2

Taxonomy of environmental models in the spatial sciences

Andrew K. Skidmore

2.1 INTRODUCTION

Environmental models simulate the functioning of environmental processes. The motivation behind developing an environmental model is often to explain complex behaviour in environmental systems, or improve understanding of a system. Environmental models may also be extrapolated through time in order to predict future environmental conditions, or to compare predicted behaviour to observed processes or phenomena. However, a model should not be used for both prediction and explanation tasks simultaneously.

Geographic information system (GIS) models may be varied in space, in time, or in the state variables. In order to develop and validate a model, one factor should be varied and all others held constant. Environmental models are being developed and used in a wide range of disciplines, at scales ranging from a few meters to the whole earth, as well as for purposes including management of resources, solving environmental problems and developing policies. GIS and remote sensing provide tools to extrapolate models in space, as well as to upscale models to smaller scales.

Aristotle wrote about a two-step process of firstly using one's imagination to inquire and discover, and a second step to demonstrate or prove the discovery Britannica 1989 14:67). This approach is the basis of the scientific approach, and is applied universally for environmental model development in GIS. In the section on empirical models, the statistical method of firstly exploring data sets in order to discover pattern, and then confirming the pattern by statistical inference, follows this process in a classical manner. But other model types also rely on this process of inquiry and then proof. For example, the section on process models shows how theoretical models based on experience (observation and/or field data) can be built.

Why spend time developing taxonomy of environmental models – does it serve any purpose except for academic curiosity? In the context of this book, taxonomy is a framework to clarify thought and organize material. This assists a user to easily identify similar environmental models that may be applied to a problem. In the same way, model developers may also utilise or adapt similar models. But taxonomy also gives an insight to very different models, and hopefully helps in transferring knowledge between different application areas of the environmental sciences.

2.2 TAXONOMY OF MODELS

Using terminology found in the GIS and environmental literature, models are here characterized as 'models of logic' (inductive and deductive), and 'models based on processing method' (deterministic and stochastic) (see Table 2.1). The deterministic category has been further subdivided into empirical, process and knowledge based models (Table 2.1). The sections of this chapter describe the individual model type; that is, a section devoted to each column of Table 2.1 (e.g. see 2.3.2 for inductive models) or row (e.g. see 2.4.1 for deterministic-empirical models). In addition, an example of an environmental application is cited for each model.

An important observation from Table 2.1 is that an environmental model is categorized by both a processing method and a logic type. For example, the CART model (see 2.3.2) is both deterministic (empirical) as well as inductive. In categorizing models based on this taxonomy, it is necessary to cite both the logic model and the processing method.

Finally, a model may actually be a concatenation of two (or more) categories in Table 2.1.

Table 2.1: A taxonomy of models used in environmental science and GIS.

			Model of logic (see Section2. 3)	
			Deductive (see Section 2.3.1)	**Inductive (see Section 2.3.2)**
Model based on processing method (see Section 2.4)	**Deterministic (see Section 2.4)**	**Empirical (see Section 2.4.1)**	Modified inductive models (e.g. R-USLE); process models classification by supervised classifiers (model inversion)	Statistical models (e.g. regression such as USLE); training of supervised classifiers (e.g. maximum likelihood) threshold models (e.g. BIOCLIM) rule induction (e.g. CART) Others: geostatistical models, Genetic algorithms
		Knowledge (see Section 2.4.2)	Expert system (based on knowledge generated from experience)	Bayesian expert system; fuzzy systems
		Process (see Section 2.4.3)	Hydrological models Ecological models	Modification of inductive model coefficients for local conditions by use of field or lab data
	Stochastic (see Section 2.5)		Monte Carlo simulation	Neural network classification; Monte Carlo simulation

For example, a model may be a combination of an inductive-empirical and a deductive-knowledge method. Care must be taken to identify the components of the

model, otherwise the taxonomic system will not work. This point is addressed further in the chapter.

2.3 MODELS OF LOGIC

2.3.1 Deductive models

A deductive model draws a specific conclusion (that is generates a new proposition) from a set of general propositions (the premises). In other words, deductive reasoning proceeds from general truths or reasons (where the premises are self-evident) to a conclusion. The assumption is that the conclusion necessarily follows the premises; that is, if you accept the premises, then it would be self-contradictory to reject the conclusion.

An example of deduction is the famous Euclid's 'Elements', a book written about 300 BC. Euclid first defines fundamental properties and concepts, such as point, line, plane and angle. For example, a line is a length joining two points. He then defines primitive propositions or postulates about these fundamental concepts, which the reader is asked to consider as true, based on their knowledge of the physical world. Finally, the primitive propositions are used to prove theorems, such as Pythagoras' theorem that the sum of the squares of a right-angled triangle equals the square of the length of the hypotenuse. In this manner, the truth of the theorem is proven based on the acceptance of the postulates.

Another example of deduction is the modelling of feedback between vegetation cover, grazing intensity and effective rainfall and development of patches in grazing areas (Rietkerk 1998). In Figure 2.1 (taken from Rietkerk *et al.* 1996 and Rietkerk 1998), the controlling variables are rainfall and grazing intensity, while the state variable is the vegetation community. State I in Figure 2.1 are perennial grasses, state II are annual grasses and state III are perennial herbs. The diagram links together a number of assumptions and propositions (taken from the literature) about how a change in rainfall and grazing intensity will alter the mix of the state variables (*viz.* perennial grass, annual grass, and perennial herbs). For example, it is assumed that the three vegetation states are system equilibria. Rietkerk *et al.* (1996) show that according to the literature this is a reasonable assumption; the primeval vegetation of the Sahel at low grazing intensities is a perennial grass steppe. They go on to discuss the various transition phases between the three vegetation states and to support their conclusion that Figure 2.1 is reasonable they cite propositions from the literature.

