COMMUNITY ECOLOGY 4(2): 129-139, 2003
1585-8553/$20.00 © Akadémiai Kiadó, Budapest

Continuum or community: *a priori* assumption or data-dependent choice?

M. B. Dale

Australian School of Environmental Studies, Griffith University, Nathan, Qld. 4111, Australia
Phone: +61 7 3371 4414, Fax: +61 7 3870 5681, E-mail: m.dale@griffith.edu.au

Keywords: Continuum, Community, Mixture model, Model quality, Complexity, Fit, Precision, Minimum mesage length, Static, Dynamic, Hidden Markov processes.

Abstract: This paper examines how we might test the continuum theory against the community unit theory. Adherence to one or other of these models without testing is simply an assignment of an extreme prior probability to the preferred option. The question can be rephrased to ask whether, for a set of observations, a single model is adequate or whether a mixture of models would be preferable. To judge between them involves first defining the nature of the model(s) to be fitted in each case and then comparing the complexity and quality of fit. Occam's razor suggests that we should seek the simplest model with adequate fit, with parameters estimated with optimal precision. The simplest comparison of the two theories thus requires only the estimation of the number of clusters for the chosen model(s) of within-cluster variation. If a single cluster is of adequate quality then the continuum model is appropriate, while if several are needed then the community model is preferable for that particular dataset. To establish universal applicability of either model involves investigation of many datasets.

There are several ways in which model quality can be assessed, and here I concentrate on the minimal message length principle which is a function of the prior probability of the model and its fit to the observed data, assuming the model to be correct. This principle has been shown to perform well when compared with other possibilities.

I first illustrate the procedure for making a choice between models, using a simple model, then examine two alternative formulations of within-cluster models which seem more appropriate, one static, the other dynamic.

Introduction

Intellectual progress can often occur through abandonment of questions together with the alternatives they assume, an abandonment that results from their decreasing vitality and a change of urgent interest – we do not solve them so much as get over them. To some extent the arguments concerning continuum theory and community unit theory have followed this path, with attention in ecology now being largely directed elsewhere. Indeed, Shipley and Keddy (1987) have argued that the continuum and community unit hypotheses are not falsifiable so that further debate is pointless. This is a pessimistic, not to say myopic, view of how we build models which may not be totally appropriate (cf. Fisher 1992). Falsification looks for what is not, not for what is. In any case, statistical tests do not so much falsify hypotheses as indicate the risks we take by accepting or rejecting them.

There is a further consideration. In order to test such models we need a formal representation and neither theory explicitly determines a single model. Instead they represent classes of models. For example, if I assume that some set of axes represents a formal continuum model, then there is a whole class of models parameterised by the number of axes and several other parameters may also need estimation. For the community unit model, one parameter would be the number of classes, but again further parameters are possible: crisp versus fuzzy, hierarchical versus non-hierarchical and so on. In both cases, there are further alternatives leading to further model classes. Thus, the task is actually to determine the 'best' model class and within that the 'best' model. In fact, both hypotheses are 'collections of families of classes of …' of models. Notice that the selection of model $m[j]$ within class C[i] is equivalent to the estimation of a parameter (j) within a model $m[i]$. Estimating these required parameters is not

logically different from simple parameter estimation within a single model.

Rissanen (1978) proposed Minimum Description Length (MDL) as a means of quantifying the quality of an entire model class, based on an average over all models within a class and hence of discriminating between classes. He has since (1987, 1996) considerably modified his original views. In any case, although MDL might solve our problem, in general we want to identify a specific model and not just the class. To obtain this latter we can determine a 'best' model for each class separately and then make the comparison, which is the approach adopted in Minimum Message Length (MML) studies (Wallace and Dowe 2000, Dale 2002).

In this latter, which is adopted here, we need to provide prior probabilities for the model classes although MML is not strictly Bayesian. So if a user wishes to express some preconceived preference the prior probabilities can be adjusted, otherwise they can be made bland enough to avoid establishing any preferences. Wallace and Georgeff (1983) note that any model accepted using MML is *necessarily* falsifiable.

In this paper, I examine the MML method by which the hypotheses might be tested. This requires a more precise specification of the two models (or rather model classes) and I have assumed, for simplicity, that the required model is such that within any cluster the variation can be described as a multivariate Gaussian or Poisson distribution with uncorrelated attributes. The number of clusters can run from 1 upwards. It is *not* suggested that such a model is an adequate representation of either theory, but the computation is considerably simplified thereby. In fact, recognition of deficiencies in this overly simple model leads to some suggestions of other possible models which might be more worthy of investigation.

