Transactions in GIS, 2002, 6(4): 383–402

Research Paper

Knowledge Based Soil Attribute Mapping In GIS: The Expector Method

Robert J. Corner
Department of Spatial Sciences
Curtin University of Technology

Robert J. Hickey
Department of Geography and Land Studies
Central Washington University

Simon E. Cook
Centro Internacional de Agricultura Tropical (CIAT)
Cali, Colombia

Abstract
EXPECTOR is a method of combining data and 'expert' knowledge within a Geographic Information System to provide information on the occurrence of spatially distributed attributes. It was developed to predict soil property values from spatially variable input data. Although initially developed to provide soil surveyors with a quantitative soil mapping method, it also has applications in land evaluation, land capability assessment, geological mapping and in precision agriculture. It operates on the basis that the state of a particular property, which may be difficult to measure directly, can be inferred from other (more measurable) entities and a knowledge of their inter-relationships. The method has been implemented as a stand-alone 'Knowledge Editing' module for the PC that can be linked to raster GIS packages. This paper describes the basis of the method and illustrates its use with an example describing the production of a surface clay content map for a small catchment in south-western Western Australia.

1 Introduction

Land resource assessment agencies in Australia are moving from a regime of traditional soil mapping towards quantitative methods of land resource assessment. Techniques suggested as candidates for quantitative land evaluation include visible and near-

Address for correspondence: Robert J. Corner, Department of Spatial Sciences, Curtin University of Technology, P.O. Box U1987, Perth, Western Australia 6845 Australia. E-mail: rcorner@vesta.curtin.edu.au

infrared remote sensing, gamma radiometrics (Wilford et al. 1997) and terrain analysis (Gessler et al. 1995, McKenzie et al. 2001). Methods using such techniques rely either on the existence of particular spectral signatures which are diagnostic of particular physical properties or on direct regression relationships between signal (or terrain attribute) and the soil attribute being mapped. However, these relationships are frequently masked by other effects and distortion of the signal (or relationship) may occur. A human interpreter of such data is capable of making value judgements to assist in the interpretation. The same applies to 'traditional' soil surveying based on site observations and air-photo interpretation. This use by the interpreter of poorly formalised ancillary data and prior knowledge has led to soil surveying being regarded as more art than science (Hewitt 1993).

Land evaluation and land capability classification have as their goals the matching of land use to land resources. The evaluation of land for any purpose requires an examination of land qualities, which are defined by the FAO (1976) as: 'Attributes of land which act in a distinct manner in their influence on the function of land for a specific kind of use'. Examples of land qualities include moisture supply, susceptibility to wind erosion and trafficability. Land qualities are complex attributes themselves composed of simpler land characteristics (FAO 1976). Land characteristics are, by this definition, either measurable or can be estimated.

Geographic Information Systems (GIS) have been used for land use inventory purposes since their inception. Historically however, in many agencies which are charged with the collection and dissemination of such information, they have been used only as a tool for the cartographic compilation of the data. This under-utilises the power of GIS to perform spatial analyses and the ability to link them with decision-making structures in order to assist in the task of collecting and quantifying resource data. The Expector method takes advantage of the power of GIS, using a probabilistic representation of 'expert' knowledge to combine disparate and sometimes conflicting, data. This enables soil surveyors, for example, to produce quantitative estimates of the occurrence of soil properties.

Expector has been designed to facilitate the quantitative combination of diverse streams of evidence into maps that represent the distribution of individual soil properties. The evidence streams take the form of GIS data layers. Typical inputs to a soil property mapping exercise might be topographic data such as slope and landform, native vegetation cover and remotely sensed data. The method has been designed to use raster data with integration and synthesis being carried out using map algebra. Although this paper focuses on the use of Expector to map soil properties, it could also be used for other applications that rely on the integration of map data.

Agriculture is a major user of soil information. In that context, some of the most important, and often most limiting, soil characteristics are those that control soil moisture. These include saturated hydraulic conductivity and infiltration rate. These characteristics can in turn be related to primary properties such as particle size distribution, organic matter content, mineralogy, etc. A truly quantitative soil survey method would require that the values of such primary attributes be mapped. Then, providing the relationships are understood and have been quantified as pedo-transfer functions, any of the other qualities may be synthesized from the primary attributes.

2 Why Map Soil Properties Rather Than Soil Map Units?

Traditionally, soils are mapped as polygons within which the soil mapping unit is supposedly homogenous. Soil surveyors know that in reality there is variability within these map units and endeavour to represent this variability in legends and reports. However, it has generally been difficult to represent this variability spatially. The degree to which it is desirable to understand and map this variation depends on the intended use of the soil map. This is to some extent a question of scale. At small scales, such maps are likely to form the basis of land evaluation and capability classification exercises. At a larger scale, the recent advent of precision agriculture techniques allows precise targeting of agricultural inputs, and precise mapping of agricultural production. Models that create control maps for inputs such as fertilisers require large-scale quantitative information about primary soil properties.

On a map produced using traditional soil survey techniques, the hard boundaries between map units are more an imposition of the techniques used than of reality. The nature of soils is such that membership of a particular type or class encompasses a range of physical and chemical properties. There will in some cases be hard distinctions and boundaries but there are also many cases where properties are gradational between points in the landscape. There is therefore a fuzziness in terms of class membership. When a surveyor is interpreting the various data sources, air-photos, field work, etc. which go into the production of the map there is similarly a degree of fuzziness in the models used. Thus, conceptual models are not always crisply defined.

Irrespective of the scale, economic constraints limit the number of sample points in an area of interest at which primary soil attributes can be measured. Some means is required of spatially extending the detailed information available at those sample points over the whole area to be mapped.

It has been suggested that the statement that a particular 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 (Bouma 1988). In short, information about both the accuracy of the information and the direction in which errors might lie is extremely valuable. Quantitative survey tools may be designed to work directly with soil properties, but if we divide those properties up into classes, a suite of probabilities of membership of each of those classes may be assigned to any particular location.

Whilst there is a practical alternative of assigning absolute values to raster grid cells, together with an associated expectation of accuracy, we are still left in doubt as to the distribution of that error. A system that uses a suite of classes enables us to state, for example, that there is a 75% chance that an attribute exceeds a value of 4 but is less than 5, and to also state the probabilities that it exceeds 5 and that it is less than 4. This tells a potential user of the information something about the distribution of the variable in attribute space. To a decision maker, this is of more use than a simple statement that the attribute value is 4.5, with a probability of 75%.