For example, transition 'T2a' in Figure 2.1, is a catastrophic transition where low rainfall is combined with high grazing, leading to rapid transition of perennial grass to perennial herbs, without passing through the annual grass stage II. Such deductive models have been rarely extrapolated in space.

In all these examples, the deductive model is based on plausible physical laws. The mechanism involved in the model is also described.

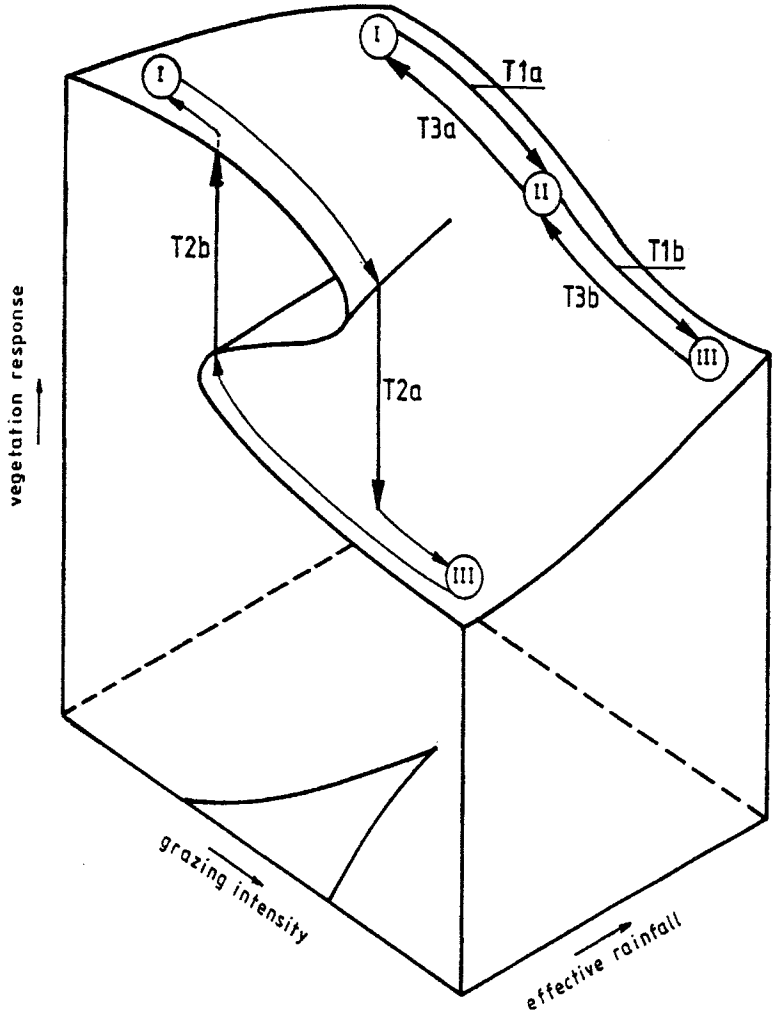

Figure 2.1: The cusp catastrophe model applied to the Sahelian rangeland dynamics (from Rietkerk *et al.* 1998).

2.3.2 Inductive models

The logic of inductive arguments is considered synonymous with the methods of natural, physical and social sciences. Inductive arguments derive a conclusion from particular facts that appear to serve as evidence for the conclusion. In other words, a series of facts may be used to derive or prove a general statement. This implies

that based on experience (usually generated from field data), induction can lead to the discovery of patterns. The relationship between the facts and the conclusion is observed, but the exact mechanism may not be understood. For example, it may be found from field observation or sampling that a tree (*Eucalyptus sieberi*) frequently occurs on ridges, but such an observation does not explain the occurrence of this species at this particular ecological location.

As noted above, induction is considered to be an integral part of the scientific method and typically follows a number of steps:

- Defining the problem using imagination and discovery.
- Defining the research question to be tested.
- Based on the research question, defining the research hypotheses that are to be proven.
- Collecting facts, usually by sampling data for statistical testing.
- Exploratory data analysis, whereby patterns in the data are visualized.
- Confirmatory analysis rejects (or fails to reject) the research hypothesis at a specified level of confidence and draws a conclusion.

The inductive method as adopted in science, and formalized in statistics, claim that the use of facts (data) leads to an ability to state a probability (that is a confidence or level of reasonableness) about the conclusion.

An example of an inductive model is the classification and regression tree (CART) method also known as a decision tree (Brieman *et al.* 1984; Kettle 1993; Skidmore *et al.* 1996). It is a technique for developing rules by recursively splitting the learning sample into binary subsets in order to create the most homogenous (best) descendent subset as well as a node (rule) in the decision tree (Figure 2.2a) (see Brieman *et al.* 1984; and Quinlan 1986 for details about this process). The process is repeated for each descendent subset, until all objects[1] are allocated to a homogenous subset. Decision rules generated from the descending subset paths are summarized so that an unknown grid cell may be passed down the decision tree to obtain its modelled class membership (Quinlan 1986) (Figure 2.2b). Note that in Figure 2.2a, the distribution of two hypothetical species ($y = 0$ and $y = 1$) is shown with gradient and topographic position, where topographic position 0 is a ridge, topographic position 5 is a gully, and values in between are midslopes. The data set is split at values of gradient = 10° and topographic position = 1.

The final form of the decision tree is similar to a taxonomic tree (Moore *et al.* 1990) where the answer to a question in a higher level determines the next question asked. At the leaf (or node) of the tree, the class is identified.

[1] For example in the paper by Skidmore *et al.* (1996), the objects were kangaroos.

