



Models, measures and messages: an essay on the role for induction

M. B. Dale

Australian School of Environmental Studies, Griffith University, Nathan, Qld 4111. Australia.

Phone: +61 7 33714414, Fax: +61 7 38705681, E-mail: m.dale@mailbox.gu.edu.au

Keywords: Falsification, Model classes, Model selection, Null hypothesis, Prediction, Valorisation.

Abstract: In this essay, I examine the role of induction in developing vegetation models. Falsification is a necessary component of model building but is not itself sufficient. Induction provides a necessary complement and one that dethrones the null hypothesis from its privileged state. After examining the role of description and environment, I examine several possible criteria useful for valorising models so that we may choose the 'best'. These criteria include fit, simplicity, precision and interest. Predictability, which is given overwhelming importance in a falsification approach, is found to be ambiguous. It may be obtained by using multiple models without regard to the processes active in the real system. In addition movement towards a model which does reflect the 'real' processes can result in loss of predictivity. Finally, some comments are made on what we can infer and how this relates to our understanding of living systems.

Abbreviations: MML- Minimum message length, GUHA - General Unary Hypothesis Automaton.

*Between the idea and the reality,
between the motion and the act,
falls the shadow
T. S. Eliot
Four Quartets*

Introduction

Models (theories, hypotheses) pervade science. Some are mental, others physical, some are deterministic, others stochastic, and some are qualitative and others quantitative. In this paper, I want to examine statistical models which are derived from observed data by induction. For some, the fundamental proposition in science is the use of falsification (Popper 1968). Several authors have argued strongly that ecology will progress only through the testing of clear null hypotheses (Simberloff 1980, Wilson et al. 1994, Wisheu and Keddy 1992). However falsification is certainly not the only criterion, a fact known to Cicero, who comments in "*De Natura Deorum*", '*Would that I could discover truth as easily as I can uncover falsity*'. The call for simple, clear hypotheses to be tested may not be answerable in the complexity and murk of ecological situations and will depend markedly on being able to choose a method of suitable method of description of the system, or more succinctly an alphabet.

Other views suggest that there are several stages in the process of scientific endeavour; for example, the four stages of saturation, incubation, illumination and verification attributed to Von Helmholtz and Poincaré. The problem here is that we are not sure how illumination proceeds from incubation! This can be alternatively interpreted as meaning that we first make observations, in which we seek patterns. These patterns form the basis of our models, which are then subject to verification. Falsification applies only to the last stage, while induction is pertinent to the detection of pattern and perhaps to the formation of models. So clearly the induction of patterns is a crucial stage, and falsification contributes nothing to it!

A more modern phraseology was adopted by Hájek and Havránek (1977), who argued that '*the logic of discovery is twofold: to choose a language in which to express the hypothesis and to choose a satisfactory sentence which solves it*'. Unfortunately they provide no definition of 'satisfactory' nor do they suggest how the language choice might be made. In fact they are asking for a means of determining a class of languages based on some specific lexicon which is acceptable and within that class the specific language which best fits the data we have (cf. Dale 1980). Thus, if we seek to decide between individualistic and community-unit concepts of vegetation (Shipley and Keddy 1987), falsification is largely irrele-

vant. The problem is one of trying to distinguish between two classes of models, not individual models within a class. —*perhaps a comment here on the importance of perception in model-building and model-choice*—. To do this we need some means of valorising models, and classes of models, which falsification does not supply.

The arguments for falsification have about them the same scent of illusion as that attributed by Williams (1972) to the scientific paper. We first state a null hypothesis, triumphantly demonstrate its falsity and present, as from a hat, an alternative which we claim is acceptable, though, in due modesty, we may add some expression of possible dubiety! Where did the null hypothesis come from? Why is it so special? From where did the alternative hypothesis appear? Is there only a single such alternative and, if not, why choose a particular one? In complex problems such as are found in ecology, what we need is some means of constructing models as well as disposing of them by falsification, and such construction is a problem in induction not deduction.

There are more subtle effects as well. In the development of any science, as in other human activities, there is a tendency for people interested in a subject to develop their own interpretative community (Carley and Palmquist 1992). Such a group builds up, in its own field of endeavour, a series of propositions that enable its members to communicate effectively, inventing its own jargon, and perhaps initiating its own journal. Members are more or less agreed about the nature of their topic and what form interesting questions should take. What is less commonly realised is that, in so doing, it is also the case that propositions not within this community are regarded as irrelevant, second class, even indicative of loose thinking and hence ignorable. As an example, the case for conservation was so regarded until quite recently. Proposals for evaluating the ecological impact of any development were regarded as trivialities propounded by weak and woolly-minded thinkers, pejoratively termed ‘Greenies’. In some cases, as Thucydides noted, this results in ‘*words become fair phrases used to cover guilty ends*’. Biology provides an excellent example in Bishop Wilberforce’s attack on Darwin¹!

Falsification is, then, a necessary component of scientific method but it is not itself sufficient. **It tells us what is not, but not what is.** There may be several different classes of alternatives, any of which would cause rejection of the null hypothesis (see e.g., Paluš 1996a). Overall it is a pessimistic approach (Fisher 1992), based on accepting only propositions for which there is positive evi-

dence in favour. In contrast, for building models with sparse and noisy data, we may profit by employing optimistic approaches, which reject only propositions for which there is strong negative evidence. In this paper I suggest that at least three stages are needed; description, pattern-formation and verification. I shall be primarily concerned with pattern finding and assessment of models, with the comparison of multiple models, and with the role of prediction in these tasks. I shall not address verification.

Focus of attention and description

One of the very basic questions is determining what type of hypothesis space must we assume so that we are assured of finding a developmental system consistent with our observations? In other words, the hypothesis space itself is dependent on the input data. By changing those data we implicitly change the hypothesis space as well, and it is desirable that we change it to make it both simpler and smaller! (cf. Domingos 1996). There are two aspects involved here:

- How to determine the descriptors to be used to describe the system since these determine the space in which the system is embedded? This also involves selecting the scale and precision of measurement.
- How to determine the focus of attention, which governs the specific objects examined, what they are, where they are, when they are and what sizes and shapes we take them to have? This also involves discriminating characteristics of these objects from those of the environment, thus separating the two.

Such selection questions will generally be answered in the light of some general theory. It is possible to assess how effective such selection may be, in comparison with some other choice but analysis cannot commence without some initial choices.

In vegetation studies, the selection of the ‘things’ in which patterns are sought is itself a difficult task. The importance of scale in model inference is well illustrated by May’s (1995) examination of measles epidemics. Indeed, Hogeweg (2002) comments that “processes do not, in biotic systems, operate in isolation and the existence of entanglement at different time and space scales does not need explanation, being there by default. Ignoring it by segregating time and space scales is simply a modelling artefact”. In vegetation studies we often use samples which exhibit dependence, spatial, temporal or both. Such

1 This notion also seems related to Bacon’s *idola tribus*, the boundaries to our understanding imposed by human limitations.

dependence must be accounted for in any analysis (see eg. Wallace, 1998, and Dale et al. 2002a). Such dependence is related to questions of scale.

Description

We have first to identify a descriptive language. We must first recognise that choosing descriptors for vegetation carries semantic implications. Riddle and Hafner (1999) argued recently that species are not the most appropriate units for all purposes. Taxonomic species import phylogenetic information, whereas structural descriptions more likely reflect local habitat factors and biogeographic or other descriptors will carry still other information. Thus, in place of the traditional floristic descriptions, Dale and Barson (1989, see also Dale et al. 1988), used cluster labels in grammatical studies of vegetation, Dale et al. (1984) used structural features of the canopy, Noble and Slatyer (1980) isolated features which were deemed to have dynamic importance (vital attributes) while Webb et al. (1976) found floristic descriptions desirable for simple communities although structural form was effective in complex ones. Pillar (1999a) has even considered how we might optimise the description by selection within the class of functional types, using correlation between plant traits and environment.

