.

Cook, S. E., Corner, R. J., Grealish, G., Gesler, P. E. and Chartres, C. J. (1996) A Rule-based System to Map Soil Properties, *Soil Science Society of America Journal*, Vol. 60, pp. 1893-1900.

Cook, S,E. Corner, R.J, Grealish, G. and Groves, R. Use of Airborne Gamma Radiometrics for Soil Mapping, *Australian Journal of Soils Research*, 1996 Vol. 34, pp 183-194.

Cowen, D. J. (1988) GIS versus CAD versus DBMS: What are the differences?, *Photogrametric Engineering and Remote Sensing*, Vol. 56, No. 11, pp. 1551-1555.

Crisp, N. (1998) Open GIS: The key to a new generation of GIS, *Adding a Spatial Dimension to Business*, Mapping Sciences Institute of Australia National Conference, Fremantle, Western Australia, May.

Davis, J. C. (1986) *Statistics and Data Analysis in Geology*, John Wiley, New York, New York, 646 pp.

Davis, J. R. and Nanninga, P. M. (1985) GEOMYCIN: Towards a Geographic Expert System for Resource Management, *Journal of Environental Management*, Vol. 21, pp. 377-390.

Davis, R. (1987) Interactive Transfer of Expertise, in: *Rule Based Expert Systems*, Buchanan, B. G. and Shortliffe, E. H. (eds.), Addison-Wesley, Reading, Massachusetts, pp. 171-205.

Dempster, A. P. (1967) Upper and Lower Probabilities induced by a Multivalued Mapping, *Annals of Mathematical Statistics*, Vol. 38, pp. 325-339.

Dent, D. and Young, A. (1981) *Soil Survey and Land Evaluation*, George Allen and Unwin, London, U.K., 278 pp.

Duda, R. O. and Gaschnig, J. G. (1981) Knowledge-Based Expert Systems Come of Age, *BYTE*, No. 9, pp. 238-281.

Duda, R. O. and Shortliffe, E. H. (1983) Expert Systems Research, *Science*, Vol. 220, No. 4594, pp. 261-268.

Duda, R. O., Hart, P. E., Barrett, P., Gaschnig, J. G., Klige, K., Reboh, R. and Slocum, J. (1978) Development of the Prospector Consultation System for Mineral Exploration, *Final report for SRI Projects 5821 and 6415*, Artificial Inteligence Center, SRI International, Stanford, California, 193 pp.

Dunn, M. and Hickey, R. (1998) The effect of slope algorithms on slope estimates within a GIS, *Cartography*, Vol. 27, No. 1, pp. 9-15.

Eastman, J. R. (1997) IDRISI for Windows, Clark Labs for Cartographic technology and Geographic Analysis, Worcester, Massachusetts, (computer program).

ESRI (1997) ARC/INFO, Environmental Systems Research Institute, Redlands, California, (computer program).

FAO (1979) *Soil survey investigations for irrigation.* FAO Soils Bulletins 42. FAO, Rome, Italy, pp198.

FAO (1999) *Down-to-Earth for food security*, FAO Soil Resources Management and Conservation Service Fact sheet No. 1, FAO, Rome, Italy, pp 2.

Fedra, K. (1993) GIS and environmental Modelling, in: *Environmental Modelling with GIS*, Goodchild, M. F., Parkes, B. O. and Steyaert, L. T. (eds.), Oxford University Press, Oxford, U.K., pp. 35-50.

Ferrier, G. and Wadge, G. (1997) An integrated GIS and knowledge -based system as an aid for the geological analysis of sedimentary basins, *International Journal of Geographic Information Science*, Vol. 11, No. 3, pp. 281-297.

Flowerdew, R. (1991) Spatial Data Integration, in: *Geographical Information Systems*, Maguire, D. J., Goodchild, M. F. and Rhind, D. W. (eds.), Longman, London, U.K., pp. 375-87.

Fox, D.(1996) *Personal Communication*, Biometrics Unit, CSIRO Centre for Mediterranean Agricultural Research, Perth, Western Australia.

Gallant, J. C. and Wilson, J. P. (1996) TAPES-G: A grid-based terrain analysis program for the environmental sciences, *Computers and Geosciences*, Vol. 22, No. 7, pp. 713-722.

Garcia, O, N., and Chien, Y-T., (eds.) (1992) *Knowledge-Based Systems: Fundamentals and Tools*, IEEE Computer Society Press, Los Alamitos, California, 495 pp.

Gessler, P. E., Moore, I. D., McKenzie, N. J. and Ryan, P. J. (1995) Soil-landscape modelling and the spatial prediction of soil attributes, *International Journal of Geographic Information Systems*, Vol. 9, pp. 421-432.