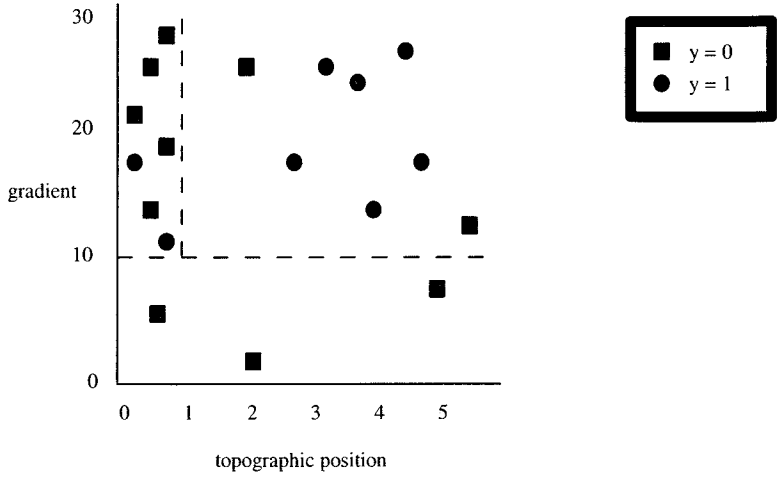

Figure 2.2a: The distribution of two hypothetical species (y = 0 and y = 1) is shown with gradient and topographic position, where topographic position 0 is a ridge, topographic position 5 is a gully, and values in between are midslopes.

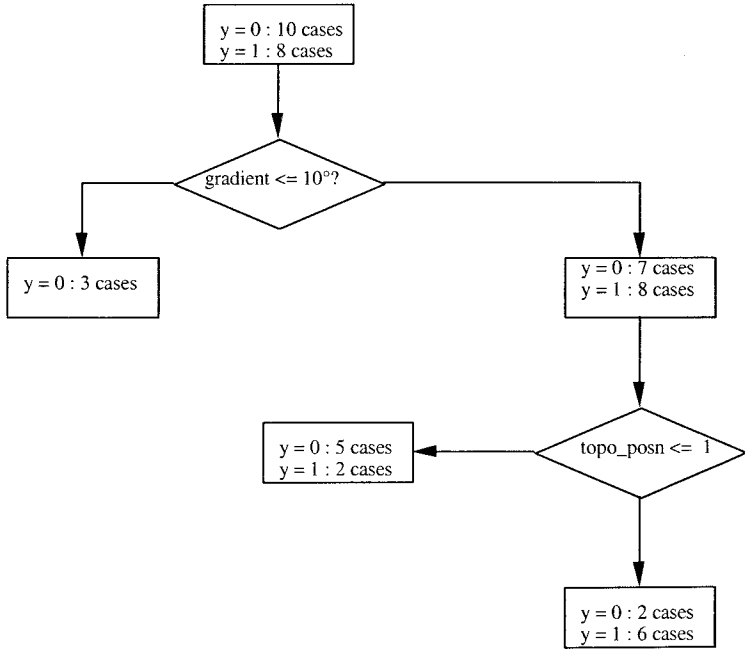

Figure 2.2b: The decision tree rules generated from the data distribution in Figure 2.2a.

2.3.3 Discussion

Both inductive and deductive methods have been used for environmental modelling. However, inductive models dominate spatial data handling (GIS and remote sensing) in the environmental sciences. As stated in 2.2, some models are a mix of methods; a good example of a mix of inductive and deductive methods is a global climate model (see also Chapter 4 by Reed *et al.* as well as Chapter 5 by Los *et al.*). In these models, complex interactions within and between the atmosphere and biosphere are described and linked. For example, photosynthesis is calculated as a function of absorbed photosynthetically active radiation (APAR), temperature, day length and canopy conductance of radiation. A component of this calculation is the daily net photosynthesis, the rationale for which is given by Hazeltine (1996). Some of the parameters in this calculation of daily net photosynthesis may be estimated from remotely sensed data (such as the fraction of photosynthetically active radiation) or interpolated from weather records (such as daily rainfall), while other constants are estimated from laboratory experiments (e.g. a scaling factor for the photosynthetic efficiency of different vegetation types). Thus the formula has been deduced, but the components of the formulae that include constants and variable coefficients are calculated using induction.

Classification problems may be considered to be a mix of deductive and inductive methods. The first stage of a classification process is inductive, where independent data (usually collected in the field or obtained from remotely sensed imagery) are explored for possible relationships with the dependent variable(s) that is to be modelled. For example, if land cover is to be classified from satellite images, input data are collected from known areas and used to estimate parameters of a particular image classifier algorithm such as the maximum likelihood classifier (Richards 1986). The second stage of the supervised classification process is deductive. The decision rules (premises) generated in the first phase are used to classify an unknown pixel element, and come up with a new proposition that the pixel element is a particular ground cover. Thus, the classification of remotely sensed data is in reality two (empirical) phases – the first phase (training) uses induction and the second phase (classification) uses deduction.

Another example of a combined inductive-deductive model in GIS may be based on a series of rules (propositions) that a GIS analyst believes are important in determining a process or conclusion. For example, a model has been developed to map the dominant plant type at a global scale (Hazeltine 1996). The model is deduced from propositions linking particular biome types (e.g. dry savannas) to a number of independent variables including:

- leaf area index
- net primary production
- average available soil moisture
- temperature of the coldest month
- mean daily temperature
- number of days of minimum temperature for growth.

The thresholds for the independent variable determining the distribution of the biome type are induced from observations and measurements by other ecologists.

For example, dry savannas are delineated by a leaf area index of between 0.6 and 1.5, and by a monthly average available soil moisture of greater than 65%.

A well-known philosophy in science, developed by Popper, rejects the inductive method for the physical (environmental) sciences and instead advocates a deductive process in which hypotheses are tested by the 'falsifiability criterion'. A scientist seeks to identify an instance that contradicts a hypothesis or postulated rule; this observation then invalidates the hypothesis. Putting it another way, a theory is accepted if no evidence is produced to show it is false.

2.4 DETERMINISTIC MODELS

A deterministic model has a fixed output for a specific input. Most deterministic models are derived empirically from field plot measurements, though rules or knowledge may be encapsulated in an expert system and will consistently generate a given output for a specific input. Deterministic models may be inductive or deductive.