Composite descriptions are also possible; Lux and Bemerle-Lux (1998) address the equilibration of time responses between trees and ground flora by using mixed descriptions, structural for the former, floristic for the latter. But how might we compare the effectiveness of these alternatives, which differ in their semantics? If we can identify a value for a model then such comparisons could be made in terms of the amount of pattern captured by the various choices. This can be measured using the difference between the value of the optimal model and the value of the null model, appropriately normalised although other suboptimal models could be used if there were some specific reason for so doing. The larger the difference, the 'better' the optimal model.

Such considerations also apply wherever dissimilarities are involved, as shown by Watanabe's (1969) '*ugly duckling*' theorem. This proves that the investigator must choose the attributes for describing a system, for without such a selection *all* pairs of objects have the *same* dissimilarity. This means that the adoption of lazy, case-based approaches (see Lekkas and Avouris 1994) instead of eager, rule-based systems does not avoid the choice, although the former may be preferred in that the choice can be made to vary depending on where you are in the space.

Distinguishing the environment

I do not propose to provide here a detailed examination of all aspects of determining the focus of attention. Instead, I shall concentrate on one major feature of this choice, the distinction between the system and its environment.

Herman and Rozenberg (1975) indicate that if the environment is unspecified then we can *always* choose a grammatical model (specifically a D1L parallel grammar) simply by assigning any abnormalities and infelicities to environmental events. Brokaw and Busing (2000) recently emphasised the importance of contingent effects in maintaining diversity (see also Niven's, 1992, malenticities). Antonelli (1990) goes further and suggests that it is *desirable* to export complexity from the ecosystem into the environment. He replaced a complex graph of population dynamics by more or less straight lines by embedding his data in Finsler space (which is a generalisation of Riemannian space). In other words, he had a very twisted environmental space, but this permitted the population dynamics to be expressed simply. Such a suggestion relates to the problems of selecting an appropriate metric for the space in which we embed our model (cf. Dale 1994).

Environment is not everything, however. Austin (1970) argued that interpretation of patterns in vegetation will be accomplished by developing environmental correlates, but we know that autopoietic generation of patterns is possible (Boerlijst and Hogeweg 1991), so that not all patterns in vegetation are necessarily environmentally determined. Dale and Hogeweg (1998) and Dale (1999) identified three distinct dynamic types of vegetation without any initial environmental variation, which indicates that there may be no single model suitable for modelling all vegetation. Once patterns have formed, selection can take place, which may *result* in environmental differentiation (Savill et al. 1997). If the vegetation patterns precede the environmental differences, Austin's (1970) reliance on environment as a means of interpreting vegetation patterns becomes suspect.

Measuring models

I shall confine myself to the following six topics, each of which has a general application to a variety of models:

- degree of fit to observed data;
- consistency of estimation;
- simplicity and/or complexity of the model;
- precision. of measurement;

- prediction and causality; and
- interest and ‘actionability’.

It is also necessary to be able to compute the model. Posse (1995) has shown how projection pursuit methods were handicapped initially by the use of inexact and inappropriate algorithms. But there can be more subtle effects arising from the choice of computational procedure. Howard and Oakley (1994) suggested the use of genetic programming for finding patterns in short time series. Pagie and Hogeweg (1997) show that by using sparse evaluation in a genetic algorithm, a more generalisable result can be obtained than can be obtained using a global evaluation; the manner of the search affects the quality of the results.

The models with which I shall be concerned are statistical models. This class includes such different types as regression, decision trees and graphs, Linear Structural models, unsupervised clustering, ordination, directed acyclic graphs and Markov models. It is perhaps obvious that there can be no universally applicable method of determining the ‘best’ model from observational data. Consider a series starting 2, 4, 6, 8. What number do we expect to follow? Schmidhuber (1994) shows two models which provide different answers. For the i th position return $2i$ (which gives 10 as the next number) or use $i^4 - 10i^3 + 35i^2 - 48i + 24$ (which gives 34!). There are, of course, still other expressions that might lead to still other answers yet still fit the observed series. We need criteria that permit us to rank the models so we can choose between them.

Degree of fit

The quality of fit (also known as coverage) is perhaps the commonest criterion we use to assess a model. Many papers contain references to R^2 values, many applications use forms of stepwise regression where selection is based on ‘better’ fit while ordinations commonly use ‘proportion of captured variance’ to determine dimensionality. But fit is a dangerous ally, for if we add sufficient complexity to our model we can fit anything perfectly! Thus, if we have n observations, a polynomial of degree $(n-1)$ will fit them perfectly. This is a major problem with methods such as Ivakhnenko’s (1971) which use extremely high order polynomials. Wallace (1996) adds a variant of fit which he terms *plausibility a priori*. He suggests that a theory should be less favoured if it is at odds with prior knowledge. This means data should fit not only the given data but also any previous data that seem relevant. And if prediction is important, then the model must fit future data as well! Obviously fit is of interest – it is not much use having a result which does not fit the observations at all!

But equally obviously, it cannot be the sole criterion of model value.

Consistency of estimation

It is desirable that the estimates we obtain for parameters are consistent and unbiased. Usually we adopt maximum likelihood as the means of selecting estimators. Yet, it is well known that the maximum likelihood estimator for variance is biased; the correction is simple in that case involving replacement of the denominator n by $(n-1)$. There are other estimators that are unbiased for this case (minimum message length, for example), though they may have other problems. It is also known that for estimating large numbers of parameters, maximum likelihood estimation does not perform very well. We must therefore take care that our choice of estimation procedure is appropriate.

As an example, consider multiple factor analysis. Jöreskog (1966) presented a maximum likelihood estimation procedure that has been widely used. However, Wallace (1995) has examined this and compared the results with an alternative estimation procedure using minimum message length (MML) principles (Wallace and Freeman 1987; see also Wallace 1996, Wallace and Dowe 2000). Wallace (1995) found that the maximum likelihood procedure had several drawbacks. First, it does not consistently estimate the factor scores, since such estimates are *conditional* on the estimates of factor loadings and of dimensionality. Second, maximum likelihood did not perform well in estimating the dimensionality. Third, even in those cases where the dimensionality was estimated correctly, the results were biased, with axes often set parallel to one or other of the variables and thus essentially useless! In contrast the MML estimation was consistent, could estimate the factor scores unconditionally, gave on average better estimates of the number of factors and had no tendency to identify useless axes.

Consistency can assist us in choosing between classes of models. For example, most clustering methods identify crisp clusters with sharp boundaries (segmentation) although some permit fuzzy assignments of various kinds. In Figure 5 of Dale et al. (2001), we show two overlapping clusters and remark that the parameters for these clusters will be inconsistently estimated if a crisp separation is imposed. However, if a fuzzy assignment is used, whereby the things forming the clusters are but partially assigned to clusters, then consistent estimation is possible. Obviously, this is a considerable advantage and suggests that, for vegetation study, fuzzy clustering should be preferred to crisp. Since the samples we take of vegetation can easily result in elements of several clusters being present in

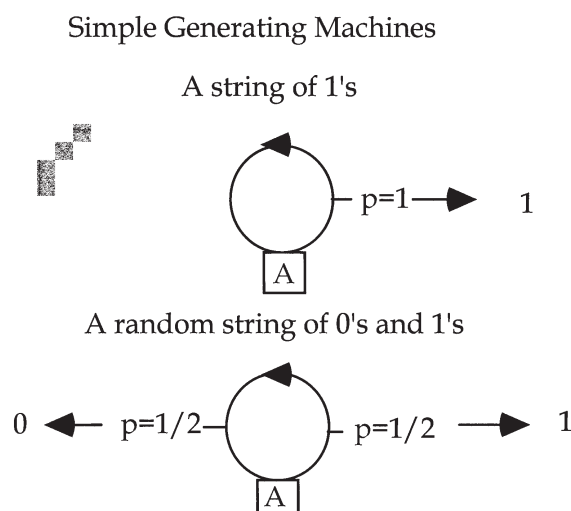


Figure 1. A machine with a single state A emits a symbol with probability p and remains in state A. The first machine emits only the symbol 1, while the second emits symbols 0 or 1 randomly.

a single plot (Dale 1988) we should generally avoid crisp clustering solutions.