3 Probability as a Knowledge Representation Mechanism

Human experts are capable of drawing inferences from streams of disparate data, and synthesizing meaningful combinations from, say, a topographic map and an airborne gamma radiometric image. This process does not lend itself to modelling as a simple mathematical relationship. It makes no sense to perform a mathematical operation on the values of the topographic map and the radiometrics! However, provided certain conditions relating to independence are met (Bonham-Carter 1994), they can be transformed into probabilities that some other attribute, common to both, exists. We then gain access to a range of mathematical techniques.

In order for data combination methods in GIS to be meaningful, it is necessary that there be some common logic associated with the data, rather than mere statistical (and not necessarily causal) relationships. The human interpreter of such data essentially uses probabilistic reasoning in combining different data sources. There are a number of parallels here with activities such as medical diagnosis and a lot of the early work on combination of probabilities was done in that context (see Spigelhalter et al. 1993 for a review). The general area of study is that of causal probabilistic networks of which the method described here is a simplified version. Jensen (1996) provides a useful introduction to the field.

We call our method Expector. This is partly because it reflects the surveyors *expectations* and partly to acknowledge the impetus given to this work by the considerable body of literature surrounding the early 'expert system' PROSPECTOR, (e.g. Duda et al. 1978, Katz 1991). A number of discussions of probability and its application to mapping are available. Davis (1986) provides a simple treatment, whilst Skidmore (1989) looks at its application to mapping forest types. The following provides an overview of probability theory as it pertains to the work described in this paper.

3.1 The Total Probability Rule

A basic premise of probability theory is that probabilities are additive, and that the probability of an event and its converse sum to unity. That is, for an event **A**:

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

By extension of this, it can be assumed that, if there exists a set **A** of mutually exclusive, and collectively exhaustive events A_1, A_2, to A_N, then:

$$\mathrm{P}(\mathbf{A}) = \sum_{1}^{N} \mathrm{A_i} \tag{2}$$

This is sometimes referred to as the Law of Total Probability.

3.2 Bayes Theorem

Probability theory, whose history has been tied to games of chance, dates in its present form from the late eighteenth century. The general rules that allow combination of probabilities were laid down then, with the basic codification of the principles used here being provided by Bayes (1763; Reprinted 1958).

Bayes stated the following premise, and went on to prove it by means of frequency counts:

> '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**'

When restated in modern notation, this gives us the relationship that defines conditional probability. If we have two events **A** and **B**, then if we write **P(A) = a/n, P(B,A) = p/n** and **P(B|A)** as **p/a** then:

$$\mathbf{P(B|A) = \frac{P(A,B)}{P(A)}} \tag{3}$$

This relationship is generally known as Bayes' rule. A more general version of this may be written:

$$\mathbf{P(B_i|A = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}} \tag{4}$$

where $\mathbf{B_i}$ is one element of a set **B** of mutually exclusive and exhaustive events. The derivation of this can be found in standard texts such as Hacking (1965) or, with an example drawn from the earth sciences, Davis (1986).

3.3 Data Combination

In order to visualise the application of Bayes' rule in soil mapping we can replace the events **A** and **B** above with real world constructs. **A** becomes some hypothesis **H**, such as the hypothesis that gravel exists in the topsoil. The set of events **B** becomes a number of pieces of evidence, **E** which support **H** to varying degrees. That is to say, there is a predictive relationship between each of those pieces of evidence and the hypothesis. The evidence could, for example be topographic and vegetative factors such as those that might be considered by a surveyor. In the simplest of cases, the hypothesis will have a converse, in this case that gravel does not exist. A more complex situation would result in a suite of mutually exclusive hypotheses. These might for example be that gravel exists at a number of defined proportions of the total soil mass.

Equation 4 may be expanded, under an assumption of conditional independence, 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** pieces of evidence:

$$\mathbf{P(H_j|E_1, E_2, \ldots E_m) = \frac{P(H_1).\prod_{i=1}^{m} P(E_i|H_1)}{\sum_{j=1}^{n}\{P(H_j).\prod_{i=1}^{m} P(E_i|H_j)\}}} \tag{5}$$

Equation 5 is the standard method used for updating of probabilities in Bayesian networks (Cohen 1985). 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*.

It can be shown (Corner 1999) that by a further application of Bayes' rule and maintaining the assumption of conditional independence, Equation 5 may be rewritten as:

$$\mathbf{P(H_1|E_1, E_2, \ldots E_m) = \frac{P(H_1).\prod_{i=1}^{m}\left\{\frac{P(H_1|E_i)}{P(H_1)}\right\}}{\sum_{j=1}^{n}\{P(H_j|E_1, E_2, \ldots Em)}} \tag{6}$$

Equation 6 now gives us a mathematical means of combining the individual estimates of the probability of our hypothesis occurring given **m** pieces of evidence. We simply need to provide conditional probability distributions linking each evidence map and our hypothesis.

These distributions can be provided either from sample data or, in its absence, from expert knowledge. Since sample data is often somewhat skewed or biased, a mechanism is required whereby distributions based on sample data can be edited in the light of expert knowledge. Sample data might include, for example, measurements from soil pits at which both soil properties and topographic indicators have been measured. Data from these can be used to develop joint probability tables. We then require only information about the prior probability, essentially the frequency of occurrence, of each of our evidence data sets in order to convert the joint distributions to conditional probability distributions for use with Equation 6. The information on frequency of occurrence is embodied in the very digital map being used as evidence.

3.4 Dealing with Uncertainty in Input Data

The joint probability distributions derived from sample data or from expert knowledge represent real world relationships. They will be used in a combinational calculus which employs digital maps that may be less than perfect representations of the real world.

The mental analogue of Expector, that a traditional soil surveyor uses, is capable of dealing with input data of different degrees of precision by using a scheme of weighting or discounting. Expector formalises the surveyor's knowledge about the quality of the input data by using a map purity concept that populates a confusion matrix for all classes in each input data layer rather than assigning a single weight to the entire layer. This enables us to develop a calculus which overcomes the disparities between the real situation in the field and that represented on the input maps. To illustrate this, we will use the example of a multi-state piece of evidence such as a map of slope classes.

We will refer to this piece of evidence generally as **E**. The prior probabilities of the classes (their frequency of occurrence) represent a probability distribution. We can call this distribution $P(E)$ for the real or field situation and $P(E')$ for the input map data. If the map were a completely accurate representation of the real world, then the two distributions would be the same. Expector gives the user the opportunity to input a matrix of Map Purity values such as those shown in Table 1. These are the conditional probability distribution $P(E|E')$.