2.4.1 Empirical models

Empirical models are also known as statistical, numerical or data driven models. This type of model is derived from data, and in science the model is usually developed using statistical tools (for example, regression). In other words, empiricism is that beliefs may only be accepted once they have been confirmed by actual experience. As a consequence, empirical models are usually site-specific, because the data are collected 'locally'. The location at which the model is developed may be different to other locations (for example, the climate or soil conditions may vary), so empirical models of the natural environment are not often applicable when extrapolated to new areas.

For empirical models used in the spatial sciences, models are calculated from (training) data collected in the field. Recall that inductive models also use training data, so a model may be classified as inductive-empirical (see 2.3.2). However, not all inductive models are empirical (see Table 2.1)!

Statistical tests (usually employed to derive information and conclusions from a database) require a proper sampling design, for example that sufficient data be collected, as well as certain assumptions be met such as data are drawn independently from a population (Cochran 1977). A variety of statistical methods have been used in empirical studies, and some authors have proposed that empirical models be subdivided on the basis of statistical method. Burrough (1989) distinguished between regression and threshold empirical models; these are two dominant techniques in GIS. An example of a regression model is the Universal Soil Loss Equation (USLE), which was developed empirically using plot data in the United States of America (Hutacharoen 1987; Moussa, *et al.* 1990). In contrast, threshold models use boundary values to define decision surfaces and are often expressed using Boolean algebra. For example, dry savannas in the global vegetation biome map cited in 2.3.3 (Hazeltine 1996) are defined using a number of factors including the leaf area index of between 0.6 and 1.5. Other examples of

empirical models where thresholds are used include CART (see 2.3.2) and BIOCLIM.

The BIOCLIM system (see also Chapter 8 by Busby) determines the distribution of both plants and animals based on climatic surfaces. Busby (1986) predicted the distribution of *Nothofagus cunninghamiana* (Antarctic Beech), the Long-footed Potoroo (*Potorous longipes*), and the Antilopine Wallaroo (*Macropus antilopinus*), and inferred changes to the distribution of these species in response to change in mean annual temperature resulting from the 'greenhouse effect'. Nix (1986) mapped the range of elapid snakes. Booth *et al.* (1988) used BIOCLIM to identify potential *Acacia* species suitable for fuel-wood plantations in Africa, and Mackay *et al.* (1989) classified areas for World Heritage Listing. Skidmore *et al.* (1997) used BIOCLIM to predict the distribution of kangaroos.

The basis of BIOCLIM is the interpolation of climate variables over a regular geographical grid. If a species is sampled over this grid, it is possible to model the species response to the interpolated climate variables. In other words, the (independent) climate variables determine the (dependent) species distribution. The climate variables used in BIOCLIM form an environmental envelope for the species. Firstly, the BIOCLIM process involves ordering each variable. Secondly, if the climate value for a grid cell falls within a user-defined range (for example, the 5th and 95th percentile) for each of the climatic variables being considered, the cell is considered to have a suitable climate for the species. Using a similar argument, if the cell values for one (or more) climatic variables fall outside the 95th percentile range but within the (minimum) 0-5th percentile and (maximum) 95-100th percentile, the cell is considered marginal for a species. Cells with values falling outside the range of the sampled data (for any of the climatic variables) are considered unsuitable for the species (Figure 2.3).

In practice, there are other types of empirical models, including genetic algorithms (Dibble and Densham 1993) and geostatistical models (Varekamp *et al.* 1996). These, and other, models do not fit into the regression or threshold categories for inductive and empirical models as proposed by Burrough (1989), so it is considered simpler and more robust not to subdivide empirical models further.

Bonham-Carter (1994) grouped empirical and inductive models into two types, *viz.*, exploratory and confirmatory. This follows the established procedure in statistics of using exploratory data analysis (EDA) followed by confirmatory methods (Tukey 1977). In exploratory data analysis, data are examined in order that patterns are revealed to the analyst. Graphical methods are usually employed to visualize patterns in the data (for example, box plots or histograms). Most modern statistical packages permit a hopper-feed approach to developing insights about relationships in the data.

In other words, all available data are fed in the system, data are explored, and it is hoped that something meaningful emerges.[2] Once relationships are discovered, data driven empirical methods usually confirm rules, processes or relationships by statistical analysis.

---

[2] An approach frowned upon by some scientists who believe that science should be driven by questions and hypotheses that determine which data are collected, and pre-define the statistical methods used to confirm relationships within the data set.

An example is taken from Ahlcrona (1988) who identified a linear relationship between the normalized difference vegetation index[3] (NDVI) calculated using Landsat MSS (multispectral scanner) imagery and wet grass biomass (Figure 2.4).

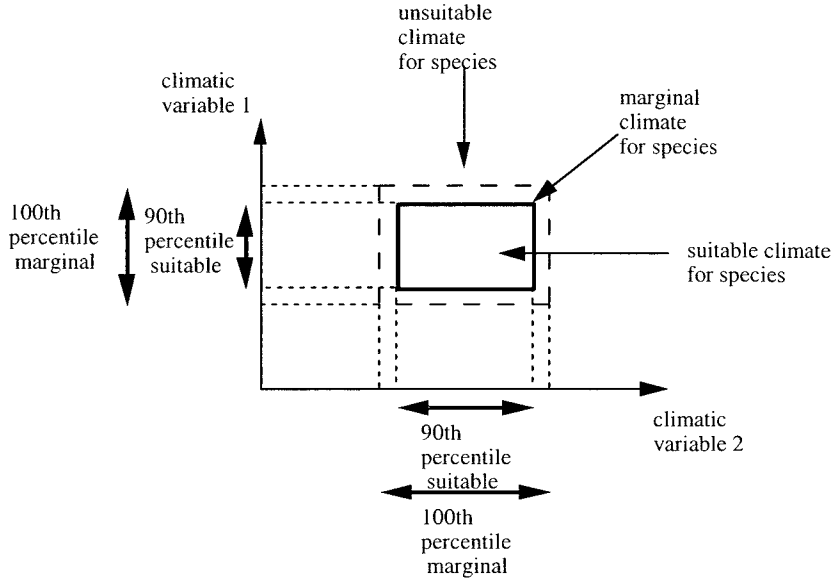

Figure 2.3: Possible BIOCLIM class boundaries for two climatic variables.