Simplicity and complexity

Since complex models improve fit, several authors (e.g., Iba et al. 1988) have suggested a trade off between simplicity and degree of fit. This is an appeal to William of Occam's famous razor² and does not enjoy universal support (see Webb 1994, 1996). Indeed, Wittgenstein (1995) states that '*It is clear there are no grounds for believing that the simplest course of events will really happen*'. Domingos (1999) identifies two aspects to using the razor. One is for understanding and this is generally a good reason. The other is to increase accuracy (or generalisability) at which it may fail. But predictivity is not independent of simplicity and van den Bosch (1997) provides strong support for seeking simple solutions for prediction problems.

To make use of the razor we need a measure of simplicity or its inverse, complexity. There seem to be two major approaches to measuring complexity. Kolmogorov (1965) introduced a measure of information that treats uniformity as the simplest situation and randomness as the most complex and this has been widely adopted. Kol-

mogorov's information measure cannot be calculated directly but can be approximated (see Wallace and Dowe 2000, who base their estimation of complexity on prior probabilities).

Kolmogorov's information complexity is not the only such measure. Gunther et al. (1994) and Wackerbauer et al. (1994) examine other various choices. Lloyd and Pagels (1988), Gell-Mann (1994) and Grassberger (1991) concentrate on complexity of a model and not the observed data, emphasising time required or resources necessary to accomplish a computation. This turns out to be related to the difference between coarse- and fine-grained entropy; simple systems have small differences, complex ones have large differences. This view is largely shared by Crutchfield and Shalizi (1999) although the details are different. Indeed, there is a well-known theoretical and measurable relationship between the entropy and Kolmogorov complexity of systems (Anand and Orlóci 1996).

One result of this model-based approach is that randomness is no longer the most complex situation. Consider binary patterns being generated by two finite state machines, each of which has only one state. The first emits a value, say 1, with probability 1 while the second emits a value of 1 with probability of 0.5, and a value of 0 with probability 0.5. This is illustrated in Fig. 1. Both models are very simple, although one involves randomness. Calculating Kolmogorov's information for the series produced, the first is extremely simple whereas the second is extremely complex. Crutchfield and Shalizi (loc. cit.) are concerned with prediction and want the shallowest past for causal states that provide predictions of the future. But whichever view we adopt, the complexity is estimable.

Choice between these alternatives may rest with the observation that complexity is in part a function of the observer (Löfgren 1974) and what is appropriate in one circumstance is not in another. For example, many studies in phytosociology concentrate on deriving a static taxonomy, whereas we may desire a dynamic representation.

Precision and fit

Much data in vegetation science is measured using coarse scales, such as ordered categories. When estimating parameters we therefore need to maintain a balance between excessive precision and lack of fit due to coarse-

2 William of Occam might have said, but probably did not, something like '*Entia non sunt multiplicanda praeter necessitatem*' which can be translated as '*No more things should be presumed to exist than are absolutely necessary*'. Similar statements have been recorded much earlier. Occam's razor still remains a heuristic, advice to aid us in discovery and there certainly exist situations where it will fail. The real question is whether those situations are sufficiently common for us to reject the principle entirely. Domingos (1999) argues against this possibility, van den Bosch (1997), among others, supports it.

ness of estimates. Viswanathan et al. (1999) note that when segmenting a series, there is no point in specifying the cut-points to a precision greater than can be estimated. Georgeff and Wallace's (1984) MML-based procedures encode parameters and real numbers with a finite precision only, based on a Dirichlet tessellation. An optimal precision is determined as a trade-off between the cost of encoding extra precision and the cost of encoding the data using a suboptimal hypothesis due to rounding. Other arguments against overmuch precision deprecate the sensitivity of a model to extremely precise values for parameters

Paluš (1996b) has examined the case of quantised variables and provided several estimators for their characterisation, though most of them still require relatively long and noise-free observations, a situation not likely to occur in vegetation studies. Oates and Jensen (1998) have shown that simplicity can be compromised if the dataset used for induction is very large so there may be a conflict. It is odd, perhaps, to find too much data a problem!

Prediction and cause

Prediction is generally regarded as of great importance, since it is required for applying falsification, but it is not without its own problems. Yamada and Amoroso (1971) distinguished two sorts of models, termed 'behaviourally isomorphic' and 'structurally isomorphic'. In behaviourally isomorphic models we adopt a black-box approach, being concerned solely with the accuracy of the output. Such a view generally restricts generalisability, that is the range of prediction, permitting interpolation but not extrapolation. In contrast, structurally isomorphic models seek to capture, in some idealised way, the 'real' structure of our system and as such can be used for extrapolation as well (see Dale 1970).

Behavioural models

The 'black box' approach can work well as a means of making predictions, until we consider the difficulties resulting from model uncertainty. A single behaviourally isomorphic model is not the optimal method of making predictions, and several alternative models may give very similar results. In such circumstances, we can do better by introducing model averaging (Mac Nally 2000, Hoeting et al. 1998) which uses a weighted average of predictions from all possible models, or at least some large sample of them. Usually the weights are derived from the posterior probabilities. But how do we assign blame or credit to any particular model or attempt to falsify it? If prediction is our sole aim we should certainly seek the optimal result, but this means falsification is not applicable at all!

Model averaging can also be useful when estimating dissimilarities. Yee and Allison (1993) found that better estimates of the evolutionary distance between organisms could be derived from model averaging than from a single 'best' model. The superiority was maintained over a wide range of values of the actual dissimilarity.

Structural models

Farrands (1990) has given an account of the traditional view of how models are developed. We start with a general model then cycle between complication to take account of other factors, thereby gaining in precision but losing in generality and further generalisation to regain elegance and understanding. If we can map the phenomena of the perceived world onto a plane where it can be manipulated intellectually, then we say we understand the phenomenon. If we cannot, we review matters and may introduce new explanations. If the model explains our previous experience, that is good. If it leads necessarily to a new testable hypothesis that is better; if it leads to a new view of the world that is best. But note that in moving from behavioural to structural isomorphism we may temporarily lose quality of prediction (Yamada and Amoroso 1971).

Wallace (1996) has examined the relationship of prediction and induction and concludes that using a specific method of valuing models, the MML principle, 'minimises the degree to which future data will surprise us'. A theory is to be preferred if it is applicable to a wide range of, possibly yet unseen, instances. A theory which is restricted to just the instances in known data is not of much use. We do not, after all, set much store by a degree $(n-1)$ polynomial as a model of n observations! Wallace further points out that if we are seeking a minimal cost prediction, induction is not suitable. The conventional Bayesian minimum-cost estimation avoids commitment to any model of the real world at all. However, it is often infeasible and the true cost functions may be unknown. In this case, induction can be used but it will rarely be optimal for any specific cost function. This situation can sometimes be improved if we average over several 'reasonably good' models but this again shifts us away from structural isomorphism..