Two families of theories

Before we can attempt to discriminate between the two models, we need to make a clear definition of what each of them entails. The available data consist of records of the performance of species in a set of samples from the area of interest. The continuum theory requires an estimation of the number of (possibly correlated) axes, the nature of the response curve and the mathematical nature of the embedding space. For example, Dale (1994) has considered using Riemannian space instead of the more usual Euclidean space and this would necessitate estimating the curvature, while correspondence analysis employs a chi-square metric rather than a Euclidean one. Analogously, community theory will require estimation of the number of communities and also of any parameters necessary to characterise them including any within-cluster variation we choose to permit.

A major difficulty with both theories is that neither adequately addresses questions of process but rather depend on a single snapshot. If a stand of vegetation is in the process of changing from one community to another, our observations will contain elements of both. There is of course no guarantee that the initial communities will have the same boundaries as the final ones and blurring may well occur. Shalizi and Crutchfield (1999) argue that we should define units which have similar predicted paths into the future irrespective of their history but a single snapshot will generally be inadequate for this purpose.

Three further points need brief consideration. First, maximum likelihood estimation does not perform well with problems where the dimensionality may change. For example, in clustering the maximum likelihood solution distinguishes many clusters, only identical samples forming groups. Some other method of estimation is needed and MML provides such estimates.

Second, there is the question of hierarchy in clustering, a feature often adopted by community unit supporters. Hierarchical clustering is simply one subclass of possible clustering models, and by comparing different subclasses we can determine if a hierarchy, or any other specified structure, is desirable (see Boulton and Wallace 1973). In this study, I shall use only non-hierarchical clustering and more specifically, the method used is a fuzzy clustering, because the use of crisp clusters leads to inconsistency in estimating cluster parameters (Dale 2002).

Third, there are always questions of scale and of interactions between processes operating at different scales, and neither theory addresses these directly. Hogeweg (2002) notes that interactions between processes at different scales are to be expected in biological systems. Bar-Yam (2002) considers the problem of characterising the multi-scale behaviour of non-equilibrium systems so as to relate descriptions of a system as a function of the scale of observation. He regards clustering as one way of polarising scales, by separating macro-scale variation – between clusters – from micro-scale clustering – within clusters. However it would be preferable if the scale question were more directly addressed. I am presently studying some possibilities using MML approaches.

The continuum theory

The continuum theory assumes that the performance of plant species is a continuous function of some under-

lying gradients of (often unspecified) factors. To represent this, it employs an axis-based model which is selected to capture the observed correlations between species in some low-dimensional representation. Stone and Porrill (1998) indicate that for many methods, such as component analysis or projection pursuit (Posse 1995), there is an implicit assumption that different processes tend to generate signals that are statistically independent of one another. Thus, one way of finding these processes is to use methods that identify independent signal components or sources. In ecology, this is known as ordination.

A sufficiently complicated axis model will be able to describe arbitrarily complicated distributions, but such complex models do not instil confidence in the generality of their application any more than an O(n-1) polynomial is regarded as a suitable model for n points in regression. 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 exactly and only the instances in known data is not of much use. We have to avoid such overfitting, which means avoiding overly complex models. This avoidance of overly complex solutions also applies to the choice of the class of axis model. A linear solution should be preferred to a unimodal one if both are more or less equivalent in goodness of fit and generalisability, since a linear model requires fewer parameters than a unimodal one. Generalisability here means the ability to fit future data. Of course we also want the axes to be simply related to other information so that we can ascribe some meaning to the axes. We want interesting axes.

Axis models are not the only means by which we can capture correlation structure within a dataset, for Bayesian belief networks can also be inferred (Wallace et al. 1996, Neil et al. 1999) and provide great generality; an ecological example is given in Grace et al. (2000). Wisheu and Keddy (1992) suggested a model for vegetation similar to the Rasch model used by Dale and Anderson (1973) in an hierarchical clustering; mixture models of these have also been examined by Uebersax and Grove (1993). This model assumes that the performance of a species in a particular sample is a function of the species ability, a, and the sample difficulty, b, but is restricted to qualitative or frequency data. We can also consider mixtures of factor models including independent factor analysis (Attias 1999), independent component analysis (Stone and Porrill 1998) and hierarchical versions of these, that combine clustering with within-cluster axis models. These have been suggested by several authors including Hanson et al. (1991) and Edwards and Dowe (1998).