Regression was used to calculate a linear model between the dependent (wet grass biomass) and independent (MSS NDVI) variables with a correlation coefficient of 0.61.

A derivative of the Universal Soil Loss Equation (USLE) is the Revised Universal Soil Loss Equation (RUSLE), which is used to calculate sheet and rill erosion (Flacke *et al.* 1990; Rosewell *et al.* 1991). The RUSLE model is an interesting example of a localized empirical model that has been modified (using deduction) and then reapplied in new locations.

[3] NDVI is a deduced relationship between the infrared and red reflectance of objects or land cover.

$$\text{NDVI} = \frac{\text{NIR} - \text{red}}{\text{NIR} + \text{red}}$$

where NIR is the reflectance in the near infrared channel and red is the reflectance in the red channel.

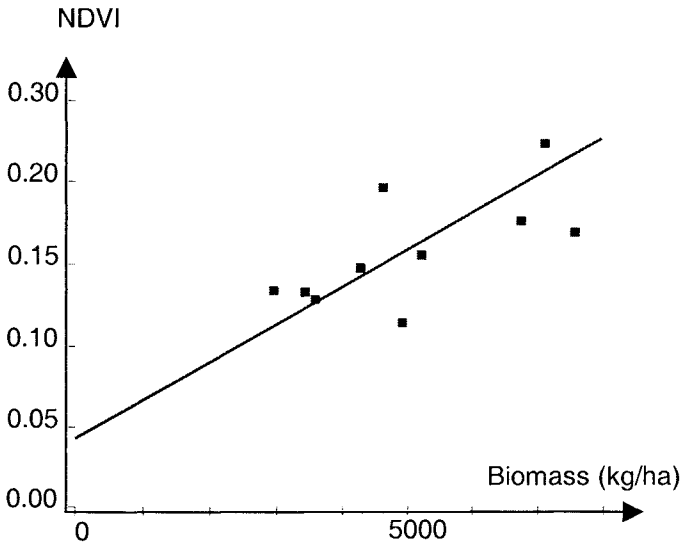

Figure 2.4: The relationship between MSS NDVI and wet grass biomass (from Ahlcrona 1988).

2.4.2 Knowledge driven models

Knowledge driven models use rules to encapsulate relationships between dependent and independent variables in the environment. Rules can be generated from expert opinion, or alternatively from data using statistical induction (such as CART described in 2.3.2). The rules can directly classify (unknown) spatial objects (grid cells or polygons) by deduction, or the rules may be input to an expert system. An expert system is a type of knowledge driven model.

An expert system comprises a knowledge base of rules, a method for processing the rules (the inference engine), an interface to the user, and the (independent) spatial data that are usually stored in a GIS. The structure of the knowledge base largely determines the appropriate inference technique required to generate a conclusion from the expert system. One common method for representing knowledge is the frame (Forsyth 1984), while a method called a probability matrix has also been developed (Skidmore 1989).

The advantage of the frame structure is that knowledge is organized around objects, and knowledge may be inherited from one frame to the next. This is similar to our own 'memory', where knowledge or facts are often remembered through association with other knowledge. The frame structure has been utilized in some expert system applications (Skidmore *et al.* 1992). A second method of representing knowledge in a GIS, called a probability matrix, links the probability of a species occurring at different environmental positions (Skidmore 1989).

Expert systems have been developed from, and given a theoretical foundation based on the field of, formal logic. Following the definitions given in the 'inductive logic' section above (see 2.3.2), formal logic is used to infer a conclusion from facts contained within the knowledge base. For example, given the evidence that a location is a ridge top, and given that if there is a ridge then *Eucalyptus sieberi* occurs, it is possible to infer (conclude) that *Eucalyptus sieberi* is present on the ridge. Using this flow of logic (*modus tollens*), the evidence (E) that a ridge occurs may be linked with a hypothesis (H) that *Eucalyptus sieberi* is present, using an expert system. In expert systems, the evidence (E) is often called an antecedent, and the hypothesis (H) the consequent. In other words, given evidence (E) occurs then conclude the hypothesis (H):

GIVEN	→	E	→	THEN	→	H
		antecedent				consequent
		evidence				hypothesis

where E is the evidence, H is the hypothesis.

Two methods exist for linking the evidence with the hypotheses. The first is forward chaining, where the inference works forward from the evidence (e.g. data represented at a grid cell) to the hypothesis. This is a 'data driven' process, where given some evidence, a hypothesis is inferred from the expert's rules and is an inductive model. The second method is simply the reverse, and is called backwards chaining. In other words, given a hypothesis, the expert system examines how much evidence there is to support the hypothesis. Backwards chaining is obviously a hypothesis driven process, and is akin to the deductive model as described in 2.3.1. But what happens when you do not know with 100 per cent confidence whether the rules are true? For example, *Eucalyptus sieberi* may be present only on some ridges in an area of interest. In such a case you need a method to handle uncertainty in the rules, so that the rules may be weighted on the basis of the uncertainty.

The basis of the Bayes' inferencing algorithm is that knowledge about the likelihood of a hypothesis occurring, given a piece of evidence, may be thought of as a conditional probability. For example, a user may not be certain whether *Eucalyptus sieberi* always occurs on ridges — it may sometimes occur on midslopes. This knowledge may be expressed as the user being reasonably certain (e.g. a weight of 0.9) that *Eucalyptus sieberi* occurs on ridges. By linking the knowledge (weights) with GIS layers, the attributes of the raster cell or polygon are matched with the information in the knowledge (rule) base. The expert system then infers the most likely class at a given cell, using Bayes' Theory.