The process of converting the values in $P(E')$ generated by a frequency or cell count on a map can be illustrated with reference to a two class case where **E** has two

Table 1 Conditional probabilities for a three class map

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

states $\mathbf{E_A}$ and $\mathbf{E_B}$. From a cell count of the map we know the probabilities of occurrence $\mathbf{P(E_A{}')}$ and $\mathbf{P(E_B{}')}$ of map classes **A** and **B** respectively. From the map purity table we know the following:

$\mathbf{P(E_A\|E_A{}')}$	Conditional probability that field class is A if map class is A
$\mathbf{P(E_B\|E_A{}')}$	Conditional probability that field class is B if map class is A
$\mathbf{P(E_A\|E_B{}')}$	Conditional probability that field class is A if map class is B
$\mathbf{P(E_A\|E_B{}')}$	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 $\mathbf{E_A{}'}$ and $\mathbf{E_B{}'}$ are mutually exclusive and are exhaustive:

$$\mathbf{P(E_A) = P(E_A, E'_A) + P(E_A, E'_B)} \tag{7}$$

The joint probabilities in Equation 7 can be calculated by inverting the definition of conditional probability. That is:

$$\mathbf{P(E_A, E_A{}') = P(E_A|E'_A)\, {}^*\, P(E'_A)} \tag{8}$$

Hence Equation 7 expands to:

$$\mathbf{P(E_A) = P(E_A|E'_A)\, {}^*\, P(E'_A) + P(E_A|E'_B)\, {}^*\, P(E'_B)} \tag{9}$$

A similar relationship can be derived to solve for $P(E_B)$. Generalising this for more than two classes we can derive:

$$\mathbf{P(E) = \sum_i [P(E|E'_i)\, {}^*\, P(E'_i)]} \tag{10}$$

This enables us to convert the 'impure' map probability distribution to that which would pertain for a 'pure' map. This operation needs to be performed for each individual class within the evidence. It is a matrix operation that does not, at this stage, have a spatial context.

3.5 Data Uncertainty and Predictive Relationships

The effect of data uncertainty on relationships between evidence and hypothesis may be dealt with using an analogous process to that described for the determination of the the 'pure' map probability distribution in the previous section.

In order that Equation 6 may be used to combine individual estimates of probability, we require, for each possible combination of evidence class and hypothesis class, a value for the probability distribution $P(H|E)$. That distribution assumes purity of the map input. In fact what we really require is the distribution $P(H|E')$ which reflects the uncertainty in the map. For any one grid cell, $P(H|E)$ may be written long hand as:

$$\mathbf{P(Hypothesis = Class\ j\,|\,Field\ class\ is\ Class\ i)}$$

Similarly the quantity of interest $\mathbf{P(H|E')}$ is:

$$\mathbf{P(Hypothesis = Class\ j\,|\,Map\ class\ is\ Class\ i)}$$

The uncertainty in class membership has been quantified by the distribution $P(E|E')$ describing the probabilities of occurrence of the field classes at that location. By analogy with Equations 7 to 9 above, we can calculate the probability that a particular hypothesis class exists at a particular location by summing the contributions made to it by all of the possible field classes represented by the mapped class and their associated map purity. The result can be written as:

$$\mathbf{P}(\mathbf{H}|\mathbf{E}') = \sum_{\mathbf{e}} [\mathbf{P}(\mathbf{H}|\mathbf{E}) * \mathbf{P}(\mathbf{E}|\mathbf{E}')] \tag{11}$$

4 The Expector Method

The stages of the Expector process are illustrated in Figure 1. They are divided into a problem definition phase, which includes knowledge definition and data preparation, and a data processing phase which can result in the production of a map. Alternatively, the output may be passed on digitally for further processing. For example, the output from a process to map rooting depth may be passed on to a process to map suitability for a particular crop.

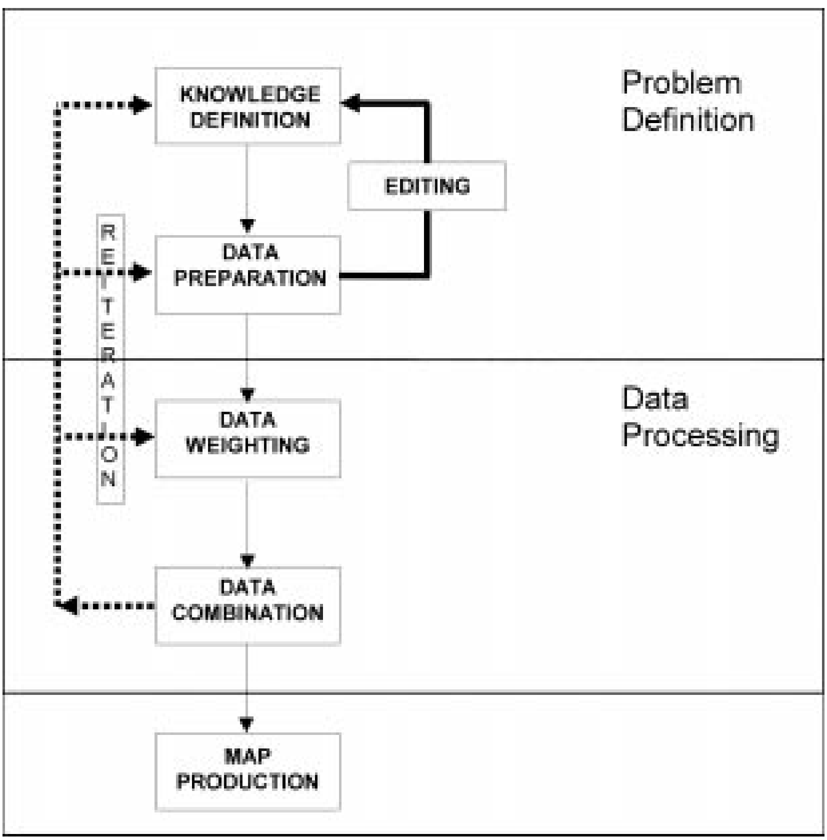

Figure 1 A schematic representation of the Expector method

4.1 Expector and GIS

The problem definition and knowledge editing part of Expector is performed using a stand-alone program coded in Visual Basic for the PC. It interfaces with a GIS both to derive information about the input spatial data and to communicate back the results of its calculations. Whilst this process can be done manually, routines have been incorporated to interface it to ArcView. The data processing and combination phases are carried out in ArcView using routines written in Avenue (the ArcView scripting language). Whilst the development of Expector has been carried out in the ArcView environment, it is capable of being interfaced to any raster GIS. The relationship between Expector and its host GIS is illustrated in Figure 2.