Most of the commonly used axis models are additive, and either linear or logarithmic. Another possibility is due to Gilbert and Wells (1966), which models the co-occurrence frequencies with a multiplicative model. I have found that this generally fits very well, much better than additive models, although it is restricted to presence/absence data.

It is obvious from this discussion that there are in fact many possible classes of model which might be used to capture the ecological notion of continuum. Each such model class may itself have parameters, most notably the number of axes, thus providing several models. To choose between them we might rely on *a priori* ecological arguments regarding the prior probability of particular choices; our models should not incorporate ecological impossibilities if we are aware of such.[1] We might also seek to provide some measure of model quality which allows us to determine the optimal model class, and within such a class the optimal model itself. Such an evaluation might be expected to involve a prior probability, which when combined with the fit to the data leads us to a posterior probability as an evaluation metric.

The community unit theory

The community unit theory proposes that description of vegetation is best made in terms of several distinct 'communities' with largely abrupt boundaries These boundaries are in the embedding space defined by the descriptors and need not be environmentally determined; indeed the existence of the Modifiable Areal Unit Problem (MAUP; Openshaw 1984, Jelinski and Wu 1996) and of possible vegetation-generated spatial patterns (Boerlijst 2000, see also Rietkerk et al. 2002) make boundaries difficult to identify and independent of the environment.

The sampling of extant vegetation often involves selectivity based on the recognition in the field of 'homogeneous areas'. There has been much debate concerning the definition of homogeneity and how it might be established other than by subjective visual assessment. Most of the problems can in fact be resolved by accepting that any clusters obtained through analysis of observational data will be fuzzy, that is the actual vegetation samples will

1 There are some deep philosophical considerations associated with the rejection of 'abnormal' solutions which I shall not consider here (see e.g., Kreinovich and Kunin 2003). Essentially they concern differences between mathematical solutions and 'real-world' solutions, the later being constrained to 'normal' values.

belong more or less to an abstract community defined by specific parameter values. We can then say that the community unit theory is basically an argument that axis-based descriptions become excessively complex and that this can be avoided by developing an overall model as a (piecewise) mixture of simpler ones. Neal (1998) indicates that we can have a countably infinite number of components in such models but in general we would wish to keep the number of clusters relatively small.

This leaves us with a problem concerning the nature of variation within a cluster. To determine how many clusters provide the optimal fit requires within cluster variation to be suitably defined. With a few exceptions, the proponents of community theory have been silent on the nature of acceptable within-cluster variation, although in practice they agree that the performance measure for each attribute (species) should be invariant within a cluster. This is the model I propose to use.

Another possibility is that an axis representation might be used, possibly with different dimensionalities in each cluster. But note that relationships within clusters can be taken to be linear - the clusters themselves provide a mechanism for dealing with non-linearity. It would also be possible to have *different* axis models within each cluster allowing avoidance of the use of axes altogether in some clusters.

To assess a cluster or ordination model, we resort to notions of complexity. In general, more clusters require more parameters and should be regarded as a more complex description. However, things are a little more difficult if the within-cluster variation is considered, for more clusters may allow simpler within-cluster models with fewer parameters; in small enough pieces everything can be regarded as linear.

In the end, the comparison of these two theories depends on two things. First, there is a choice of some model of correlation structure within a single cluster, a choice required by *both* hypotheses. Second, the investigator must determine whether to accept a mixture of simple models instead of a, possibly complex, single model. If the same model is used of within-cluster variation is used in both, then this is nothing more than a test for the optimal number of clusters! A single cluster possibly with many axes is sufficient for the continuum theory, whereas the community unit theory argues for several clusters with a simple within-cluster model for each.

Comparing models and model classes

The problem has now become a question of how we determine an optimal model class or model? I have previously (Dale 2002) examined this question in more detail so only a summary will be given here.

We could certainly obtain a very simple model which fitted our observations badly and equally we could obtain a very complex model which fitted our data exactly. Neither of these seems attractive. The usual procedure, which I follow here, is to invoke Occam's (or Aristotle's) razor (Young et al. 1999, Gamberger and Lavra 1997, Domingos 1999; but see Webb 1996, for arguments against) and seek the simplest model with adequate fit when the parameters of the model are estimated with optimal precision; too low a precision will impact on the quality of fit, too high will increase the complexity. Simplicity is commonly associated with generalisability and Wallace (1996), in discussing the relationship between prediction and induction, concludes that the MML principle "minimises the degree to which future data will surprise us".