The expert system was executed and a soil type map predicted by an expert system was plotted for a catchment in south eastern Australia (Skidmore *et al.* 1996). When compared with a soil type map of the same soil classes as prepared by a soil scientist, it was obvious that the two results are similar. 53 soil pits were dug through the area, and 73.6 per cent of the pits were correctly predicted by the expert system. There was no statistically significant difference between the accuracy of the expert system map and the map prepared by the soil scientist, as tested by the Kappa statistic (Cohen 1960).

The Bayesian expert system described above is inductive, as input data from field plots are used to develop rules. It is also possible to develop rules for an

expert system based only on existing knowledge; that is an expert would deduce a model about an environmental system. Such an expert system is deterministic, knowledge based, and of course deductive (see Table 2.1). As noted in 2.2, environmental models may be a mix of categories (Table 2.1).

2.4.3 Process driven models

Process driven models, also known as conceptual models, physically based models, process driven systems, white box models (as opposed to 'black box' because the process is understood) or goal driven systems, use mathematics (often supported by graphical examples) to describe the factors controlling a process. Process driven models are mostly deductive, and to a large extent the features of deductive models described in 2.3.1 are applicable. This class of models describe a process based on understanding and established concepts (prepositions), though parameter values may be estimated from data. In many respects, a process model is a pure science product. However, induction is also frequently used to support the development of process driven models particularly to estimate the value of the model parameters, or to refine the underlying concepts (or factors) on which the model is constructed. The necessity to input detailed parameters that are frequently not available make the task of operating and validating process-models difficult. In practice, most process models are limited to small, relatively simple areas (Pickup and Chewings 1986; Pickup and Chewings 1990; Moore *et al.* 1993; Riekerk *et al.* 1998)

Process models may be static or dynamic with respect to time. Static process models split complex areas of land into relatively homogeneous sub-units, and then use the output from one sub-unit as an input to the next sub-unit (e.g. O'Loughlin 1986). Dynamic process models iterate the process over time and typically attempt to represent a continuous surface.

An example of a process model based on deduction is the Hortonian overland flow model (Horton 1945):

$$Q = (I-F)A \quad (2)$$

Where Q is the surface runoff rate, I is the rainfall intensity and F represents the infiltration rate and A is the catchment area. The generality of Hortonian overland flow has been criticised because:

- surface runoff is dependent on ground conditions, which vary spatially and over time
- that the calculation of surface runoff from comparisons of rainfall intensity and infiltration rates holds good only for very small areas
- that the Hortonian overland flow assumes average conditions over an entire catchment
- the independent parameters (i.e., I, F and A) in equation 2 require induction to estimate their coefficients.

Hortonian overland flow is an example of a lumped empirical model, where the output is calculated for a region based on average input values for the region and is akin in GIS to polygon data structures.

In contrast to lumped models, distributed process models assume that space is continuous, and calculations are made for each element within the area. The elements may be linked in order to estimate the movement between elements (for example, the flow of water between elements in a hydrological model, or the movement of air in a global climate model). Distributed models are developed using raster GIS. The technology makes it simple to spatially and temporally link elements, allowing models to describe the flow of materials or water over a landscape. Such grid based models have been widely developed in hydrology (e.g. TOPMODEL, SHE, ANSWERS).

The problem with distributed models is that they frequently require a large number of input variables of a specific resolution. Remote sensing data, or geostatistics, therefore generate these spatially distributed variables. However, major obstacles exist to the use of distributed models including:

- scaling up (e.g. from points to catchments to continents)
- models based on point data may not be applicable
- input data vary in scale and accuracy (garbage in — garbage out).

As a number of researchers have noted, there is little evidence that complex process models are superior to simple empirical models for many environmental modelling applications (Burrough *et al.* 1996).

Based on the evidence presented in 2.3.1 and 2.4.3, it would be tempting to simplify the taxonomy system and merge 2.4.3 into 2.3.1 (Table 2.1). However, the widespread use of the term 'process driven model' in hydrology, and the fact that process driven models is a hybrid consisting of a concatenation of a number of models (see 2.2), on balance resulted in this category of model remaining separate.

2.5 STOCHASTIC MODELS

If the input data, or parameters of the model itself, are (randomly) varied then the output also varies. A variable output is the essence of a stochastic model.

An example of a stochastic model increasingly used in environmental modelling is the neural network model, commonly implemented using the back-propagation (BP) algorithm. The structure of a typical three-layered neural network is shown in Figure 2.5; however networks may easily be constructed with more than three layers.

To train a network, a grid cell is presented with values derived from a GIS. For example, in Figure 2.5, the values for a cell may be elevation equal to 0.8, aspect equal to 0.3 and SPOT visible band equal to 0.5 (note the input values are normalized to range between 0 and 1). Simultaneously, an output class is presented to the network; the output node has an associated output, or target, value. In other words, an output class, such as water, may be assigned to an output node number (for example node 3 in Figure 2.5), and given a target value of, for example, 0.90. Clearly, the neural network is trained using induction (see 2.3.2).

The BP algorithm iterates in a forward and then in a backward direction. During the forward step, the values of the output nodes are calculated from the input layer. Phase two compares the calculated output node values to the target (i.e. known) values. The difference is treated as error, and this error modifies connection weights in the previous layer. This represents one epoch of the BP algorithm. In an iterative process, the output node values are again calculated, and the error is propagated backwards. The BP algorithm continues until the total error in the system decreases to a pre-specified level, or the rate of decrease in the total system error becomes asymptotic. Prior to the first epoch, the neural network algorithm assigns random weights to the nodes and introduces the stochastic element to the neural network model.

Node weights are an interesting neural network parameter to adjust (Skidmore *et al.* 1997). An experimental set up was chosen that produced an accurate map of forest soil, and the network parameters were noted (Skidmore *et al.* 1997).