I suggest that predictivity is not a primary characteristic of models which provides a unique valuation, a view supported by Provost et al. (1998) for different reasons. Good predictions can be obtained from models that are far from structurally isomorphic with the real system, relying instead on the occurrence of strong correlations. This is not to say that prediction is uninteresting, only to regard its role in structural modelling as belonging to a

O	X	O	X	O
X	O	X	O	X
O	X	?	X	O
X	O	X	O	X
O	X	O	X	O

Figure 2. Dissimilarity and pattern prediction.

later stage in the process of model forming and verifying. As our structural model becomes perfected, the quality of its predictions will generally improve, but not monotonically. As an example consider Figure 2. Attempting to predict the unknown value using its immediate neighbours will fail and we shall only obtain an accurate prediction once we have grasped the periodic nature of the pattern.

Causality

Mac Nally (2000) identified the search for structural isomorphism with search for cause (or explanations) while Domingos (1998) suggested that it was desirable to combine prediction with causality, thus emphasising their distinctiveness. Hume (1999) provided a definition of deterministic causation as follows: '*We may define a cause to be an object, followed by another, and where all the objects similar to the first kind are followed by objects similar to the second*'. Recent logical approaches would probably rephrase this by a counterfactual definition that if A causes B then without the event A we never have event B; i.e., *not A implies not B*. In fact, we cannot realistically assume that any such simple causal relationships exist in ecology. Pagie and Hogeweg (1997) indicate that biologically there may be several 'causes' for an observed phenomenon. For example, there can be many-one mappings from genotypes to phenotype; and in ecology there may be several different mechanisms for the maintenance of high diversity.

Such deterministic definitions are unsuitable for stochastic modelling such as path analysis. In contrast to the determinism of Hume's definition, probabilistic causation characterises the relationship between cause and effect using the tools of probability theory. The central idea behind these theories is that causes raise the probabilities of their effects, all else being equal. But some causes can reduce the immediate probability. A golfer slicing a ball is not expecting to improve the chances of getting an

eagle, but if the ball hits a tree and bounces into the hole it still may succeed. The usual cause-effect relationship is defeated by a contingent event.

The usual methods of examining causality involve linear structural models. Procedures such as path analysis (Wright 1934) rely on correlations, which are known to be deficient in capturing causality. The techniques of path analysis can be used to estimate the strength of various connections, given a particular graph structure showing inter-relationships between variables. Hájek and Havránek (1977) in the GUHA programs, rely on the concept of 'almost implication', again a correlation measure. Wallace et al. (1996) infer Bayesian networks, which identify the most probable *a posteriori* model for some set of models, while Neil and Korb (1998) extend the methodology to examine ways in which the connection pattern of variables can also be inferred (at least within the limits of statistically equivalent classes). Chambers (1991) has proposed the method of 'Corresponding Regressions', a method which does allow causal inference.

Measures of interestingness

Interpretability and expectation

One of the more obvious properties we might like of our models is that they be interesting and interpretable, and also that they form a reasonable basis for actions. Interpretability probably implies conformity with known beliefs or facts (Pazzani and Kibler 1992) and (possibly) with simplicity, while interestingness relates to deviation from expectation and, in goal-directed situations, to actionability (Barsalou 1995, Hilderman and Hamilton 1999), i.e., given that our focus of interest is in obtaining some specific ends, how easy it is to use the result for those purposes? MacKay (1969) argued that all patterns were ultimately for an agent, who presumably will have some personal agenda. How much ontogenetic complexity an agent is willing to accept in pursuing this agenda

then becomes a personal choice although there are grounds for desiring simplicity (van den Bosch 1994).

We mostly do not know what criteria have actually been used in published studies, although in some cases users may have recorded their expectations prior to analysis. However, I have one example where expectation can be reconstructed. Williams et al. (1969) analysed secondary successional data from subtropical rainforest using cluster analysis. Presence data were analysed using Agglomerative Information Analysis (Williams et al. 1966) and the number of clusters chosen (7) was a subjective decision. Surprisingly, a re-analysis using density data and an objective determination of the number of clusters (Dale 2000), showed more or less the same number of clusters (8), while re-analysis of the presence data showed 3 clusters but with the density-based clusters neatly nested within them. Williams et al. (1969) did not provide any precise information on the criteria actually used to establish the number of clusters. The clustering method is both crisp and hierarchical so it would be possible to work down the hierarchy examining the resultant groupings for spatial or other environmental coherence as Austin (1970) recommends.

This does not appear to have happened. Webb (pers. comm.) has indicated that the choice was made largely on the basis of habitat types visible on site after the initial clearance; for example areas with soil scraped by bulldozers and with timber heaps. To these observers, such differences were likely to result in different development patterns. Webb's observation indicates that conformity with such expectation was a primary reason for fixing the number of clusters at 7, perhaps a surprising conclusion. Temporal and/or spatial coherence were, apparently, not regarded as important, nor did the interpreters investigate alternative possibilities implied by the hierarchical clustering scheme, though this may have helped confirm their initial presumptions and was certainly a component of their subsequent interpretations.

Klemettinen et al. (1994) point out that formation of rules is easy, finding interesting ones is not. There are many proposals for measuring interest, some objective, some subjective; Schmidhuber (1994) lists 14 measures and Hilderman and Hamilton (1999) present others. Most of the objective rules involve GUHA-like (Hájek et al. 1966) 'almost implication', that is rules of the form 'the set of conditions A 'almost always' implies the result B'. Objective measures indicating potential interest of such rules include coverage (fit), certainty, strength, sensitivity, predictiveness, statistical significance and simplicity and most have already been considered here as part of the model evaluation process.

Actionability and agents

Subjective indicators of interestingness include both unexpectedness (Padmanabhan and Tuzhilin 1999) and actionability (Adomavicius and Tuzhilin. 1997). In order to determine unexpectedness an obvious approach is to adopt a Bayesian viewpoint, and incorporate our expectations in the prior probabilities. Bayesian analysis in effect asks how much we have to change our prior view in the light of the observed data. This would require a user to explicitly state expectations, but this is complicated by the possibility that the user will regard some things as 'too obvious for comment' and will in consequence ignore them.

Mackay's (1969) notion of 'patterns for agents' suggests that every investigator has a singular agenda which will be reflected in the selection of the focus of attention. One aspect of this will commonly be the desire to accomplish some aim, perhaps to maximise productivity, maintain diversity or ensure survival. Appraisal of a model will then involve assessment of its value in meeting these ends. Such actionability is difficult to assess, even if we have prior information on the user's interests; we need to know what the user desires to accomplish. Perhaps multiple predictive analysis (Beeston and Dale 1975) provides a possible mechanism for assessing the potential usefulness since it incorporates information on alternative possibilities for modifying the system.

Another aspect of actionability which might be quantified, concerns the extent to which there is consistency with prior knowledge (Murphy and Allopenna 1994). Thus, in some decision making situations it may be desirable that the rules to be used do not appear counter-intuitive to a human user otherwise reliant on previous knowledge (cf. Pazzani et al. 1997). This means adopting rules that fit human expectations in preference to those that do not. Psychological investigations can indicate the forms of acceptable rules (see e.g., Kelley 1971). But while this may seem attractive in a behaviourally isomorphic model, in a structurally isomorphic model we might well seek innovation and revolution. Actionability, I suggest, like predictivity is not a primary criterion in valorising models

Aesthetics

As a final criterion we might look to aesthetics (Reich 1993). Exactly how a valuation could be developed to capture the concept of a 'beautiful' or 'elegant' model is not clear, though it probably would include features of simplicity, fit, predictiveness and interest. But if mathematical proofs and computer programs can be deemed beautiful, then why not ecological models of romantic and

classical kinds, one emphasising function, the other form! That beauty is truth and truth beauty may be more than a poetic fantasy.