4.2 A Sample Expector Calculation

Figure 3 shows a schematic representation of the calculus used in Expector. This traces, for one individual pixel, the calculations involved in determining its probability of membership of two hypothesis classes (H1 and H2). The evidence in this case is a map of some classified topographic attribute, with the pixel under consideration being mapped as belonging to Class 2 of that attribute.

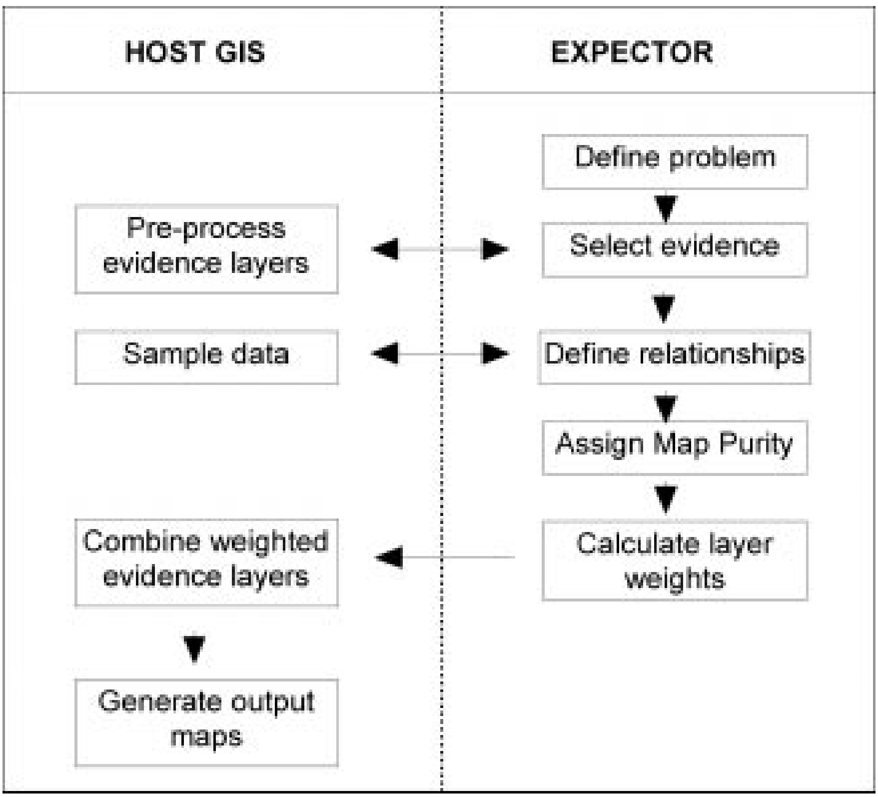

Figure 2 The relationship between Expector and its host GIS

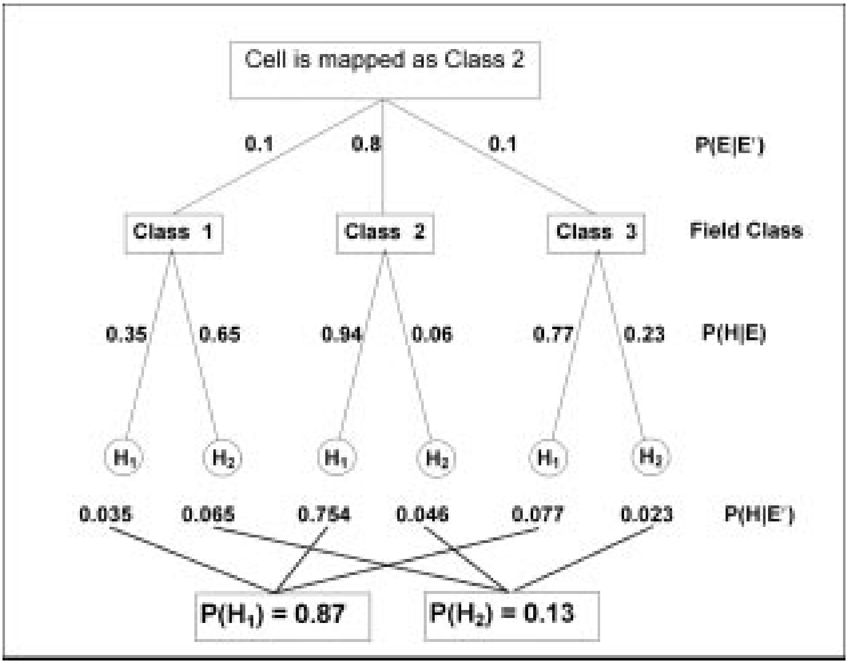

Figure 3 Schematic representation of Expector calculations

The surveyor carrying out the analysis has created a map purity table for this three class map; it is the same as that shown in Table 1 above. Referring to Figure 3, we see that any cell mapped as Class 2 has a probability **P**(**E**|**E**′)= 0.8 of actually being Class 2 in the field. There is a probability of 0.1 assigned to its membership of either Class 1 or Class 3.

Using sample data and expert knowledge, the surveyor has also developed a joint probability table linking membership of the three evidence attribute classes to the two hypothesis classes. These have been converted, using Equation 3, to conditional probabilities and are shown as **P**(**H**|**E**) in Figure 3. The setting of these joint probabilities **P**(**H,E**) is central to the method and is described below with reference to an example.

The next level of calculation in Figure 3 determines the probabilities **P**(**H**|**E**′) of the pixel belonging to each of the two hypothesis classes based on its possible membership of each of the three attribute classes. These are derived from the **P**(**E**|**E**′) and **P**(**H**|**E**) values by means of Equation 11. This also pools the contributions through each possible attribute class to give the probabilities 0.87 and 0.13 as shown in the bottom line. By contrast, the values for **P**(**H**|**E**) suggest that had the pixel belonged unequivocally to Class 2, the final probabilities would have been 0.94 and 0.06 respectively.