When introducing Occam's razor, though, we must be aware that, in reality, there are no grounds for believing that the simplest course of events did really happen. Any model we select may be erroneous. It may not even be possible to determine the 'true' model, as in the case in factor analysis where only some of an infinite number of alternatives can be chosen. Such possible model error is a major contributor to uncertainty.

Here I employ the minimal message length principle (MML) to assess both complexity and fit and thus quantify the quality of any model. It is sufficient to note that an event of probability p will have a message length of $-\log(p)$ so that the shortest message length is associated with the largest (posterior) probability. The test of the two hypotheses then becomes a matter of determining if the message length for a model using a single cluster is smaller than that for models with any other number of clusters, for some model of within-cluster variation. The difference in message lengths is related to the odds in favour of the model with shorter message length; for difference d we calculate the odds as $e^{-d} : 1$.

For clustering, Boulton and Wallace (1970) described and implemented a procedure for estimating the optimal number of clusters for a limited set of models of within -cluster variation; Dirichlet, Poisson, Gaussian and Von Mises distributions, all with uncorrelated attributes, are included. The *a priori* distribution of the number of clusters ranges from 1 to some appropriately large number. If we obtain evidence of more than one cluster, we would accept that a community unit hypothesis has merit. If there is but a single cluster, then the chosen axis-based model is clearly sufficient.

For ordination, Wallace (1995) provided an MML procedure for estimating the number of factors, the factor loadings and the factor scores simultaneously and consistently in linear factor analysis. We do not necessarily need different procedures for all alternative axis models. For example, Legendre and Gallagher (1999) showed that it is possible to preprocess the data and change the underlying implied metric for the embedding space, while Ihm and Groenewoud (1975) used a logarithmic transform to approximate unimodal response curves.

It would be possible to fit a multi-axis model to the one-cluster case and compare this to a mixture model with a simpler within-cluster axis model, perhaps even one which assumes that there is no correlation, and hence no axes, within clusters. It seems to me to be fairer to allow all clusters to use the same within-cluster model class, and allow the analysis to determine the number of axes within each cluster, including none. To use an arbitrarily complex model in fitting a continuum theory but to limit the cluster model would not, I suggest, be a fair test. In any case, the mechanism for comparison is the same – we would accept that model which has the shortest message length.

Finally, there is a question of finding the optimal solution. The search space is very large, and heuristic methods are employed, and it is necessary to use multiple random starts to avoid local minima. On the positive side, it is sometimes possible to prove that a procedure for induction converges to the optimal solution in the limit of very large amounts of data. More precisely (Barron and Conover 1991), if the data come from a finitely complex model within the family considered then, with probability one, it will be discovered for sufficiently large sample size; the rate of convergence is not predicted. MML in principle will discover the optimal solution given sufficient data, as when estimating fractal dimensionality, may necessitate several thousands of samples (Ramsey and Yuan 1990).

Analyses and data

For the example analyses presented here, I have adopted an overly simplistic model, which assumes that there is *no* within-cluster correlation of attributes. The distribution within a cluster is regarded as random variation obtained by sampling from a Gaussian distribution. From this, it is obvious that I do not regard this paper as providing a proper test of the two competing hypotheses, only as an example of the manner in which such a test might be performed. If there exists within-cluster correlation, the results will favour clustering. I shall later examine two more realistic models of within cluster variation, which I am presently investigating.

For each analysis, the results obtained include the message length for a 1-cluster (i.e., the continuum) solution, an estimate of the optimal number of clusters, their sizes, and means and standard deviations for all species in all clusters, as well as a (fuzzy) assignment of all samples to all clusters.

I have used 5 sets of data to investigate the choice between 1-cluster and multi-cluster solutions. These are:

1. Successional data from transects in transitional boreal forest near Sudbury, Ontario, Canada, with vegetation affected by pollution. The data consist of 1200 samples recorded in two consecutive years. The samples are arranged in transects of 100 quadrats, each transect running down a south-facing slope. There are 2 transects in each of 6 plots with the latter arranged along a pollution intensity gradient. In all, 119 understorey species were recorded using a 0-7 cover-abundance code. For further details, see Tucker and Anand (2002) or Desrochers and Anand (2003).