Goodchild, M. F. (1991) The technological setting of GIS, in: *Geographical Information Systems*, Maquire, D. J., Goodchild, M. F. and Rhind, D. W. (eds.), Longman, London, U.K., pp. 45-54.

Goodchild, M. F. and Gopal, S. (1989) *The Accuracy of Spatial Databases*, Taylor and Francis, London, U.K.,290 pp.

Grealish, G. G., Cook, S. E. and Corner.R.J. (1994) Testing the ability of existing soil maps to predict soil properties, *Proceedings of Soils '94*, Busselton, Western Australia, Australian Society of Soil Science, pp. 171-175.

Hacking, I. (1975) *The Emergence of Probability*, Cambridge University Press, London, U.K., 209 pp.

Hacking, I. M. (1965) *Logic of Statistical Inference*, Cambridge University Press, Cambridge, U.K. 232 pp.

Hart, P. E., Duda, R. O. and Einaudi, M. T. (1978) PROSPECTOR - A Computer-Based Consultation System for Mineral Exploration, *Mathematical Geology*, Vol. 10, No. 5, pp. 589-610

Heckerman, D. (1997) Bayesian Networks for Data Mining, *Data Mining and Knowledge Discovery*, Vol. 1, pp. 79-119.

Herring (1987) TIGRIS: Topologically integrated geographic resource information systems, *Proceedings of Auto Carto 8*, Falls Church, Virginia, pp. 282-91.

Hewitt, A., E. (1993) Predictive modelling in soil survey, *Soils and fertilisers*, Vol. 56, pp. 305-314.

Isbell, R. F. (1996) *The Australian soil classification*, Australian soil and land survey handbook series, V. 4, CSIRO Publishing, Melbourne, Victoria, 143 pp.

Jankowski, P. (1995) Integrating Geographical Information Systems and Multiple Criteria Decision-making methods, *International Journal of Geographical Information Systems*, Vol. 9, No. 3, pp. 251-273.

Jensen, F. V. (1996) *An Introduction to Bayesian Networks*, University College London Press, London, U.K., 178 pp.

Katz, S. S. (1991) Emulating the Prospector Expert System with a Raster GIS, *Computers and Geoscience*, Vol. 17, No. 7, pp. 1033-1050.

Maguire, D. J. (1991) An overview and definition of GIS, in: *Geographical Information Systems; V. 1: Principles*, Maquire, D. J., Goodchild, M. F. and Rhind, D. W. (eds.), Longman, London, U.K., pp. 9-20.

McArthur, W. M. (1991) *Reference soils of south-western Australia.* Western Australian Department of Agriculture, Perth, Western Australia,265 pp.

McArthur, W. M., Churchward, H. M. and Hick, P. T. (1977) *Landforms and soils of the Murray River catchment of Western Australia*, CSIRO Division of Land Resource Management, Adelaide, South Australia, 23pp. plus maps.

McHarg, I. L. (1969) *Design with Nature*, Natural History Press, Garden City, New York, 197 pp.

Montgomery, D. C. and Runger.G.C. (1994) *Applied Statistics and Probability for Engineers*, John Wiley, New York, New York, 895 pp.

Moon, W. O. (1990) Integration of Geophysical and Geological Data Using Evidential Belief Function, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 28, No. 4, pp. 711-720.

Moore, I. D., Gessler, P. E., Nielsen, G. A. and Peterson, G. A. (1993) Soil Attribute Prediction Using Terrain Analysis, *Soil Science Society of America Journal*, Vol. 57, pp. 443-452.

Moore, I. D., Grayson, R. B. and Ladson, A. R. (1991) Digital terrain modelling: A review of hydrological, geomorphological and biological applications, *Hydrological Processes*, Vol. 5, pp. 3-30.

Mulcahy, M. (1973) Landforms and soils of south-western Australia, *Journal of the Royal Society of Western Australia*, Vol. 56, pp. 16-22.

Northcote, K. H. (1971) *A Factual Key for the Recognition of Australian Soils*, Rellim Technical Publications, Adelaide, South Australia,123 pp.

Odeh, I. O. A., McBratney, A. B. and Chittleborough, D. J. (1995) Further results on prediction of soil properties from terrain attributes: heterotopic co-kriging and regression-kriging, *Geoderma*, Vol. 64, No. , pp 215-226.

Paracelsus (1967) *The Hermitic and Alchemical writings of A.P.T.Bombast, called Paracelcus the Great.* , University Books, New Hyde Park, New York.

Pearl, J. (1986) Fusion, Propagation, and Structuring in Belief Networks, *Artificial Inteligence*, Vol. 29, pp. 241-288.

Penrose, R. (1989) *The Emperors New Mind*, Oxford University Press, Oxford, U.K., 602 pp.