Other considerations

Dynamics and dependencies

In building models we often have observations organised in time and space, as discussed for example by Anand (1997, 2000). In such cases there exists, potentially, a dependency between the observations. Furthermore, along any gradient we can also expect a variation of interest, novelty, surprisingness, uncertainty, even conflict and we match complexity against our processing power to determine attractiveness. Anderson and McMaster (1982) examined ebb and flow of interest (in their case the affective tone of a sentence or paragraph) along a gradient, and identified fluctuations of emotional tension and the complexity of pattern as aesthetic components which correlated well with the 'popularity of the books. Similar variation in the values associated with quality of models provides a means of characterising the nature of the underlying series (Paluš 1997). Instead of purely phenomenological parameters used in stochastic methods, we have invariants characterising dynamic properties, which can be used as a first step to building a model.

Wolpert and Macready (1997) used self-dissimilarity measures based on entropy to obtain a scale-related characterisation of a complexity spectrum. Scale-dependency of this kind can be used to define information dimensions relevant to characterising chaotic series (Grassberger and Procaccia 1983). Entropy measures the error we have in determining our location in a state space while entropy rate considers how this error changes with time and can be measured using the Kolmogorov-Sinai or metric entropy rate (Paluš 1997). Entropy rate is the maximal diversity of patterns in a data stream and can be related both to mutual information between parts of the series and system memory. It is also related to Lyapunov exponents (see Osledec 1968, Devaney 1985, Abarbanel et al. 1992) being the sum of the positive exponents which is commonly just the largest of them and Kolmogorov complexity (Anand and Orlóci 1996).

For time series, one characteristic of interest is the 'memory' of the system – how far back do we have to move before the present becomes independent of the past. This is important in the definition of causal states³

(Crutchfield and Shalizi 1999) and is also related to the Effective Entropy of Grassberger (1989, 1991). Such a measure indicates the period we need to observe a series in order to make effective predictions. The technique is to divide the series into parts, assess the value of each part separately and also assess the entire series. If there is any shared pattern, then the value for the entire series should be smaller than the sum of values for the parts, at least in the simplest case. If it is not then the parts share no patterns, no common information on which correlation might rest.

Other approaches to the analysis of temporally dependent observations which employ explicit valuation of models include Edgoose and Alison (1999) and Li and Biswas (1999; see also Dale et al. 2002) while Wallace (1998) has developed a method suited to the study of spatial dependency, where there are 2-dimensional dependencies.

Practical measures

So far I have been discussing what qualities might be useful to characterise model value. Various proposals have been made as to how such properties might be quantified in practice and several authors have suggested possible schemes. Kohavi (1995, see also Pillar 1999b) has looked at cross-validation and bootstrapping for accuracy estimation and model selection; it is known that, in the limit, this converges to Akaike's (1977) criterion, which uses a complexity-fit trade-off. Vapnik and Chervonenkis (1971) suggested guaranteed risk minimisation which is a non-Bayesian criterion. Schwarz (1978) proposed the Bayesian Information Criterion, which has been used by Li and Biswas (1999). Crutchfield and Young (1989) used a causal information measure which captures aspects of the memory of the system. Related to this view is that of Badii and Politi (1997) who identify complexity with what happens to the error rate as the level of resolution of a model is increased: the slower the convergence, the higher the complexity. Wallace and Freeman (1987) and Rissanen (1995) have used estimates of Kolmogorov's information in slightly different ways, the latter using minimum description length to compare model classes, the former using MML to identify the 'best' model within a class. MML could be, but has not yet, extended to allow comparisons of model classes, and such a comparison would involve stating prior probabilities for those classes. Such priors would presumably reflect both the particular situation and the personal bias of the user. Most of these

3 Perhaps the simplest definition of a causal state is as follows (Shalizi and Crutchfield 1999): Define an equivalence relationship \sim between histories such that two histories are equivalent if and only if they have the same conditional distribution of futures. Causal states are the equivalence classes generated by \sim .

measures are concerned with the fit-complexity trade-off, though Wallace and Freeman (1987) also consider the question of precision. Kontkainen et al. (1999) discuss the impact of missing data in the estimation of complexity but found considerable robustness.

Viswanathan et al. (1999) have compared several of these estimates using the task of segmenting a series. They found in general that the MML criterion was superior. Forster and Sober (1994) found that none of the methods they examined (which did not include MML) was universally preferable, which is to be expected, since no universal pattern-finding mechanism can exist.

The mechanism of use of the valuation is simple enough in principle. We can examine a class of models with a number of parameters. For example, consider the model class of fuzzy clustering. Investigating such a class involves estimating the number of clusters, the parameters of the clusters, the precision warranted for these parameters and the (partial) assignment of the things being clustered to clusters. However, the critical parameter is clearly the number of clusters. For each model with a different number of clusters, we calculate a value and that model with the smallest (or largest) value is taken to be the optimal one. The range of models considered can include one equivalent to a null hypothesis; e.g., the null hypothesis in clustering is a single cluster and this is one possible choice for the optimal solution.

Discussion

I have indicated that it is possible to obtain measures of fit and complexity of models, and to balance the components of precision and fit. It is also possible to balance coarseness of estimate and the lack of fit that this necessitates. I have further argued that predictivity should not be a component of the valuation of models in the inductive stages, except in so far as we require some of it. Good prediction can be made with a model of no structural merit, and partial structural merit does not guarantee good prediction! Finally I have argued that various measures of interest and actionability might be available if we accept the use of prior probabilities. More mundanely I have indicated that clustering of vegetation should be 'fuzzy' rather than crisp in order to obtain consistency.

It is tempting to ask which patterns are the ones really there in the world – ours, or the cuttlefish's? or the crab's or even those of *E. coli* as it hunts for glucose – but the temptation must be resisted. All patterns provide predictive leverage, or we would be around to carve them out of 'black and jointless' continuity; and that is as far as reality extends for patterns (Dennett 1991).

Eco (1980) confirms the temporary nature of models. He argues that we build nets to enable us to scale an obstacle blocking our view. Once atop the wall, we can see where to go, and we can throw the net away. In this sense it does not actually matter whether the net represents any particular truth, provided it is useful in scaling the wall.

A model is always a simplification of a real system. In fact, a model is both an idealisation and a simplification. Real systems have the same properties 'in so far as' they approximate the idealisation, but this is never complete even if, for some purposes, it is adequate and effective. However, models also need to be interesting; a completely random model is decidedly uninteresting since it has no patterns (that we can perceive) while a regular unchanging model is equally boring.

Models are useful for comparison. I can ask if this model (in some class) is more effective at capturing patterns than some other model in some other class. And I can ask which of two classes provides the greater capture of pattern. I first ask for a value for the null case and for the case when pattern has been captured. Assuming that the procedure for capturing pattern is effective I can then calculate the change in value. This can be done for both classes of model, using the optimal model within each class, and thus we can decide both which patterns provide us with the greatest change in value and which has the greatest overall value. The next step is to use model valuations in the analysis of data and determine if the obtained patterns are useful, interesting and effective in comparisons.

The consilience of induction

How can one justify the use of induction? Basically because it works, though this, as Hume pointed out, is itself an induction. Reichenbach (1950) suggests that we are acting like fishermen who cast nets into an unknown part of the ocean. They do not know if they will catch fish, but they do know that without casting the net they will certainly catch none. They try because they wish to eat. Our predictions of the future may not be true, but they are the best available (we hope). Induction is then the best available instrument of action available to us.

But what exactly are we modelling and is it what we want to model? Taking Antonelli's (1990) example of an ant traversing a landscape. It is possible to identify sequences of events which reflect iterative application of solutions to particular local problems: such actions might include 'goes over', 'goes under', 'goes around', 'moves forward' patterns although these may be incomplete when recursively applied as when 'going over' in the middle of

‘going over’. (For finding such motifs, see Rigoutsos and Floratos 1998).