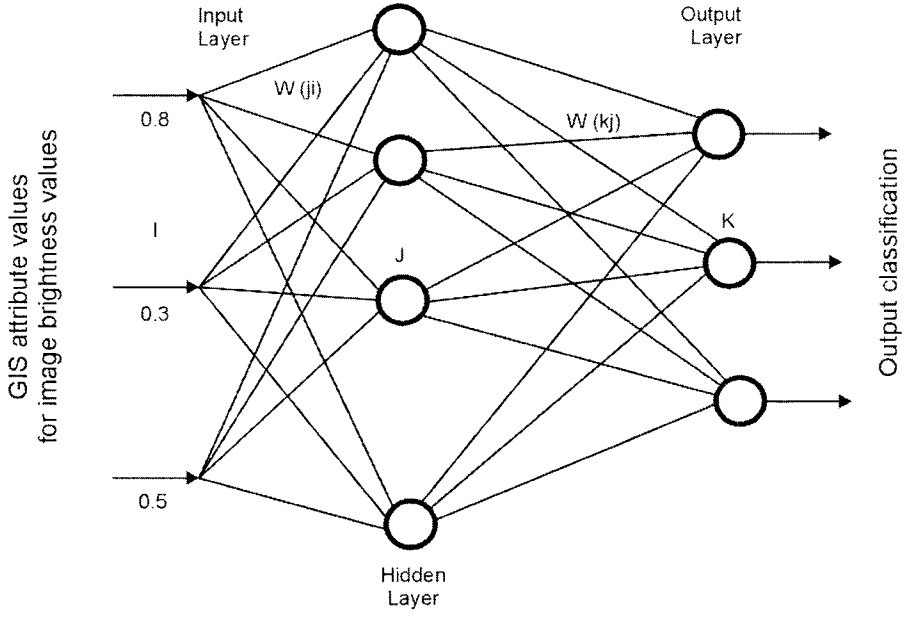

Figure 2.5: Neural network structure for the BP algorithm.

A map of the classes predicted by the neural network shows the classification was reasonable. All network parameters were then held constant (e.g. number of learning patterns, number of nodes, number of layers, learning rate, momentum etc.), except that the starting weights were randomly adjusted by ± 5%. Five different maps were produced, with each map having slightly different starting weights. Even though the accuracy of the training and test data is similar (ranging from 90 to 97 per cent training accuracy and 42 to 55 per cent test accuracy), the spatial distribution of the classes was quite different. Such a variation in mapping accuracy highlights the stochastic nature of neural networks.

Stochastic models have also been developed where the average (and variance) value for many (usually random) events are calculated. For example, randomly selecting the input data from a known population distribution, and then noting the range of output values obtained, indicates the possible range of output values, as well as the distribution of the output.

2.6 CONCLUSION

A taxonomy of GIS models has been presented with examples from various application fields in the environmental sciences. Some of the model types have had limited application in the spatial sciences. Other model types are widely applied, such as inductive empirical models.

As highlighted in this chapter, many environmental applications combine two (or more) categories (as detailed in Table 2.1), though the modelling process may appear seamless to a user. In order to use the taxonomic system, a user must deconstruct the application, and identify the taxonomic categories. This provides the user with a framework to clarify thoughts and organize material. In other words, a user, or model developer, can easily identify similar environmental models that may be applied or adapted to a problem. As taxonomy also gives an insight to very different models, a taxonomy hopefully helps in transferring knowledge between different application areas of the environmental sciences.

2.7 REFERENCES

Ahlcrona, E., 1988, *The impact of climate and man on land transformation in central Sudan.* PhD thesis. Lund University, Lund, Sweden.

Aleksander, I. and Morton, H., 1990, *An introduction to neural computing.* London, Chapman and Hall.

Bonham-Carter, G.F., 1994, Methods of spatial data integration for mineral mapping potential. *Proceedings Seventh Australasian Remote Sensing Conference,* Melbourne, Aust. Soc. for Remote Sensing and Photogrammetry.

Booth, T.H., Nix, H.A., Hutchinson, M.F., and Jovanovic, T., 1988. Niche analysis and tree species introduction. *Forest Ecology and Management* **23**:47–59

Brieman, L., Friedman, J.H., Ollshen, R.A., and Stone, C.J., 1984, *Classification and Regression Trees.*Belmont, CA, Wadsworth.

Britannica, E., 1989, *The New Encyclopaedia Britanica.* Chicago, Encyclopeadia Britannica.

Burrough, P.A., 1989, Matching spatial database and quantitative models in land resource assessment. *Soil Use and Management* **5**: 3–8

Busby, J.R., 1986. A bioclimatic analysis of *Nothofagus cunninghamia* in south eastern Australia. *Australian Journal of Ecology* **11**:1–7

Burrough, P.A., van Rijn, R. *et al.*, 1996, Spatial data quality and error analysis issues: GIS functions and environmental modelling. In: Goodchild, M.F., Steyaert, L.T., Parks B.O., (ed.). *GIS and Environmental Modelling: Progress and Research Issues.* Fort Collins, CO, GIS World Books, 29–34.

Cochran, W.G., 1977, *Sampling Techniques.* New York, Wiley.

Dibble, C. and Densham, P., 1993, Generating interesting Alternatives in GIS and SDSS Using Genetic Algorithms. In *Proceedings of the GIS/LIS '93 Conference*, Minneapolis, ACSM-ASPRS-URISA-AM/FM.

Flacke, W., Auerswald, K. *et al.*, 1990, Combining a Modified Universal Soil Loss Equation with a Digital from Rain Wash. *Catena*, **17**: 383–397.

Forsyth, R., 1984, *Expert Systems: Principles and Case Studies*. London, Chapman and Hall.

Hazeltine, A., 1996, *Modelling the vegetation of the Earth*. PhD Thesis. Lund University, Lund, Sweden.

Horton, R.E., 1945, Erosional development of streams and their drainage basins: hydrophysical approach to quantitative morphology. *Bulletin of the Geological Society of America*, **56**: 275–370.