Ragg, J. M. and Henderson, R. (1980) A reappraisal of soil mapping in an area of southern Scotland, *Journal of Soil Science*, Vol. 31, pp 559-580.

Rhind, D. W., Green, N. P. A., Mounsey, H. M. and Wiggins, J. C. (1984) The Integration of Geographical Data, *Proceedings of Austra-Carto Perth*, Perth, Western Australia, Australian Cartographic Association, pp. 273-93.

Rhodes, P. C. and Garside, G. R. (1991) Reappraisal of the use of conditional probability in early expert systems, *Knowledge Based Systems*, Vol. 4, No. 2, pp. 67-74.

Rossiter, D. G. (1990) ALES: A Framework for Land Evaluation Using a Microcomputer, *Soil Use and Management*, Vol. 6, pp. 7-20.

Rossiter, D. G. (1996) A Theoretical Framework for Land Evaluation, *Geoderma*, Vol. 72, pp. 165-190.

Rossiter, D. G. (1998) *The Automated Land Evaluation System ALES*, http://wwwscas.cit.cornell.edu/landeval/ales/ales.htm.

Shafer, G. (1976) *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, New Jersey, 297 pp.

Shafer, G. and Pearl, J., (eds.) (1990) *Readings in Uncertain Reasoning*, Morgan Kaufmann, San Mateo, California, 768 pp.

Shields, P. G., Smith, C. D. and McDonald, W. S. (1996) *Agricultural Land Evaluation in Australia*: A Review, ACLEP Report, Australian Collaborative Land Evaluation Program, Canberra, Australian Capital Territory, 152 pp.

Shortliffe, E. H. (1974) *Mycin: A rule based computer program for advising physicians regarding antimicrobial therapy selection*, Stanford University Computer Science Department, Stanford, California, 395 pp.

Shortliffe, E. H. and Buchanan, B. G. (1987) A Model of Inexact Reasoning in Medicine, in: *Rule Based Expert Systems*, Buchanan, B. G. and Shortliffe, E. H. (eds.), Addison-Wesley, Reading, Massachusetts, pp. 233-263.

Skidmore, A. K. (1989) An expert system classifies Eucalypt forest types using Thematic Mapper data and a Digital Terrain Model, *Photogrametric Engineering and Remote Sensing*, Vol. 55, No. 10, pp. 1449-1464.

Skidmore, A. K., Ryan, P. J., Dawes, W., Short, D. and O'Loughlin, E. (1991) Use of an expert system to map forest soils from a geographical information system, *International Journal of Geographical Information Systems*, Vol. 5, No. 4, pp. 431-445.

Smyth, C. S. (1998) A representation framework for Geographic Modelling, in: *Spatial and temporal reasoning in Geographic Information Systems*, Egenhofer, M. J. and Golledge, R. G. (eds.), Oxford University Press, Oxford, U.K., pp. 191-213.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. and Cowell, R. G. (1993) Bayesian Analysis in Expert Systems, *Statistical Science*, Vol. 8, No. 3, pp. 219-283.

Soil Survey Staff, (1993) *Soil Survey Manual*. United States Department of Agriculture, Washington D.C., 437 pp.

Stassopoulou, A., Petrou, M. and Kittler, J. (1998) Application of a Bayesian network in a GIS based decision making system, *International Journal of Geographic Information Science*, Vol. 12, No. 1, pp. 23-45.

Strahler, A. N. (1952) Dynamic Basis of Geomorphology, *Geological Society of America Bulletin*, Vol. 63, pp. 923-938.

Tomlin, C. D. (1990) *Geographical Information Systems and Cartographic Modelling*, Prentice Hall, Engelwood Cliffs, New Jersey, 249 pp.

van Lanen, H. A. J. and Woperis, F. A. (1992) Computer-captured expert knowledge to evaluate possibilities for injection of slurry from animal manure in the Netherlands, *Geoderma*, Vol. 54, pp. 107-124.

van Melle, W., Shortliffe, E. H. and Buchanan, B. G. (1987) EMYCIN: A Knowledge Engineer's Tool for Constructing Rule Based Expert Systems, in: *Rule Based Expert Systems*, Buchanan, B. G. and Shortliffe, E. H. (eds.), Addison-Wesley, Reading, Massachusetts, pp. 302-313.

Verboom, W. H., Galloway, P. D., Corner, R. J. and Moore, G. A. (1997) Landscape processes drive GIS mapping, *Soils'97*, Australian Society of Soil Science Inc. (WA Branch), September, pp. 209-210.

von Mises, R. (1964) *Mathematical Theory of Probability and Statistics*, Academic Press, New York, New York, 694 pp.

APPENDIX A

DATA USED IN DEMONSTRATIONS.