This does not do us much good. Teleologically, the ant is going home and its behaviour in overcoming obstacles need not be interesting to us. In fact, the mechanisms lack discriminatory power when considering what the ant is doing. They would also be used if the ant were seeking food through exploration, i.e., proceeding in an opposite direction! So how does knowledge of the local mechanism get transformed into some higher level of comprehension?

Note, too, that an individualistic approach to modelling requires us to specify exactly the interactions permitted between pairs of organisms and between organisms and environment, as is done in cellular automaton models (Dale and Hogeweg 1998). The clarity obtained makes such models very attractive, if we can recognise emergent properties of the system. This difficulty was recognised by Kaufman (2001) who defines life in terms of self-replication and ‘doing’. He then insists that we shall need a new mathematics to be able to capture this.

Dennett (1991) suggests we adopt one of three stances to explain and predict what something will do: intentional, design and physical. What is odd here is that Kaufman’s view seems to relate to the intentional stance which Dennett regards as least powerful! I would expect some relationship between Dennett’s stances and Aristotle’s four kinds of causes: material, efficient, formal and final (see Chambers 1991). Much scientific study emphasises efficient and formal causes, whereas Antonelli’s ant is seemingly concerned with final causes. Traditional methods of empirical research have a tendency to reduce telic theories to non-telic, billiard ball type models (Rychlak 1988, Chambers 1991). Is Antonelli’s ant concerned with final causes, or simply to be regarded as a machine? It is difficult, though, to conceive of vegetation ‘doing’ anything in a teleological sense, so that Kaufman’s problems may not arise in practice – but does ‘doing’ necessarily imply teleological? I do not think so, especially with respect to Kaufman’s definition of self-replication.

Acknowledgements: I would like to thank Sanyi Bartha and Pat Dale for many helpful comments made after reading earlier drafts.

References

Adomavicius, G. and A. Tuzhilin. 1997. Discovery of actionable patterns in databases: the action hierarchy approach. In: D. Heckerman, H. Mannila, D. Pregibon and R. Uthurusamy (eds.), *Proc. Third International Conference on Knowledge Discovery and Data Mining*. AAAI, pp. 111-114.

- Akaike, H. 1977. On entropy maximization principles. In: P. K. Krishnaiah (ed.), *Applications of Statistics* North Holland, Amsterdam. pp 27-41.
- Anand, M. 1997. The fundamental nature of vegetation dynamics - a chaotic synthesis. *Coenoses* 12: 55-62.
- Anand, M. 2000. Fundamentals of vegetation change: complexity rules. *Acta Biotheoretica* 48: 1-14.
- Anand, M. and Orlóci, L. 1996. Complexity in plant communities: the notion and quantification. *J. theoret. Biol.* 179:179-186.
- Anderson, C. W. and G. E. McMaster. 1982. Computer assisted modelling of affective tone in written documents. *Comput. Humanit.* 16: 1-9.
- Antonelli, P. L. 1990. Applied Volterra-Hamilton systems of the Finsler type: increased species diversity as a non-chemical defense for coral against crown-of-thorns. In: R. H. Bradbury (ed.), *Acanthaster and the Coral Reef: A Theoretical Perspective, Lecture Notes in Biomathematics* 88. Springer-Verlag, Berlin pp. 220-235.
- Abarbanel, H. D. I., R. Brown and M. B. Kennel. 1992. Local Lyapunov Exponents Computed from Observed Data. *Journal of Nonlinear Science* 2:343-365.
- Austin, M. P. 1970. An applied ecological example of mixed data classification. in: R. S. Anderssen and M. R. Osborne (eds.), *Data Representation*. University of Queensland Press, Brisbane. pp. 113-117.
- Badii, R. and A. Politi. 1997 *Hierarchical Structure and Scaling in Physics*. Cambridge University Press, Cambridge.
- Barsalou, L. W. 1995. Deriving categories to achieve goals. In: A. Ram and D. B. Leake (eds.), *Goal Directed Learning*. MIT Press, Cambridge MA. pp. 121-176.
- Beeston, G. R. and M. B. Dale. 1975. Multiple predictive analysis: a management tool. *Proceedings of the Ecological Society of Australia* 9: 172-181.
- Boerlijst, M. and P. Hogeweg. 1991. Spiral wave structure in prebiotic evolution: hypercycles stable against parasites. *Physica D* 48: 17-28.
- Brokaw, N. and R. T. Busing. 2000. Niche versus chance in tree diversity in forest gaps. *TREE* 15: 183-188.
- Carley, K. and M. Palmquist. 1992. Extracting, representing and analyzing mental models. *Social Forces* 70: 601-636.
- Chambers, W. V. 1991. Inferring formal causation from corresponding regressions. *The Journal of Mind and Behavior* 12:49-70.
- Crutchfield, J. P. and C. R. Shalizi. 1999. Thermodynamic depth of causal states: when paddling around in occam’s pool, shallowness is a virtue. *Physical Review E* 59: 275-283.
- Crutchfield, J. P. and K. Young. 1989. Inferring statistical complexity. *Physical Review Letters* 63: 105-108.
- Dale, M. B. 1970. Systems analysis and ecology. *Ecology* 51: 2-16.
- Dale, M. B. 1980. A syntactic basis for classification. *Vegetatio* 42: 93-98.
- Dale, M. B. 1988. Some fuzzy approaches to phytosociology: ideals and instances. *Folia Geobotanica Phytotaxonomica* 23: 239-274.
- Dale, M. B. 1994. Straightening the horseshoe: a Riemannian resolution? *Coenoses* 9: 43-53.
- Dale, M. B. 1999. The dynamics of diversity: mixed strategy systems. *Coenoses* 13:105-113
- Dale, M. B. 2000. Mt Glorious revisited: Secondary succession in subtropical rainforest *Community Ecol.* 1: 181-193.