The calculation shown in Figure 3 is performed for each class in each evidence data layer and produces a series of virtual data layers, one for each hypothesis class – evidence class combination. These are held as attribute tables and a final combination of these on a pixel by pixel basis is carried out using a derivative of Equation 6. This estimates the probability that attribute H (Class j) exists, based upon a series of evidence layers $\mathbf{E_1}$ to $\mathbf{E_n}$ as shown in Equation 12 below:

$$P(H_j|E_1, E_2, \ldots E_n) = \frac{\left\{\prod_i \frac{P(H_j|E_i)}{P(H_j)}\right\} P(H_j)}{\sum_j P(H_j)} \tag{12}$$

5 An Example of the Use of Expector Method

The Expector method is perhaps best illustrated by an example. We will consider an exercise to map surface clay content across the East Yornaning catchment, located in the southwest of Western Australia. It has an area of approximately 200 km^2 and is representative of the dissected lateritic landscape which occurs extensively within throughout the region (Mulcahy 1973).

5.1 Determining Prior Probability Values

The data available comprised a number of spatially distributed datasets, as well as point data representing some 200 site descriptions. We decided that clay content could be divided into three classes, 0–5%, 5–10%, and >10%. Analysis of the sample site data enabled us to set the prior probability of occurrence of each of these classes $\mathbf{P(H_i)}$ at 0.14, 0.51 and 0.35, respectively, using an assumption that the sites well represented the area being mapped. A grid resolution of 25 m was chosen, based on the size of the area to be mapped and the spatial resolution of the input data.

5.2 Evidence Data Sets

The data layers used as evidence are listed in Table 2, together with a brief description of their predictive relationship to clay content. Selecting the data sets and classifying them constitutes the first part of the problem definition phase. Figure 4 shows all of the data sets together with the classes used. Some data sets, such as the geology map, were already classified, others, such as the terrain attributes, required classification. Classifying such datasets requires some skill on the part of the expert soil surveyor working in conjunction with a GIS analyst to convert vague notions such as 'low slopes' or 'close to ridge top' into meaningful classes and to then prepare the data sets

Table 2 Data used to map surface clay content

Data set	Basis of relationship to surface clay content
Catchment stream order	Relative position in landscape, indicative of physical weathering status (fluvial processes)
Topographic Curvature	Partitions area into depositional and erosional environments
Geology	Parent material for insitu weathering or short range transport
Distance from rock outcrop	Depth of weathering, degree of removal
Airborne γ Radiometrics (Classified)	Parent material, not necessarily in situ
Slope	Surface transport (colluvial processes)
Stream:ridge distance ratio	Colluvial movement

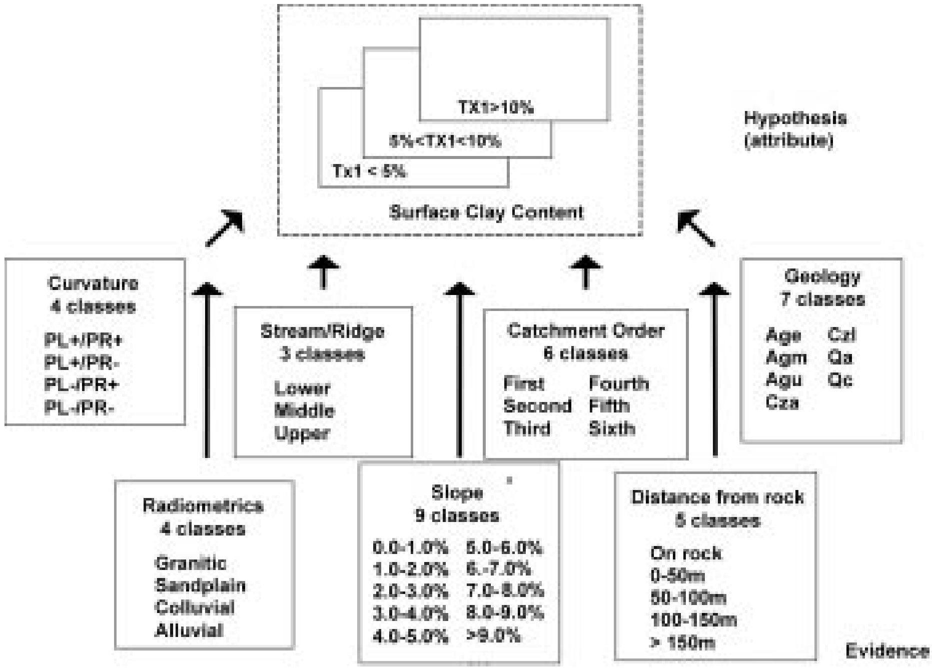

Figure 4 Evidence data classes used at East Yornaning

accordingly. Formal objective schemes for classifying landscapes into geomorphic units do exist (Dikau 1989) and others are the subject of recent and ongoing research (Burrough et al. 2001, Shary et al. 2002). It is important, if expert knowledge is to be used to develop or influence the predictive relationships between these classes and the soil attributes being mapped, that those classes should be relevant to the experience of the expert involved in this process. Each input data set was therefore classified according to its own characteristics.

5.2.1 Curvature

Four curvature classes were used representing the four possible combinations of plan and profile curvature. Positive curvature is indicative of convex and potentially erosive areas whilst negative curvature indicates concave and potentially depositional areas.

5.2.2 Airborne Gamma Radiometric Data

The airborne gamma radiometric data was acquired as four band data representing total gamma counts and counts in spectral windows sensitive to Uranium, Potassium and Thorium. This dataset data was classified using standard supervised classification techniques. A full description of the treatment of this dataset is given in Cook et al. (1996). The four classes were attributed by that study as being representative of granitic areas, sand plain, and areas of colluvium and alluvium.

5.2.3 Stream to Ridge Distance Ratio

This is a measure of position in the landscape and was determined by taking the ratio of distance to nearest stream and distance to nearest ridge. Distances were determined over a 'cost surface' of slope and thus approximate to flowpath distances. These were divided into three classes with class boundaries at 0, 0.33, 0.67 and 1.0.

5.2.4 *Slope*

The slope map was derived from a high resolution Digital Elevation Model (DEM) of the catchment. This was created using an analytical stereoplotter from 1:25,000-scale photography as spot heights with a horizontal resolution of 25 m. The catchment is an area of generally low relief with slope values ranging from 0 to 9%. These were classified at 1% intervals in order to allow the expert latitude in the assignment of probabilities.

5.2.5 *Catchment Order*

The stream network data for the catchment was classified using the Strahler stream ordering method. Watersheds and sub-basins were defined using the DEM and each sub-basin was assigned a value indicating the order of the stream that drains it. This provides an indication of the erosive potential of the sub-basin.

5.2.6 *Distance from Rock*

Rock outcrops were digitised from 1:25,000 scale colour aerial photography and used as the basis for distance zones. These zones were classified at 50 metre intervals up to 150 metres. Expert opinion suggests that in this particular landscape the influence of rock outcrop on soil forming processes diminishes over that range.