Conditional Probability Printout
13/02/99 11:22:42
Evidence : catpos with 6 classes
Prior probabilities

PE 1	PE 2	PE 3	PE 4	PE 5	PE 6
0.64	0.11	0.11	0.06	0.07	0.00

Conditional Probabilities

First	Second	Third	Fourth	Fifth	Sixth
0.900	0.050	0.000	0.000	0.000	0.000
0.100	0.900	0.050	0.000	0.000	0.000
0.000	0.050	0.900	0.050	0.000	0.000
0.000	0.000	0.050	0.900	0.090	0.000
0.000	0.000	0.000	0.050	0.900	0.000
0.000	0.000	0.000	0.000	0.010	1.000

Joint Probability Printout
13/02/99 11:51:05
Evidence : catpos with 6 classes
Hypothesis : btx1 with 3 classes
Prior probabilities

PH 1	PH 2	PH 3
0.14	0.51	0.35

PE 1	PE 2	PE 3	PE 4	PE 5	PE 6
0.581	0.171	0.107	0.070	0.069	0.001

Joint Probabilities

First	Second	Third	Fourth	Fifth	Sixth
0.070	0.031	0.019	0.012	0.010	0.000
0.326	0.063	0.049	0.035	0.029	0.000
0.186	0.077	0.039	0.023	0.030	0.000

Joint Probabilities as percent of Evidence

class 1	class 2	class 3	class 4	class 5	class 6
12%	18%	18%	17%	15%	14%
56%	37%	46%	50%	42%	51%
32%	45%	36%	33%	43%	35%

Data Panel 10.1 East Yornaning - Catchment position layer

```
Conditional Probability Printout
13/02/99      11:22:33
Evidence  :   curv with  4  classes
Prior probabilities
PE 1   PE 2    PE 3    PE 4
0.27  0.26   0.35   0.12

Conditional Probabilities
PL+/PR+  PL+/PR-   PL-/PR+   PL-/PR-
0.700    0.100     0.100     0.100
0.100    0.700     0.100     0.100
0.100    0.100     0.700     0.100
0.100    0.100     0.100     0.700
Joint Probability Printout
13/02/99      11:49:29
Evidence  :   curv with  4  classes
Hypothesis :  btx1 with  3  classes
Prior probabilities
PH 1  PH 2   PH 3
 0.14  0.51   0.35

PE 1  PE 2   PE 3   PE 4
0.261 0.256  0.308  0.175

Joint Probabilities

PL+/PR+  PL+/PR-   PL-/PR+   PL-/PR-
0.029    0.049     0.046     0.014
0.136    0.100     0.176     0.101
0.097    0.108     0.086     0.059


Joint Probabilities as percent of Evidence

class 1  class 2   class 3   class 4
11%      19%       15%       8%
52%      39%       57%       58%
37%      42%       28%       34%
```

Data Panel 10.2 East Yornaning - Curvature layer

Conditional Probability Printout
13/02/99 11:31:18
Evidence : geol with 7 classes
Prior probabilities

PE 1	PE 2	PE 3	PE 4	PE 5	PE 6	PE 7
0.07	0.02	0.21	0.00	0.11	0.48	0.11

Conditional Probabilities

Age	Agm	Agv	Cza	Czl	Qa	Qc
0.700	0.150	0.200	0.000	0.050	0.000	0.000
0.100	0.700	0.100	0.000	0.050	0.000	0.000
0.200	0.150	0.700	0.000	0.050	0.000	0.000
0.000	0.000	0.000	0.650	0.000	0.200	0.150
0.000	0.000	0.000	0.000	0.850	0.000	0.050
0.000	0.000	0.000	0.350	0.000	0.750	0.000
0.000	0.000	0.000	0.000	0.000	0.050	0.800

Joint Probability Printout
13/02/99 12:25:11
Evidence : geol with 7 classes
Hypothesis : btxl with 3 classes
Prior probabilities

PH 1	PH 2	PH 3
0.14	0.51	0.35

PE 1	PE 2	PE 3	PE 4	PE 5	PE 6	PE 7
0.099	0.048	0.169	0.113	0.102	0.357	0.112

Joint Probabilities

Age	Agm	Agv	Cza	Czl	Qa	Qc
0.003	0.001	0.007	0.014	0.064	0.046	0.009
0.055	0.023	0.061	0.059	0.027	0.222	0.062
0.042	0.023	0.102	0.041	0.010	0.089	0.041

Joint Probabilities as percent of Evidence

class 1	class 2	class 3	class 4	class 5	class 6	class 7
3%	2%	4%	12%	63%	13%	8%
55%	49%	36%	52%	27%	62%	55%
42%	49%	60%	36%	10%	25%	37%