2. Calcareous grassland data from Slovakia (Dale et al. 2001) from a short transect aligned along a putative moisture gradient and containing 22 plots and 46 species. Again, a cover-abundance scale was used. This quantity of data is too small for secure induction where several thousands of samples are to be preferred. With these data I experimented with various transformations suggested by Legendre and Gallagher (1999) which manipulate the underlying metric so that, in addition to Euclidean, I used chord, chi-square and Hellinger distances. I also examined a solution involving Poisson variation within clusters rather than Gaussian (see also Dale 2001). The Poisson model requires specification of a single parameter rather than the two needed for Gaussian, and hence, is simpler.

3. Mallee data from Victoria, Australia, from 256 plots arranged in a stratified random sample. Percent cover values were estimated for 62 species and forms. With these data I analysed presence/absence and abundance data, and also examined a reduced species set (using only the 32 commonest species). For further details, see Goodall (1953). (See Dale 2001, for application of Poisson distribution to the mallee data)

4. Salt marsh data from Queensland, Australia, with 30 plots recorded quarterly for 14 years giving a total of 1680 plots. Two species were present and for

Table 1. MML clustering results for the 5 data sets.

Data	Number of observations	1-class ML	Number of clusters	n-class ML	% capture
Transitional boreal forest	2400	85266.2	29	34285.7	59.8
Calcareous grassland	22	1756.2	3	1424.7	18.9
Goodall presence -absence data	256	3507.6	3	2995.5	24.6
Goodall abundance data	256	45550.4	17	19968.1	56.2
Goodall reduced species	256	33631.2	18	16457.8	51.1
Salt marsh	1680	42355.7	11	28158.1	33.5
Danish grassland	620	232803.3	27	137347.2	41.0

each density and a vitality measure were recorded. For further details, see Dale and Dale (2002), Dale et al. (2002a, b) and Li et al. (2002).

5. Erjnæs and Bruun (2000, see also Bruun and Erjnæs 2000) provide data for 620 samples from 180 localities in Danish grasslands. In total, 387 species were recorded using a frequency measure. The samples were stratified to cover several important gradients.

For the Sudbury data, I used over 1000 random starts in order to get some grasp of the variation between analyses of the same data. and to be reasonably certain that a global optimum had been reached. For the remainder only a few random starts were used, so it is possible that still better solutions could be found.

Note that both the boreal forest and salt marsh datasets involve sampling through time as well as space, so that there may be temporal dependency between the observations. For the boreal forest data, the calcareous grassland data and the Danish grassland data spatial dependency may also be present since the sampling is what Gillison and Brewer (1985) term gradsects. In the present analyses such possible dependencies have been ignored although in principle they can be accommodated in the MML procedure.

For all the datasets, we are ignoring problems with scale and the possibility that descriptions other than species performance might be more effective. Hájek and Havránek (1977) note that using a relevant description is critical in obtaining a comprehensible solution.

Results

The general results are presented in Table 1 except for the special case of the calcareous grassland data which are presented in Table 2. These tables both show the number of samples, the message length for a one-cluster solution, the optimal number of clusters, the message length for the optimal cluster solution, and the reduction in message length as a percentage of the 1-class length.

Throughout the *n*-class mixture solutions are preferred to the 1-class solutions. It is apparent also that even the best cluster results do not always capture much structure. The calcareous grassland data set is the worst, while in the mallee data the presence/absence result is considerably worse than the quantitative solutions. Brokaw and Busing (2000) have previously indicated the importance of chance variation in forest dynamics, and it seems that this can be extended to other vegetation types. As might be expected, the larger datasets tend to produce more clusters.

1. Sudbury transitional boreal forest data

Using 1000 random starts, we obtain 2 competing solutions one with 29, the other with 30 clusters. The difference in message length is about 1.5, which is non-significant, and both are well separated as outliers from other results with the same number of clusters. The ambiguity in the result was unexpected and illustrates that large amounts of data may be necessary to identify optimal partitions. The two solutions generally agree on allocating species-poor samples to particular clusters but species-rich samples are assorted somewhat differently. However, since I am only interested in the 1-cluster / *n*-cluster comparison, here I show only the solution with the shortest message length. In either case, the solution favours mixtures of clusters, and hence, the community unit theory.

2. Calcareous grassland

Here we have several analyses, each with a different underlying metric and also the Poisson-Gaussian comparison. It is clear that the mixture solution is again pre-

Table 2. Slovak data message lengths for various models.