5.2.7 *Geology*

The geology was taken from the 1:250,000 map of the area (Chin 1986). These data are already categorical and required no further processing. Seven lithological types occur within the catchment. Three of these (Age, Agm and Agu) are granitic, the remainder are laterite, colluvium and alluvium of Cainozoic (Cz) or Quaternary (Q) age.

5.3 *Knowledge Definition*

Once all the data have been selected and classified, the work of knowledge definition and editing can begin. Knowledge definition and editing are facilitated by the use of simple forms to define the 'purity' of the input maps and the relationships of the various data layers to the hypothesis. Behind each of these forms is a table which holds the values entered by the expert. Figure 5 shows a data relationship entry form.

5.3.1 *Map Purity*

Table 3 shows the map purity table for the geology input layer. Map purity tables may be produced using some combination of experience and expert knowledge or by cross checking between fieldwork and the published map or by a combination of both. If read down the columns, a map purity table represents the probability that a pixel mapped in a particular class is really in any of the other available classes. In this case the map purity was determined by expert opinion. For example, the expert has determined that a grid cell mapped as being Cza (Cainozoic alluvium) has a high probability of belonging to its own class, a slight probability of belonging to the Qa class but zero probability of belonging to the others. It may be noted that the

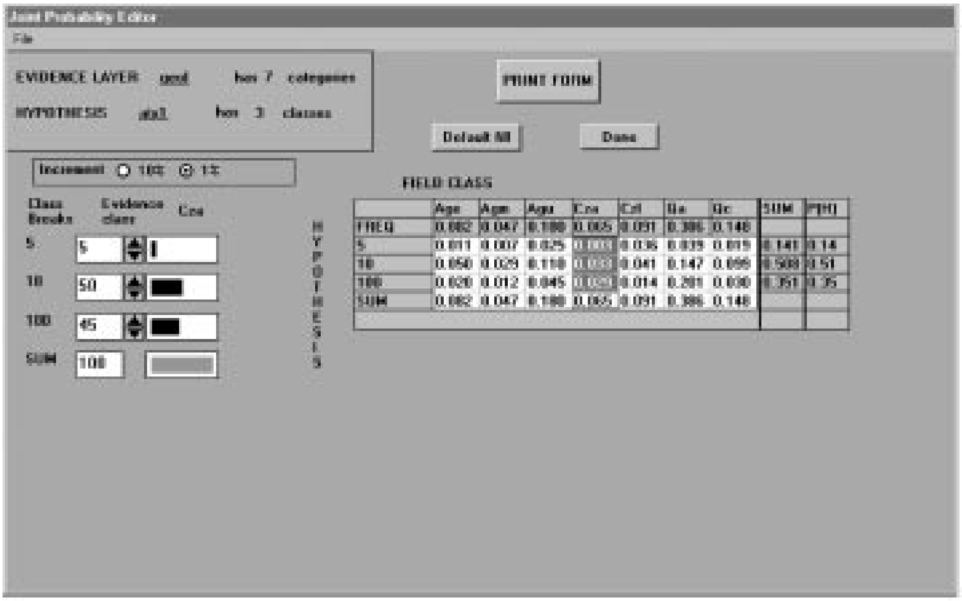

Figure 5 Joint probability editing form

Table 3 Map purity table for geology

Field Class		Map Class (Symbol)						
Unit	Symbol	Age	Agm	Agu	Cza	Czl	Qa	Qc
Biotite granite/adamellite	Age	0.80	0.10	0.10	0.00	0.03	0.00	0.00
Mixed Granitic rocks	Agm	0.10	0.80	0.10	0.00	0.03	0.00	0.00
Adamellite	Agu	0.10	0.10	0.80	0.00	0.03	0.00	0.00
Cainozoic Alluvium	Cza	0.00	0.00	0.00	0.80	0.00	0.10	0.15
Cainozoic Laterite	Czl	0.00	0.00	0.00	0.00	0.80	0.00	0.00
Quaternary Alluvium	Qa	0.00	0.00	0.00	0.20	0.00	0.80	0.05
Quaternary Colluvium	Qc	0.00	0.00	0.00	0.00	0.11	0.10	0.80

probability down any one column must sum to unity, honouring the total probability rule (Equation 2).

All other evidence datasets in this example had their map purity derived from expert opinion. In cases where mapped evidence occurs in site sample data, and the sample size is sufficiently large, then map purity may be assessed analytically. In order to do this, a coincidence matrix is developed between the sample data and the map.

5.3.2 Joint probabilities

Table 4 shows the association between the geology map classes and the expected surface clay content. Again, this table can be considered by taking each column in turn. The percentages in each cell in the column represent the joint probability that a point will belong to both a particular clay content class and to a particular geology class. The software works with raw joint probabilities but provides a facility for the expert to

Table 4 Clay content related to Geology unit

Clay Content	Field Class Age	Agm	Agu	Cza	Czl	Qa	Qc
0–5%	14%	14%	14%	5%	40%	10%	13%
5–10%	61%	61%	61%	50%	45%	38%	67%
>10%	25%	25%	25%	45%	15%	52%	20%

enter them as more readily visualised percentages. In this case, the expert has assigned grid cells mapped as Cainozoic alluvium a 50% chance of having a surface texture of 5–10% clay, with the other clay classes having probabilities of 5% and 45%. The interface form used in building this relationship is shown in Figure 5.

A key to the development of joint probability tables is the use of coincidence matrices to determine the hypothesis and evidence classes mapped at sample site locations. These matrices are generated by interrogating the input data at the sample locations, using tools contained in Expector's GIS interface. The coincidence matrices provide 'seed values' for joint probability tables such as that shown in Table 4. Bias within the samples, as a result of their spatial distribution, may lead to apparently misleading results. Expector therefore allows the expert to directly edit the joint probability tables in order to resolve any conflict between potentially biased samples and their own understanding of a particular landscape. For the example shown here most relationships required some editing. This is largely a function of the relatively small number of sample sites. It is also an indication of the power of Expector in that it permits the expert to intervene and override sample data that does not accord with their mental model of the landscape. It is of course possible that, in the presence of overwhelming evidence based on sample data, the expert's mental model could be modified. In the example discussed here the sample size was small enough to require expert intervention.