Data Panel 10.3 East Yornaning - Geology layer

```
Conditional Probability Printout
13/02/99      11:33:11
Evidence  :   radiom with  4  classes
Prior probabilities
PE 1   PE 2    PE 3    PE 4
0.06  0.02   0.17   0.75

Conditional Probabilities
Granitic Sandplain Colluvial Alluvial
0.800    0.050     0.050     0.050
0.050    0.800     0.050     0.050
0.050    0.050     0.800     0.050
0.050    0.050     0.050     0.800
Joint Probability Printout
13/02/99      12:02:11
Evidence  :   radiom with  4  classes
Hypothesis :  btx1 with  3  classes
Prior probabilities
PH 1  PH 2   PH 3
 0.14  0.51   0.35

PE 1  PE 2   PE 3   PE 4
0.092 0.067  0.177  0.613

Joint Probabilities


Granitic Sandplain Colluvial Alluvial
0.031    0.067     0.032     0.055
0.046    0.000     0.051     0.356
0.015    0.000     0.094     0.202



Joint Probabilities as percent of Evidence


class 1  class 2   class 3   class 4
34%      100%      18%       9%
50%      %         29%       58%
16%      %         53%       33%
```

Data Panel 10.4 East Yornaning - Radiometrics layer

```
Conditional Probability Printout
13/02/99      11:34:08
Evidence  :   rock with  5  classes
Prior probabilities
PE 1   PE 2    PE 3    PE 4    PE 5
0.01  0.02   0.03   0.04   0.89

Conditional Probabilities
On       0-50m     50-100m   100-150m  >150m
1.000    0.000     0.000     0.000     0.000
0.000    1.000     0.000     0.000     0.000
0.000    0.000     1.000     0.000     0.000
0.000    0.000     0.000     1.000     0.000
0.000    0.000     0.000     0.000     1.000
Joint Probability Printout
13/02/99      12:05:17
Evidence  :   rock with  5  classes
Hypothesis :  btx1 with  3  classes
Prior probabilities
PH 1  PH 2   PH 3
 0.14  0.51   0.35

PE 1  PE 2   PE 3   PE 4   PE 5
0.012 0.025  0.034  0.041  0.889

Joint Probabilities

On       0-50m     50-100m   100-150m  >150m
0.012    0.016     0.011     0.010     0.089
0.000    0.008     0.023     0.020     0.462
0.000    0.000     0.000     0.010     0.338


Joint Probabilities as percent of Evidence

class 1  class 2   class 3   class 4   class 5
100%     67%       33%       25%       10%
%        33%       67%       50%       52%
%        %         %         25%       38%
```

Data Panel 10.5 East Yornaning - Rock layer

Conditional Probability Printout
13/02/99 11:36:00
Evidence : slope with 9 classes
Prior probabilities

PE 1	PE 2	PE 3	PE 4	PE 5	PE 6	PE 7	PE 8	PE 9
0.12	0.20	0.22	0.20	0.11	0.06	0.03	0.02	0.04

Conditional Probabilities

0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	>8
0.900	0.050	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.100	0.900	0.050	0.000	0.000	0.000	0.000	0.000	0.000
0.000	0.050	0.950	0.050	0.000	0.000	0.000	0.000	0.000
0.000	0.000	0.050	0.900	0.050	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.050	0.900	0.050	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.050	0.900	0.050	0.000	0.000
0.000	0.000	0.000	0.000	0.000	0.050	0.900	0.050	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.050	0.900	0.100
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.050	0.900

Joint Probability Printout
13/02/99 12:30:30
Evidence : slope with 9 classes
Hypothesis : btx1 with 3 classes
Prior probabilities

PH 1	PH 2	PH 3
0.14	0.51	0.35

PE 1	PE 2	PE 3	PE 4	PE 5	PE 6	PE 7	PE 8	PE 9
0.119	0.201	0.229	0.195	0.116	0.064	0.033	0.022	0.033

Joint Probabilities

0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	>8
0.020	0.040	0.032	0.020	0.010	0.009	0.000	0.000	0.008
0.069	0.111	0.112	0.107	0.043	0.022	0.022	0.011	0.016
0.030	0.050	0.085	0.068	0.063	0.033	0.011	0.011	0.008

Joint Probabilities as percent of Evidence

class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8	class 9
17%	20%	14%	10%	9%	14%	%	%	25%
58%	55%	49%	55%	37%	34%	67%	50%	50%
25%	25%	37%	35%	54%	52%	33%	50%	25%