Hutacharoen, M., 1987, Application of Geographic Information Systems Technology to the Analysis of Deforestation and Associated Environmental Hazards in Northern Thailand. In *Proceedings GIS '87, San Francisco, California.* Washington, American Society of Photogrammetry and Remote Sensing.

Kettle, S., 1993, CART learns faster than knowledge seeker. In *Proceedings of the Conference on Land Information Management, Geographic Information Systems and Advanced Remote Sensing* **2**: 161–171. School of Surveying, UNSW, P.O. Box 1, Kensington, NSW 2033, Australia.

Kosko, B., 1992, *Neural networks and fuzzy systems: a dynamical systems approach to machine intelligence.* Englewood Cliffs, New Jersey, Prentice Hall.

Kramer, M.R. and Lembo, A.J., 1989, *A GIS Design Methodology for Evaluating Large-Scale Commercial Development.* In *Proceedings GIS/LIS '89*, Orlando, Florida, ASPRS, AAG, URPIS, AM/FM International.

Mackay, B.G., Nix, H.A., Stein, J.A., Cork, S.E., and Bullen, F.T., 1989. Assessing the representatives of the wet tropics of north Queensland world heritage property. *Biological Conservation* **50**:279–303

Maidment, D.R., 1993, GIS and hydrologic modelling. In: Goodchild, M.F., Parks, B.O., and Steyaert, L.T (eds.). *Environmental modelling with GIS*. Oxford Oxford University Press.

Moore, D.M., Lees, B.G. and Davey, S.M., 1990, A New Method for predicting vegetation distributions using decision tree analysis in a GIS. *Environmental Management*, **15**: 59–71.

Moore, D.M, Lees, B.G., *et al.*, 1991, A New Method for Predicting Vegetation System. *Environmental Management,* **15**: 59–71.

Moore, I.D., Turner, A.K. *et al.*, 1993, GIS and land-surface-subsurface process modelling. In: Goodchild, M.F., Parks, B.O. and Steyaert, L.T. (ed.). *Environmental modelling with GIS*. Oxford, Oxford University Press.

Moussa, O.M., Smith, S.E. *et al.*, 1990, GIS for Monitoring Sediment-Yield from Large Watershed. In *Proceedings of GIS/LIS '90*, Anaheim California, ASP&RS, AAG, URPIS and AM/FM International.

Nix, H., 1986. A biogeographical analysis of Australian elapid snakes. In: Longmore, R., (ed.), *Atlas of Elapid Snakes of Australia*, Bureau of Flora and Fauna, Canberra, Australian Government Publishing Service, 4–15.

O'Loughlin, E.M., 1986, Prediction of surface saturation zones in natural catchments by topographic analysis. *Water Resources Research*, **22**: 794 – 804.

Pao, Y.H., 1989, *Adaptive pattern recognition and neural networks*. Reading, Addison-Wesley.

Pickup, G. and Chewings, V.H., 1986, Random field modelling of spatial variations in erosion and deposition in flat alluvial landscapes in arid central Australia. *Ecol Model*, **33**: 269–69.

Pickup, G. and Chewings, V.H., 1990, Mapping and Forecasting Soil Erosion Patterns from Landsat on a Microcomputer-Based Image Processing Facility. *Australian Rangeland Journal*, **8**: 57–62.

Quinlan, J.R., 1986, Introduction to decision trees. *Machine Learning*, **1**: 81–106.

Richards, J.A., 1986, *Remote Sensing – digital analysis*. Berlin, Springer-Verlag.

Rietkerk, M., 1998, *Catastrophic vegetation dynamics and soil degradation in semi-arid grazing systems*. PhD Thesis. Wageningen Agricultural University, Wageningen, The Netherlands.

Rietkerk, M., Ketner, P., Stroosnijnder, L., and Prins, H.H.T., 1996, Sahelian rangeland development: a catastrophe? *Journal of Range Management*, **49**: 512–519.

Rosewell, C.J., Crouch, R.J. *et al.,* 1991, Forms of erosion. In: Charman, P.E.V. and Murphy, B.W. *Soils – their properties and management: A soil conservation handbook for New South Wales*. Sydney, Sydney University Press.

Skidmore, A.K., 1989, An expert system classifies eucalypt forest types using Landsat Thematic Mapper data and a digital terrain model. *Photogrammetric Engineering and Remote Sensing*, **55**: 1449–1464.

Skidmore, A.K., 1990, Terrain Position as Mapped from a Gridded Digital Elevation Model. *International Journal of Geographical Information Systems*, **4**: 33–49.

Skidmore, A.K., Baang, J. and Luchananurug, P., 1992, Knowledge based methods in remote sensing and GIS. *Proceedings Sixth Australasian Remote Sensing Conference*, Wellington, New Zealand, **2**: 394–403.

Skidmore, A.K., Ryan, P.J., Short, D. and Dawes, W., 1991, Forest soil type mapping using an expert system with Landsat Thematic Mapper data and a digital terrain model. *International Journal of Geographical Information Systems*, **5**: 431–445.

Skidmore, A.K., Gauld, A., and Walker P.A., 1996, A comparison of GIS predictive models for mapping kangaroo habitat. *International Journal of Geographical Information Systems*, **10**: 441–454.

Skidmore, A.K., Turner, B.J., Brinkhof, W. and Knowles, E., 1997, Performance of a neural network mapping forests using GIS and remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, **63**: 501–514.

Tukey, J., 1977, *Exploratory data analysis*. Reading, Addison-Wesley.

Varekamp, C., Skidmore, A.K. *et al.*, 1996, Using public domain geostatistical and GIS software for spatial interpolation. *Photogrammetric Engineering and Remote Sensing*, **62**: 845–854.

Walker, P.A., and Moore, D.M., 1988, SIMPLE: An inductive modelling and mapping tool for spatially-oriented data. *International Journal of Geographical Information Systems*, **2**: 347–364.