5.4 Calculation of Hypothesis Probabilities Based on Single Evidence Layers

Once the map purity and joint probability tables have been created for each input evidence layer, the Expector software produces from them a third table. In the case of the Geology input layer, this represents the probability that a grid cell mapped as being in a particular lithology class will have a particular surface clay content. Table 5 is an example of one such table. The probabilities for the first three classes (all granitic parent rocks) have not changed, because they have been regarded in Table 3 (map purity) as being confused only with each other and in Table 4 as having identical relationships with clay content. For other classes, the change in values reflects the degree of purity assigned to them.

This process is repeated for all the evidence layers in Table 2. We have now remapped the classes in the input (evidence) data as probabilities of occurrence of classes in the predicted attribute. Every grid cell now has assigned to it a probability of membership of each clay class based on each of the evidence layers. Tools are available in Expector's GIS interface to display as maps the probabilities of belonging to the various clay (or other attribute) classes associated with individual data layers.

Table 5 Probability of membership of surface texture classes taking only geology as evidence

Clay Content	Field Class Age	Agm	Agu	Cza	Czl	Qa	Qc
0–5%	14%	14%	14%	6%	35%	10%	12%
5–10%	61%	61%	61%	48%	49%	42%	63%
>10%	25%	25%	25%	46%	16%	48%	25%

5.5 Combination of Probabilities from Several Input Layers

The various 'virtual maps' represented by these tables are then combined on a pixel by pixel basis for each hypothesis class. As each map is taken as evidence, a look-up is performed on Table 5 or its equivalent and the probability values for the appropriate hypothesis are fed into Equation 11. This is the data combination part of the Expector method which takes into consideration the effect of each of those class memberships. Figure 6 is a map showing the probability of membership of clay Class 2 (5–10% clay) taking all the evidence layers shown in Table 1 into account. Similar maps were produced for the other two clay classes.

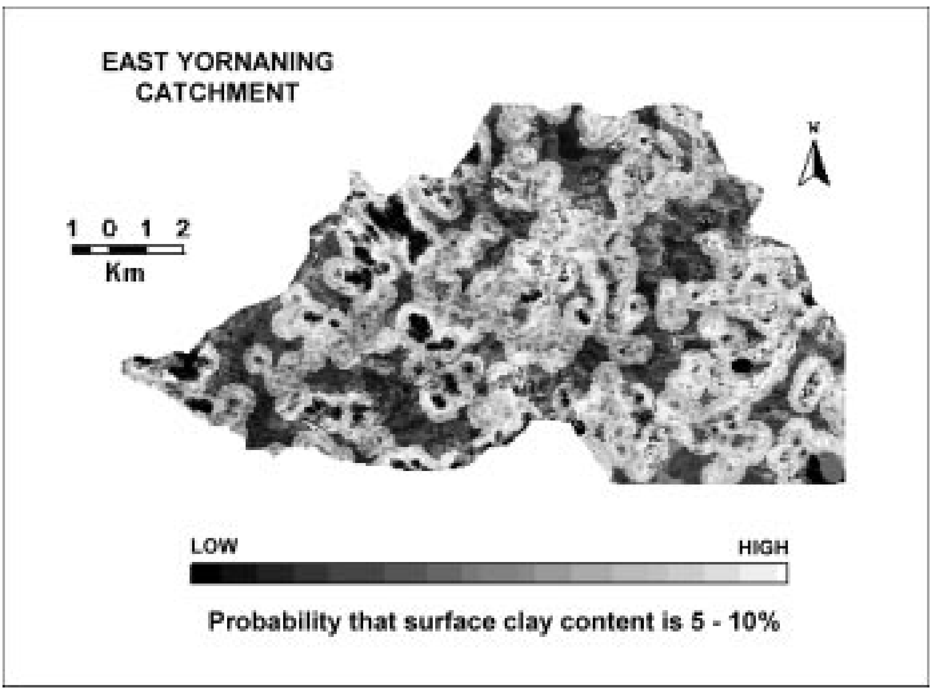

Figure 6 Map of East Yornaning catchment showing probability that surface clay content is between 5 and 10%

6 Discussion

The Expector method produces digital maps showing the probability of membership of target classes (e.g. surface clay content 5–10%). It is intended that these be used as decision support aids. It is to be expected that analysts will also wish to generate maps showing the most probable class. Tools for generating such maps are available through customised routines in the ArcView interface. Figure 7 shows the most probable class map for surface clay content in the Yornaning catchment.

Most probable class maps provide a means of readily assessing the accuracy of the output estimates. Probability of occurrence is not a field measurable attribute but the actual class present can be measured. In the case of the Yornaning data, an estimate of the accuracy of the process was gained by examining the agreement between the map of most probable clay class and the values from analyses carried out at the 200 sample sites. 51% of the sample points were correctly predicted. A similar check for the only other soil map of the area, mapped by local landowners with the assistance of a soil surveyor (Patabendige 1993) showed that only 40% of the sites are correctly attributed by that map. It should be noted, however, that in the case of the soil map, assignments to surface texture classes were made based on map unit descriptions which contained a number of broad groupings.

In other tests with larger numbers of output classes, Expector has predicted up to 95% of the input sites as being either in the correct class or within one class of their correct allocation (Corner 1999). Those tests were carried out during the development of Expector and covered relatively small areas of five to ten hectares. These areas had typically been sampled on 25 or 50 m grids and thus offered extremely rich sample data sets. The example from East Yornaning shows that this method can perform as well as a traditional survey under operational conditions that limit sampling density to approximately one observation per square kilometre. In such a case reliance must be placed on expert knowledge and the method has performed well under conditions that approach those used by a mapping agency.

A number of operational considerations need to be borne in mind when using Expector. Probably the most critical of these is that the evidence layers used in the analysis must be conditionally independent. Expector shares this constraint with other methods that are based on Bayesian calculus and in practice it is almost certain to be breached to some extent. Any evidence layers that can be used to describe the spatial distribution of a hypothesis attribute will contain some degree of statistical dependence, simply due to their co-relationship to the target attribute. Although formal tests such as entropy measures may be used to determine measures of dependence, it is frequently necessary to rely on the judgment of an expert in determining the extent to which the independence condition may be breached.

As a general guide, it is perhaps more important to consider the functional relationships between the predictive evidence and the attribute being mapped. For example, there may be a close relationship between two input data sets, such as geology and a topographic factor attribute such as profile curvature. However, if one is affecting the target attribute by virtue of differing supply of in-situ weathering products and the other through the medium of differing drainage regime, then they can be considered to be functionally independent.