Data Panel 10.6 East Yornaning - Slope layer

```
Conditional Probability Printout
13/02/99      11:37:19
Evidence  :   strdist with  3  classes
Prior probabilities
PE 1   PE 2    PE 3
0.48  0.28   0.25

Conditional Probabilities
Lower    Middle    Upper
0.950    0.050     0.000
0.050    0.900     0.050
0.000    0.050     0.950
Joint Probability Printout
13/02/99      12:31:26
Evidence  :   strdist with  3  classes
Hypothesis :  btx1 with  3  classes
Prior probabilities
PH 1  PH 2   PH 3
 0.14  0.51   0.35

PE 1  PE 2   PE 3
0.468 0.285  0.247

Joint Probabilities

Lower    Middle    Upper
0.019    0.031     0.089
0.248    0.154     0.109
0.201    0.100     0.049


Joint Probabilities as percent of Evidence

class 1  class 2   class 3
4%       11%       36%
53%      54%       44%
43%      35%       20%
```

Data Panel 10.7 East Yornaning - Stream/Ridge Distance layer

```
Conditional Probability Printout
10/3/98        6:38:27 PM
Evidence  :    os_tpos with  3  classes
Prior probabilities
PE 1    PE 2     PE 3
0.53   0.36    0.11

Conditional Probabilities
one       two       three
1.000     0.000     0.000
0.000     1.000     0.000
0.000     0.000     1.000

Joint Probability Printout
10/3/98        6:38:39 PM
Evidence  :    os_tpos with  3  classes
Hypothesis :   soil with  3  classes
Prior probabilities
PH 1   PH 2    PH 3    PH 4
 0.56   0.277   0.16

PE 1   PE 2    PE 3
0.529 0.357   0.114
Joint Probabilities

one       two       three
0.476     0.082     0.001
0.048     0.218     0.010
0.005     0.054     0.102


Joint Probabilities as percent of Evidence


class 1   class 2    class 3
90%       23%        1%
9%        61%        9%
1%        15%        90%
```

Data panel 11.1 Brookton - Topographic position layer

Conditional Probability Printout
10/3/98 6:37:24 PM
Evidence : os_5sc with 5 classes
Prior probabilities

PE 1	PE 2	PE 3	PE 4	PE 5
0.06	0.26	0.39	0.18	0.11

Conditional Probabilities

one	two	three	four	five
1.000	0.000	0.000	0.000	0.000
0.000	1.000	0.000	0.000	0.000
0.000	0.000	1.000	0.000	0.000
0.000	0.000	0.000	1.000	0.000
0.000	0.000	0.000	0.000	1.000

Joint Probability Printout
10/3/98 6:37:31 PM
Evidence : os_5sc with 5 classes
Hypothesis : soil with 3 classes
Prior probabilities

PH 1	PH 2	PH 3
0.56	0.277	0.16

PE 1	PE 2	PE 3	PE 4	PE 5
0.062	0.260	0.394	0.179	0.105

Joint Probabilities

one	two	three	four	five
0.009	0.036	0.240	0.172	0.102
0.001	0.154	0.118	0.004	0.001
0.051	0.070	0.035	0.004	0.001

Joint Probabilities as percent of Evidence

class 1	class 2	class 3	class 4	class 5
15%	14%	61%	96%	97%
1%	59%	30%	2%	1%
83%	27%	9%	2%	1%

Data panel 11.2 Brookton - Slope layer

```
 Conditional Probability Printout
10/3/98        5:52:46 PM
Evidence  :    os_sand with  2  classes
Prior probabilities
PE 1    PE 2
0.92   0.08

Conditional Probabilities
one       two
0.960     0.150
0.040     0.850

Joint Probability Printout
10/3/98        5:57:27 PM
Evidence  :    os_sand with  2  classes
Hypothesis :   soil with  4  classes
Prior probabilities
PH 1   PH 2    PH 3
 0.56   0.277   0.16

PE 1   PE 2
0.893 0.107

Joint Probabilities


one       two
0.554     0.005
0.268     0.011
0.071     0.091



Joint Probabilities as percent of Evidence


class 1   class 2
62%       5%
30%       10%
8%        85%
```

Data panel 11.3 Brookton - Sand API layer

Conditional Probability Printout
14/03/99 16:50:53
Evidence : exp_cti with 5 classes
Prior probabilities

PE 1	PE 2	PE 3	PE 4	PE 5
0.25	0.27	0.29	0.16	0.03

Conditional Probabilities

lowest	higher	middle	bigger	bigest
0.950	0.100	0.000	0.000	0.000
0.050	0.800	0.050	0.000	0.000
0.000	0.080	0.800	0.100	0.000
0.000	0.020	0.150	0.900	0.030
0.000	0.000	0.000	0.000	0.970

Joint Probability Printout
1/4/98 12:20:01 PM
Evidence : exp_cti with 5 classes
Hypothesis : s_tx with 3 classes
Prior probabilities