- Dale, M. B. and Barson, M. M. 1989. Grammars in vegetation analysis. *Vegetatio* 81: 79-94.
- Dale, M. B., R. Coutts and P. E. R. Dale. 1988. Landscape classification by sequences: a study of Toohey Forest. *Vegetatio* 29: 113-129.
- Dale, M. B., P. E. R. Dale and T. Edgoose. 2002a. Markov models for incorporating temporal dependence. *Acta Oecologica* 23: 261-269.
- Dale, M. B., P. E. R. Dale, C. Li and G. Biswas. 2002b. Assessing impacts of small perturbations using a model-based approach. *Ecological Modelling* 156: 185-199.
- Dale, M. B. and Hogeweg, P. 1998. The dynamics of diversity: a cellular automaton approach. *Coenoses* 13:3-15.
- Dale, P., K. Hulsman, B. R. Jahnke and M. B. Dale. 1984. Vegetation and nesting preferences of Black Noddies at Masthead Island, Great Barrier Reef. Part I. Patterns at the macro scale. *Australian Journal of Ecology* 9: 335-341.
- Dale, M. B., L. Salmina and L. Mucina. 2001. Minimum message length clustering: an explication and some applications to vegetation data. *Community Ecol.* 2:231-247.
- Dennett, D. 1991. Real patterns. *J. Philosophy* 88:27-51.
- Devaney, R. L. 1985. *An Introduction to Chaotic Dynamical Systems*. Benjamin/Cummings, Menlo Park.
- Domingos, P. 1996. Two-way induction. *Internatl. J. Artificial Intelligence Tools* 5: 113-125.
- Domingos, P. 1998. When and how to combine predictive and causal learning. Proc NIPS-98 Workshop on Integrating Supervised and Unsupervised Learning, Breckenridge, CO. NIPS Foundation.
- Domingos P. 1999. The role of Occam's Razor in knowledge discovery. *Data Mining and Knowledge Discovery* 3: 409-425.
- Eco, U. 1980. *Il Nome della Rosa*. Gruppo Editoriale Fabbri-Bompiani, Sonzogno, Etas S. p. A.
- Edgoose, T. and L. Allison. 1999. MML Markov classification of sequential data. *Statistics and Computing* 9: 269-278.
- Farrands, J. L. 1990. On modelling. In: R. H. Bradbury (ed.), *Acanthaster and the Coral Reef: A Theoretical Perspective Lecture Notes in Biomathematics* 88. Springer-Verlag, Berlin. pp.1-5.
- Fisher, D. 1992. Pessimistic and optimistic induction. TR CS-92-12 Dept. Comput. Sci., Vanderbilt Univ.
- Forster, M. P. and E. Sober. 1994. Key concepts in model selection: performance and generalization. *Brit. J. Philosophy Sci.* 45:1-35.
- Gell-Mann, M. 1994. *The Quark and the Jaguar*, W. H. Freeman, San Francisco.
- Georgeff, M. P. and C. S. Wallace. 1984. A general criterion for inductive inference. In: T. O'Shea (ed.), Proc. 6th European Conf. Artificial Intelligence, Elsevier, Amsterdam.
- Grassberger, P. 1989. Problems in quantifying self-generated complexity. *Helvetica Physica Acta* 62: 489-511.
- Grassberger, P. 1991. Information and complexity measures in dynamical systems. In: H. Atmanspacher and H. Scheingraber (eds.), *Information Dynamics*. Plenum Press. New York, pp. 15-33.
- Grassberger, P. and F. Procaccia. 1983. Estimation of the Kolmogorov entropy for a chaotic signal. *Phys. Rev. A* 28: 2591.
- Gunther, R., B. Shapiro and P. Wagner. 1994. Complex systems, complexity measures, grammars and model inferring, *Chaos, Solitons and Fractals* 4: 635-651.
- Hájek, P., I. Havel and M. Chytil. 1966. GUHA – the method of systematical hypotheses searching. *Kybernetika (Prague)* 2:31-39.
- Hájek, P. and T. Havránek. 1977. On generation of inductive hypotheses. *International. J. Man-Mach. Stud.* 9: 415-438.
- Herman, G. T. and Rozenberg, G. 1975. *Developmental Systems and Languages*, North-Holland, American Elsevier, Amsterdam.
- Hilderman, R. J. and Hamilton, H. J. 1999. Heuristics for ranking the interestingness of discovered knowledge. Proc. 3rd Pacific-Asia Conf. Knowledge Discovery PKDD'99, Beijing, Springer-Verlag, Berlin, pp. 204-209.
- Hoeting, J., D. Madigan, A. E. Raftery and C. T. Volinsky. 1998. Bayesian model averaging: a tutorial. *Statist. Sci.* 14: 382-417.
- Hogeweg, P. 2002. Computing an organism: on the interface between informatic and dynamic processes. *BioSystems* 64: 97-109.
- Howard, E. and N. Oakley. 1994. The application of genetic programming to the investigation of short, noisy, chaotic data series. In: T. C. Fogarty (ed.), *Evolutionary Computing. Lecture Notes in Computer Science* 865, Springer-Verlag, Berlin. pp. 320-332.
- Hume, D. 1999. *An Enquiry Concerning Human Understanding*. Oxford Philosophical Texts, Oxford University Press, Oxford.
- Iba, W., Wogulis, J. and Langley, P. 1988. Trading off simplicity and coverage in incremental concept learning. Proc. 5th Internatl. Conf. Machine Learning, Ann Arbor, Morgan Kaufman, CA. pp. 73-86.
- Ivakhnenko, A. G. 1971. Polynomial theory of complex systems I. *E. E. E. Trans. Syst. Man Cybern. SMC* 1: 364-378.
- Jöreskog, K. G. 1966. Some contributions to maximum likelihood factor analysis. Research Bulletin RB-66-41, Educational Testing Service, Princeton, N. J.
- Kaufman, S. 2001. *Investigations*. Oxford University Press, Oxford.
- Kelley, H. 1971. Causal schemata and the attribution process. In: E. Jones, D. Kanouse, H. Kelley, N. Nisbett, S. Valins and B. Weiner (eds.), *Attribution: Perceiving the causes of behavior*. General Learning Press, Morristown, NJ. pp 151-174.
- Klemettinen, M., H. Mannila, P. Ronkainen, H. Toivonen and A. I. Verkamo. 1994. Finding interesting rules from large sets of discovered association rules. In: N. R. Adam, B. K. Bhargava and Y. Yesha (eds.), *Third Internatl. Conf. Information and Knowledge Management CIKM'94*, ACM Press Association. pp. 401-407.
- Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, Proc. International Joint Conference Artificial Intelligence.
- Kolmogorov, A. N. 1965. Three approaches to the quantitative description of information. *Prob. Inform. Transmission* 1: 4-7 (translation).
- Kontkainen, P., P. Myllymäki, T. Silander and H. Tirri. 1999. On stochastic complexity approximation. In: A. Gamerman (ed.) *Causal Systems and Intelligent data Management.*, Springer-Verlag, Berlin. pp. 120-136.
- Lekkas, G. and N. Avouris. 1994. Case-Based Reasoning in Environmental Monitoring. *Applied Artificial Intelligence* 8: 359-376.
- Li, C. and G. Biswas. 1999. Temporal pattern generation using hidden Markov model based unsupervised classification. Lecture Notes in Computer Science 1662. pp. 245-257.
- Lloyd, S. and H. Pagels. 1988. Complexity as thermodynamic depth. *Ann. Physics* 188:186-212.