Model and data	1-cluster length	Number of clusters	n-cluster length	% capture
Poisson	1524.3	4	1213.3	20.4
Gaussian (squared Euclidean distance)	1756.2	3	1424.7	18.9
Chord distance	10611.4	4	7951.0	25.1
$\chi 2$ distance	7845.3	4	5786.0	26.2
Hellinger distance	19776.3	4	8646.6	56.3

ferred, but the optimal solution was obtained using Poisson within-cluster variation. Dale (2001) found the Gaussian preferable to Poisson for the mallee data, and this might be due to the difference between ordered category codes and real estimates of % cover, although it is possible that the distribution is a property of the vegetation. It is also interesting that the chi-square metric, which underlies correspondence analysis, is far from optimal as a mixture solution nor does it provide the best 1-cluster (ordination) solution. It would be interesting if this poor performance was found in other datasets.

3. Mallee

All three analyses choose the mixture solution. It is clear that the presence/absence data, which uses a Dirichlet distribution not a Gaussian, provides a much simplified solution with only 3 clusters. However, the clusters in the other analyses tend to be nested within the presence classes. Removing rare species does decrease the message length, but does not appear to change the number of clusters much, in fact, producing one more than the full data! The result may be suboptimal, and a wider search is needed to check this.

4. Salt marsh

The best result here involves a mixture of 11 clusters. However, these do not necessarily represent environmentally distinguishable communities. Some at least appear to be variations in time generated by vegetation processes of growth and replacement largely independent of environmental changes. Dale et al. (2002a), using an analysis which incorporated temporal dependency between samples, also found 11 clusters, most of which were similar to the clusters found here.

5. Danish grassland

Here the results identified 27 classes which is close to the 25 identified by Bruun and Erjnæs (2000) using Indicator Species Analysis and a permutation test. The MML solution estimates the required number of clusters directly. No detailed interpretation will be presented here, but the clusters appear to be highly correlated with the locations sampled, although some did occur at several places. If this is substantiated, it means that local conditions dominate and any communities would tend to be site-specific rather than universal abstractions.

Discussion

It is clear that in all the analyses using the within-cluster model chosen, a mixture model is always preferred. This is *not* a proof that the community unit model is to be preferred and the continuum model rejected. The within-cluster model is clearly oversimplified, and biased against continuum theory. As noted earlier, the existence of correlation between attributes within classes would lead to the identification of several classes. The results do show how a test of the competing hypotheses can be performed if more acceptable within-cluster models are provided. We need therefore to examine what alternatives are, or may soon be, available. Two possibilities will be presented here, for which the computational problems are becoming tractable.

Before examining the possibilities, it is also pertinent to note that we might question whether a single model of either kind is attainable. We may desire a single causative model, but in the light of Pagie and Hogeweg's (1999) work, this may be a forbidden fruit. Multiple causation is common enough in biology.

Static solution

The clusters developed by the analysis used here have only random variation, Gaussian for the most part. This is hardly acceptable for continua where several gradients may be expected. Even a single axis of variation would be preferable. For the mixture alternative, variation within clusters can also be allowed, although it would be desirable that the existence and number of any axes of variation

within clusters should be determined from the data rather than imposed. A method which goes some way towards meeting these requirements is that of Wallace and Freeman (1992) which permits a single linear factor, and an extension to multiple axes per cluster is theoretically possible (Wallace 1995). However, the practical use of multiple factor solutions has yet to be investigated especially with large numbers of variables. Hanson et al. (1991), in their AUTOCLAS program, do permit multivariate within-cluster axis models, as does independent factor analysis (Attias 1999). However, there seem to be a limit to 7 or less dimensions if sensible solutions are to be found, similar to those found with intrinsic dimensionality calculations (Trunk 1976). In any case high dimensional solutions would require *very* large amounts of data.