PH 1	PH 2	PH 3
0.6	0.26	0.14

PE 1	PE 2	PE 3	PE 4	PE 5
0.263	0.244	0.271	0.196	0.026

Joint Probabilities

lowest	higher	middle	bigger	bigest
0.189	0.125	0.165	0.110	0.015
0.053	0.059	0.081	0.063	0.008
0.021	0.061	0.024	0.024	0.003

Joint Probabilities as percent of Evidence

class 1	class 2	class 3	class 4	class 5
72%	51%	61%	56%	56%
20%	24%	30%	32%	32%
8%	25%	9%	12%	12%

Data panel 11.4 Bundaberg - CTI data layer

Conditional Probability Printout
14/03/99 16:51:04
Evidence : exp_geol with 6 classes
Prior probabilities

PE 1	PE 2	PE 3	PE 4	PE 5	PE 6
0.28	0.05	0.02	0.02	0.17	0.47

Conditional Probabilities

cret	Quat	Trias	Gran	BAs	TE
0.800	0.050	0.050	0.000	0.050	0.200
0.000	0.900	0.000	0.000	0.000	0.000
0.000	0.000	0.800	0.050	0.000	0.000
0.000	0.000	0.100	0.950	0.000	0.000
0.000	0.000	0.050	0.000	0.900	0.000
0.200	0.050	0.000	0.000	0.050	0.800

Joint Probability Printout
1/4/98 12:30:29 PM
Evidence : exp_geol with 6 classes
Hypothesis : s_tx with 3 classes
Prior probabilities

PH 1	PH 2	PH 3
0.6	0.26	0.14

PE 1	PE 2	PE 3	PE 4	PE 5	PE 6
0.326	0.048	0.016	0.018	0.152	0.439

Joint Probabilities

cret	Quat	Trias	Gran	BAs	TE
0.150	0.035	0.003	0.018	0.003	0.391
0.173	0.013	0.013	0.001	0.009	0.048
0.003	0.000	0.000	0.000	0.134	0.004

Joint Probabilities as percent of Evidence

class 1	class 2	class 3	class 4	class 5	class 6
46%	72%	20%	97%	2%	89%
53%	28%	80%	3%	6%	11%
1%	%	%	%	88%	1%

Data panel 11.5 Bundaberg - Geology data layer

Conditional Probability Printout
14/03/99 16:51:15
Evidence : exp_slop with 5 classes
Prior probabilities

PE 1	PE 2	PE 3	PE 4	PE 5
0.30	0.38	0.16	0.10	0.06

Conditional Probabilities

0-3	1-3	3-5	5-8	>8
0.700	0.150	0.000	0.000	0.000
0.200	0.700	0.150	0.000	0.000
0.100	0.150	0.700	0.150	0.000
0.000	0.000	0.150	0.700	0.150
0.000	0.000	0.000	0.150	0.850

Joint Probability Printout
1/4/98 12:36:42 PM
Evidence : exp_slop with 5 classes
Hypothesis : s_tx with 3 classes
Prior probabilities

PH 1	PH 2	PH 3
0.6	0.26	0.14

PE 1	PE 2	PE 3	PE 4	PE 5
0.266	0.349	0.216	0.102	0.067

Joint Probabilities

0-3	1-3	3-5	5-8	>8
0.186	0.234	0.119	0.043	0.013
0.069	0.087	0.058	0.020	0.027
0.011	0.028	0.039	0.039	0.027

Joint Probabilities as percent of Evidence

class 1	class 2	class 3	class 4	class 5
70%	67%	55%	42%	20%
26%	25%	27%	20%	40%
4%	8%	18%	38%	40%

Data panel 11.6 Bundaberg - Slope data layer

```
Conditional Probability Printout
16/02/99      18:08:50
Evidence  :   acid_g with  5  classes
Prior probabilities
PE 1   PE 2    PE 3    PE 4    PE 5
0.00  0.09   0.26   0.45   0.21

Conditional Probabilities
4.2-4.3  4.3-4.4   4.4-4.5   4.5-4.6   4.6-4.7
0.900    0.050     0.000     0.000     0.000
0.070    0.900     0.050     0.000     0.000
0.030    0.050     0.900     0.050     0.030
0.000    0.000     0.050     0.900     0.070
0.000    0.000     0.000     0.050     0.900


Joint Probability Printout
16/02/99      19:25:11
Evidence  :   acid_g with  5  classes
Hypothesis :  hiyield with  2  classes
Prior probabilities
PH 1  PH 2
 0.65  0.35

PE 1  PE 2   PE 3   PE 4   PE 5
0.005 0.093  0.263  0.429  0.209

Joint Probabilities


4.2-4.3  4.3-4.4   4.4-4.5   4.5-4.6   4.6-4.7
0.004    0.065     0.171     0.270     0.134
0.001    0.028     0.092     0.155     0.075



Joint Probabilities as percent of Evidence


class 1  class 2   class 3   class 4   class 5
80%      70%       65%       63%       64%
20%      30%       35%       36%       36%
```