- Löfgren, L. 1974. Complexity of descriptions of systems: A foundational study. *International Journal of General Systems* 3: 197-214.
- Lux, A. and F. A. Bemberlein-Lux. 1998. Two vegetation maps of the same island: floristic units versus structural units. *Applied Vegetation Science* 1: 201-210.
- Mac Nally, R. 2000. Regression and model-building in conservation biology, biogeography and ecology: The distinction between - and reconciliation of - 'predictive' and 'explanatory' models. *Biodiversity and Conservation* 9: 655-671.
- Mackay, D. M. 1969. Recognition and action. In: S. Watanabe (ed.), *Methodologies of Pattern Recognition*, Academic Press, London. pp. 409-416.
- May, R. M. 1995. Necessity and chance: deterministic chaos in ecology and evolution. *Bull. Amer. Math. Soc.* 32: 291-308.
- Murphy, G.L. and P. D. Allopenna. 1994. The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19: 203-222.
- Neil, J. R. and K. B. Korb. 1998. The MML evolution of causal models Tech. Rep. 98/17 School of Computer Science and Software Engineering, Monash University, Clayton, Victoria 3168, Australia.
- Niven, B. S. 1992. Formalization of some basic concepts of plant ecology. *Coenoses* 7: 103-113.
- Noble, I. R. and R. O. Slatyer. 1980. The use of vital attributes to predict successional changes in plant communities subject to recurrent disturbances. *Vegetatio* 43: 5-21.
- Oates, T. and D. Jensen 1998. Large datasets lead to overly complex models: an explanation and a solution. *KDD-98 Proc.* 4th Internatl. Conf. Knowledge Discovery and Datamining. pp. 294-298.
- Osledec, V. I. 1968 A multiplicative ergodic theorem: Lyapunov characteristic numbers for dynamical systems. *Trans. Moscow Math. Soc.* 19: 197-231.
- Padmanabhan, B. and A. Tuzhilin. 1999. Unexpectedness as a measure of interestingness in knowledge discovery. Decision Support Systems, 27. from <http://citeseer.nj.nec.com/padmanabhan99unexpectedness.html>
- Pagie, L. and P. Hogeweg. 1997. Evolutionary consequences of co-evolving targets. *Evolutionary Computation* 5: 401-418.
- Paluš, M. 1996a. Detecting nonlinearity in multivariate time series. *Physics Letters A* 213: 1387.
- Paluš, M. 1996b. Coarse-grained entropy rates for characterisation of complex time series. *Physica D* 93: 64-77.
- Paluš, M. 1997. Kolmogorov entropy from time series using information-theoretic functionals. *Neural Network World* 7: 269-292.
- Pazzani, M. J. and D. Kibler. 1992. The utility of knowledge in inductive learning. *Machine Learning* 9: 57-94.
- Pazzani, M., S. Mani and W. R. Shackle. 1997. Comprehensive knowledge discovery in databases. In: M. G. Shafto and P. Langley, (eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum, pp. 596-601.
- Pillar, V. D. 1999a. On the identification of optimal plant functional types. *J. Veg. Sci.* 10: 631-640.
- Pillar, V. D. 1999b. The bootstrap ordination revisited. *J. Veg. Sci.* 10: 895-902.
- Popper, K. R. 1968. *The Logic of Scientific Discovery*. Harper, New York.
- Posse, C. 1995. Projection pursuit exploratory data analysis. *Computat. Statist. Data Anal.* 20: 669-687.
- Provost, F., T. Fawcett and R. Kohavi. 1998. The Case Against Accuracy Estimation for Comparing Induction Algorithms. Presented at ICML-98 (15th Internatl. Conf. on Machine Learning).
- Reich, Y. 1993. A model of aesthetic judgment in design. *Artif. Intell. in Engineering* 8: 141-153.
- Reichenbach, H. 1950. *The Rise of Scientific Philosophy*. Univ. California Press, Los Angeles.
- Riddle, R. R. and D. J. Hafner. 1999. Species as unit of analysis in ecology and biogeography: time to take the blinkers off. *Global Ecology and Biogeography* 8: 433-441.
- Rigoutsos, I. and A. Floratos. 1998. Motif discovery without alignment or enumeration. Proc. 2nd. Ann. ACM Internatl. Conf. Computational Molecular Biology (RECOMB 98). New York, NY.
- Rissanen, J. 1995. Stochastic complexity in learning. In: P. Vitányi (ed.), *Computational Learning Theory Lecture Notes in Computer Science* 904. Springer-Verlag, Berlin. pp. 196-201.
- Rychlak, J.F. 1988. *The Psychology of Rigorous Humanism*. New York University Press, New York.
- Savill, N. J., P. Rohani and P. Hogeweg. 1997. Self-reinforcing spatial patterns enslave evolution in a host-parasitoid system. *J. theoret. Biol.* 188: 11-20.
- Schmidhuber, J. 1994. Discovering solutions with low Kolmogorov complexity and high generalization ability. Tech. Rep. FKI-194-94 Faculty of Information, Technical University, Munich.
- Schwarz, G. 1978. Estimating dimension of a model. *Ann. Statist.* 6: 461-464.
- Shalizi, C. R. and J. P. Crutchfield. 1999 Computational mechanics: pattern and prediction, structure and simplicity. Sante Fe Institute Working paper 99-07-044.
- Shipley, B. and P. A. Keddy. 1987. The individualistic and community-unit concepts as falsifiable hypotheses. *Vegetatio* 69: 47-55.
- Simberloff, D. 1980. A succession of paradigms in ecology: Essentialism to materialism and probabilism. *Synthese* 43: 3-29.
- van den Bosch, A. P. M. 1997. Simplicity and Prediction. Available electronically at <http://tcw2.ppsw.rug.nl/~vdbosch/simple.ps>
- Vapnik, V. N. and A. Chervonenkis. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probability Appl.* 16: 264-280.
- Viswanathan, M., C. S. Wallace, D. L. Dowe and K. B. Korb. 1999. Finding cutpoints in noisy binary sequences: a revised empirical examination. In: N. Foo (ed.), *AI-99 Lecture Notes in Artificial Intelligence* 1747. Springer-Verlag, Berlin. pp. 405-416.
- Wackerbauer, R., A., Witt, H. Altmanspracher, J. Kurths and H. Scheingraber. 1994. A comparative classification of complexity measures based on distinguishing partitions in phase space as well as structural v. dynamic elements. *Chaos, Solitons and Fractals* 4: 133-173.
- Wallace, C. S. 1995. Multiple factor analysis by MML estimation. Tech. Rep. 95/218, Dept Computer Science, Monash University, Clayton Victoria 3168, Australia.
- Wallace, C. S. 1996. MML Inference of predictive trees, graphs and nets. In: A. Gammerman (ed.), *Computational Learning and Probabilistic Reasoning*. John Wiley. New York. pp. 43-66.
- Wallace, C. S. 1998. Intrinsic classification of spatially-correlated data. *Comput. J.* 41: 602-611.

- Wallace, C. S. and D. L. Dowe. 2000. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing* 10: 73-83.
- Wallace, C. S. and P. R. Freeman. 1987. Estimation and inference by compact coding. *J. Roy. Statist. Soc. Ser. B* 49: 240-252.
- Wallace, C. S., K. B. Korb and H. Dai. 1996. Causal discovery via MML. Tech. Rep. 96/254 Dept. Computer Science, Monash University, Clayton, Victoria 3168, Australia.
- Watanabe, S. 1969. *Knowing and Guessing*. Wiley, New York.
- Webb, G. I. 1994. Generality is more significant than complexity: Toward alternatives to Occam's razor. In: C. Zhang, J. Debenham and D. Lukose (eds.), *AI'94 - Proceedings of the Seventh Australian Joint Conference on Artificial Intelligence*. World Scientific, Armidale. pp. 60-67.
- Webb, G. I. 1996. Further experimental evidence against the utility of Occam's Razor. *J. Artif. Intell. Res.* 4:387-417.
- Webb, L. J., J. G. Tracey and W. T. Williams. 1976. The value of structural features in tropical forest typology. *Austral. J. Ecol.* 1:3-28.
- Williams, W. T. 1972. The problem of pattern. *Austral. Mathem. Teacher* 28:103-109.
- Williams, W. T., J. M. Lambert and G. N. Lance. 1966. Multivariate methods in plant ecology V. Similarity analysis and information analysis. *J. Ecol.* 54:427-445.
- Williams, W. T., G. N. Lance, L. J. Webb, J. G. Tracey and M. B. Dale. 1969. Studies in the numerical classification of complex rain-forest communities III. The analysis of successional data. *J. Ecol.* 57: 515-535.
- Wilson, J., A. D. Q. Agnew and T. R. Partridge. 1994. Carr texture in Britain and New Zealand: community convergence compared with a null model. *J. Veg. Sci.* 5:109-116.
- Wisheu, I. and P. A. Keddy. 1992. Competition and centrifugal organisation of plant communities: theory and tests. *J. Veg. Sci.* 3: 147-156.
- Wittgenstein, L. 1995. *Tractacus Logico-Philosophicus*. (trans) 5:3631 Routledge, Keagan & Paul, London.
- Wolpert, D. H. and W. G. Macready. 1997. Self-dissimilarity: an empirically observable complexity measure. In: Y. Bar-Yam (ed.), *Proc. International. Conf. Complex Systems*, New England Complex Systems Inst. pp. 1-8.
- Wright, S. 1934. The method of path coefficients. *Ann. Mathem. Statist.* 5:161-215.
- Yamada, H. and S. Amaroso. 1971. Structural and behavioural equivalences of tessellation automata. *Information and Control* 18:1-31.
- Yee, C. N. and L. Allison. 1993. Reconstruction of strings past. *J. Comp. Appl. BioSci.* 9: 1-7.