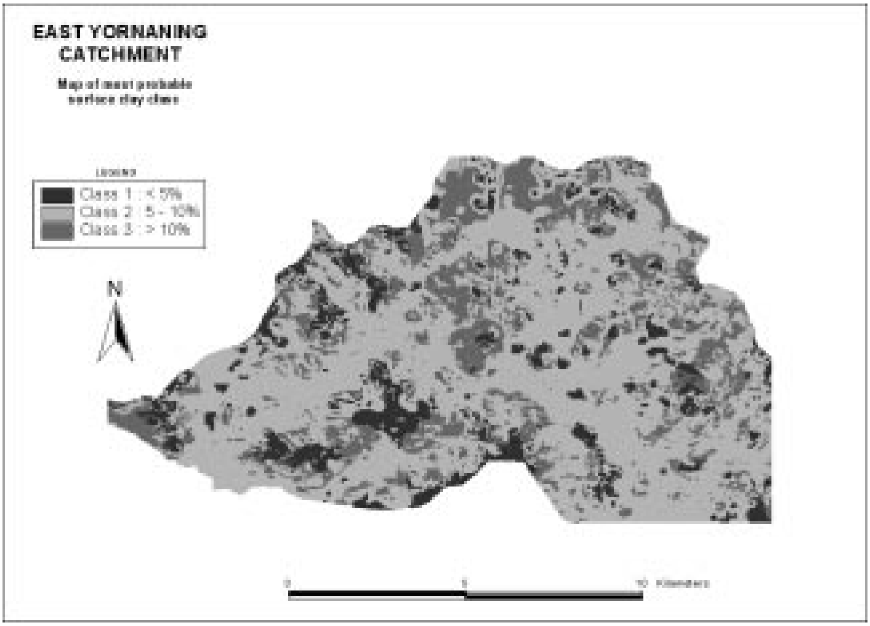

Figure 7 An example of a most probable class map

7 Conclusions

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. Conceptual models are both flexible and appealing to the natural resource surveyor and 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. A method such as Expector offers the opportunity to map the probability of occurrence of individual soil attributes, so as to preserve that information.

Expector is essentially a probabilistic interpolator. That is, it takes knowledge about relationships derived from a subset of an area (samples or surveyor impressions) and then uses spatially extensive data sets to derive maps of the whole area. As such, it is rather sensitive to bias in the input samples or knowledge, a characteristic that it shares with the traditional soil survey method.

The method has been used to produce a map of soil surface clay content over a moderately large catchment in Western Australia. In this case, the soil surveyor's knowledge was assisted by a relatively sparse sample dataset. The level of accuracy in its representation of clay content was 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.

Acknowledgements

The Development of the Expector method and its associated software was assisted by a grant from the Land and Water Resource Research and Development Corporation (LWRRDC) whilst two of the authors (RJC and SEC) were with CSIRO Land and Water. The support of both LWRRDC and CSIRO is acknowledged. Gerard Grealish, Geoff Moore and their colleagues at the Agriculture Western Australia provided invaluable assistance during the project.

References

Bayes T 1763 An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 53: 370–418

Bayes T 1958 An essay towards solving a problem in the doctrine of chances. *Biometrika*. 45: 296–315 (Reprint).

Bonham-Carter G F 1994 *Geographic Information Systems for Geoscientists: Modelling with GIS*. Kidlington, Pergamon-Elsevier Science

Bouma J 1988 Using soil survey data for quantitative land evaluation. *Advances in Soil Science* 9: 177–213

Burrough P A, Wilson J P, van Gaans P F M and Hansen A J 2001 Fuzzy k-means classification of topo-climatic data as an aid to forest mapping in the Greater Yellowstone Area, USA. *Landscape Ecology* 16: 523–546

Chin, R J 1986 Corrigin, Western Australia (Sheet SI 50-3). Perth, Geological Survey of Western Australia 1:250,000 Geological Series Maps and Explanatory Notes

Cohen P R 1985 *Heuristic Reasoning about Uncertainty: An Artificial Intelligence Approach*. London, Pitman

Cook, S E, Corner R J, Grealish G. and Groves R 1996 Use of airborne gamma radiometrics for soil mapping. *Australian Journal of Soils Research* 34: 183–194

Corner R 1999 Knowledge Representation in Geographic Information Systems. Unpublished PhD Dissertation, School of Spatial Sciences, Curtin University

Davis J C 1986 *Statistics and Data Analysis in Geology*. New York, John Wiley and Sons

Dikau, R. 1989 The application of a digital relief model to landform analysis in geomorphology. In Raper J (ed.) *Three Dimensional Applications in Geographic Information Systems*. London, Taylor and Francis: 51–77

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. Stamford, CA, SRI International Artificial Intelligence Center (Final Report for SRI Projects 5821 and 6415)

FAO 1976 *A Framework for Land Evaluation*. Rome, United Nations Food and Agriculture Organisation

Gessler P E, Moore I D, McKenzie N J, and Ryan P J, 1995 Soil-landscape models and the spatial prediction of soil attributes. *International Journal of Geographic Information Science* 9: 421–32

Hacking I M 1965 *The Logic of Statistical Inference*. Cambridge, Cambridge University Press

Hewitt A E 1993 Predictive modelling in soil survey. *Soils and Fertility* 56: 305–14

Jensen F V 1996 *An Introduction to Bayesian Networks*. London, UCL Press

Katz S S 1991 Emulating the Prospector expert system with a raster GIS. *Computers and Geoscience* 17: 1033–50

McKenzie N J, Gessler P E, Ryan P J, and O'Connell D A 2001 The role of terrain analysis in soil mapping. In Wilson J P and Gallant J C (eds) *Terrain Analysis: Principles and Applications*. New York, John Wiley and Sons: 245–65

Mulcahy M 1973 Landforms and soils of south-western Australia. *Journal of the Royal Society of Western Australia* 56: 16–22

Patabendige D 1993 Farmers soil map of the East Yornaning catchment. Perth, Western Australia Department of Agriculture (Unpublished)

Shary P A, Sharaya L S, and Mitusov A V 2002 Fundamental quantitative methods of land surface analysis. *Geoderma* 107:1–32

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* 55: 1449–64

Spiegelhalter D J, Dawid A P, Laurtizen S L, and Cowell R G 1993 Bayesian analysis in expert systems. *Statistical Science* 8: 219–83

Wilford J R, Bierwith P N, and Craig M A 1997 Application of airborne gamma-ray spectrometry in soil/regolith mapping and applied geomorphology. *AGSO Journal of Australian Geology and Geophysics* 17: 201–16