Data panel 11.7 Yield prediction data - pH layer

```
Conditional Probability Printout
16/02/99      18:08:27
Evidence  :   soil_g with  8  classes
Prior probabilities
PE 1   PE 2    PE 3    PE 4    PE 5    PE 6    PE 7    PE 8
0.38  0.09   0.02   0.09   0.17   0.11   0.12   0.03

Conditional Probabilities
DYS      SGL      MYGS     SPS      YSE      DSD      SSD      SLD
0.790    0.030    0.030    0.030    0.030    0.030    0.030    0.030
0.030    0.790    0.030    0.030    0.030    0.030    0.030    0.030
0.030    0.030    0.790    0.030    0.030    0.030    0.030    0.030
0.030    0.030    0.030    0.790    0.030    0.030    0.030    0.030
0.030    0.030    0.030    0.030    0.790    0.030    0.030    0.030
0.030    0.030    0.030    0.030    0.030    0.790    0.030    0.030
0.030    0.030    0.030    0.030    0.030    0.030    0.790    0.030
0.030    0.030    0.030    0.030    0.030    0.030    0.030    0.790


Joint Probability Printout
16/02/99      19:22:25
Evidence  :   soil_g with  8  classes
Hypothesis :  hiyield with  2  classes
Prior probabilities
PH 1  PH 2
 0.65  0.35

PE 1  PE 2   PE 3   PE 4   PE 5   PE 6   PE 7   PE 8
0.320 0.097  0.042  0.099  0.158  0.112  0.121  0.051

Joint Probabilities


DYS      SGL      MYGS     SPS      YSE      DSD      SSD      SLD
0.224    0.088    0.042    0.064    0.079    0.100    0.048    0.001
0.096    0.010    0.000    0.034    0.079    0.012    0.073    0.050



Joint Probabilities as percent of Evidence


class 1  class 2   class 3   class 4   class 5   class 6   class 7   class 8
70%      90%       100%      65%       50%       89%       40%       2%
30%      10%       %         35%       50%       11%       60%       98%
```

Data panel 11.8 Yield prediction data - Soil layer

```
Conditional Probability Printout
16/02/99      18:09:07
Evidence  :   yld5_g with  6  classes
Prior probabilities
PE 1   PE 2    PE 3    PE 4    PE 5    PE 6
0.02  0.15   0.42   0.29   0.10   0.02

Conditional Probabilities
0-0.5     0.5-1     1-1.5     1.5-2     2-2.5     >2.5
1.000     0.000     0.000     0.000     0.000     0.000
0.000     1.000     0.000     0.000     0.000     0.000
0.000     0.000     1.000     0.000     0.000     0.000
0.000     0.000     0.000     1.000     0.000     0.000
0.000     0.000     0.000     0.000     1.000     0.000
0.000     0.000     0.000     0.000     0.000     1.000


Joint Probability Printout
16/02/99      19:28:06
Evidence  :   yld5_g with  6  classes
Hypothesis :  hiyield with  2  classes
Prior probabilities
PH 1  PH 2
 0.65  0.35

PE 1  PE 2   PE 3   PE 4   PE 5   PE 6
0.025 0.148  0.420  0.288  0.096  0.024

Joint Probabilities


0-0.5     0.5-1     1-1.5     1.5-2     2-2.5     >2.5
0.023     0.126     0.277     0.164     0.048     0.011
0.001     0.022     0.143     0.124     0.048     0.013



Joint Probabilities as percent of Evidence


class 1  class 2   class 3   class 4   class 5   class 6
95%      85%       66%       57%       50%       45%
5%       15%       34%       43%       50%       55%
```

Data panel 11.9 Yield prediction data - Prior yield layer

APPENDIX B

DESCRIPTION OF CONTENTS OF CD-ROM.

The CD ROM appended to this thesis as Appendix B contains a copy of the Expector software, in its ArcView interface version. The ArcView interface requires the use of the Spatial Analyst extension to ArcView. Also on the disk are the manual for the software, some supporting files, and data for a worked example. The worked example is described in Section C of the manual. The manual is presented as an Adobe® PDF document and the Adobe® Acrobat® reader software is also provided on the disk.

The directory structure of the CD is as follows:-

v112_dist	Expector software distribution kit,
support	Additional files for the Expector software,
manual	The Expector manual,
exp_av	Expector ArcView interface.

Installation instructions are in the file *cdinstall.rft* which is on the root directory of the CD.

In case of any questions concerning the material on the CD, please contact either of the following:-

Robert Corner (Robert.Corner@per.clw.csiro.au),

Dr. Robert Hickey (rhickey@vesta.curtin.edu.au).