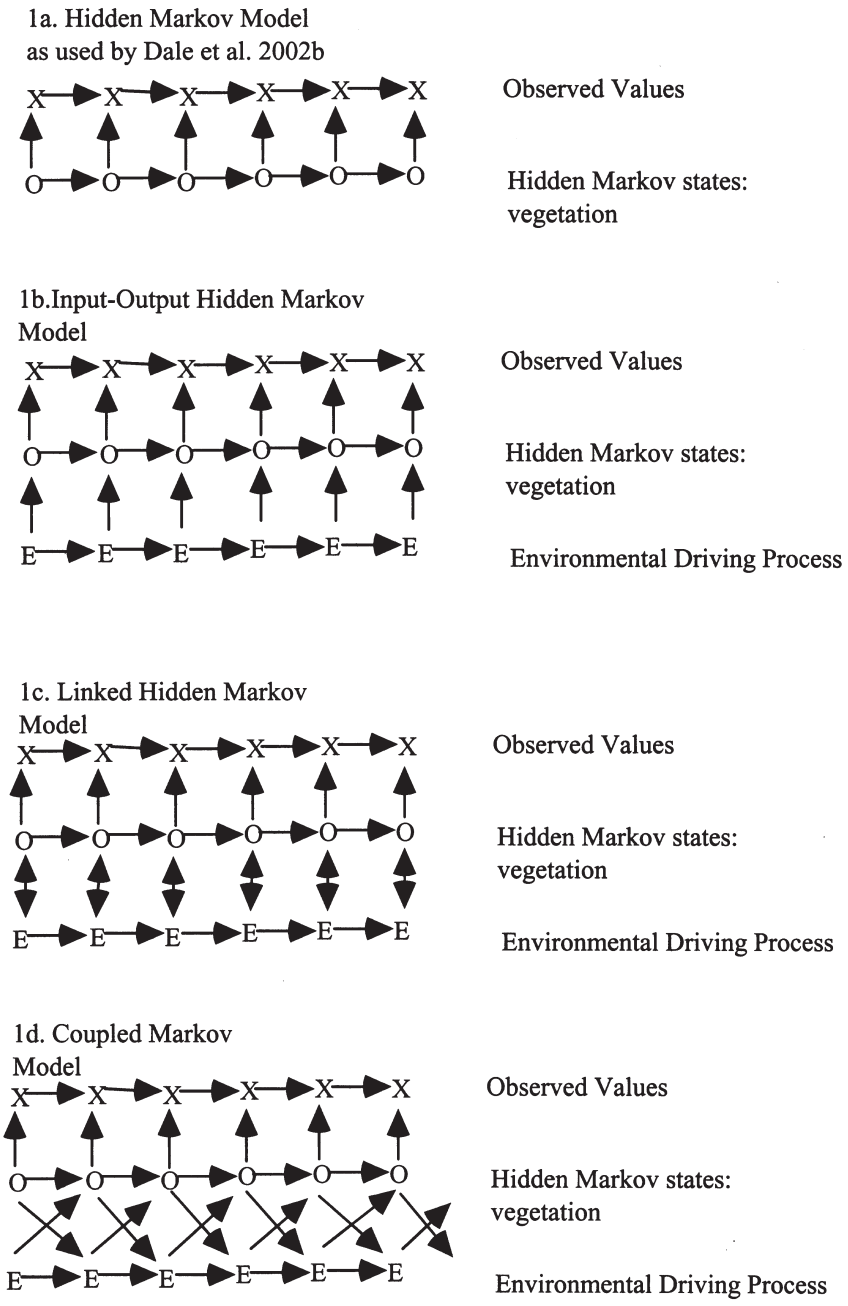

Figure 1. Some varieties of hidden Markov models applicable to studies of vegetation change. Arrows indicate direction of causal influences. O represents the state of the hidden process at some time, X the observed variables at that time and E states of an associated, and potentially causal, environmental process.

With multiple factors, it would be desirable to rotate the solutions within clusters to permit easier interpretation. Kiers' (1994) SIMPLIMAX appears to be a promising possibility for this, leading directly to an oblique simple structure solution. It would also be possible to look for axes common to subsets of clusters, which might reflect some hierarchical or other cross-cluster structure.

Dynamic solution

All of these models suffer from a major disadvantage; they are all static. It should be obvious that any structure based on a snapshot of vegetation at a particular time will be useful for prediction only if any pattern is generated by a first order process, that is by a memory-less system. Lippe et al. (1985) concluded that at Dwingelose Heide this assumption was unacceptable, but did not investigate higher order or non-Markov possibilities. In contrast, Orlóci et al. (1993) found the first order Markov model adequate, although Anand and Orlóci (1997) found adding noise to perturb the Markov process provided better modelling of transient responses.[2] The assumption that the future course of a system is solely dependent on the present state itself requires testing. The methods used to distinguish between community and continuum models can be extended to discriminating between more general models where appropriate observational data are available; for example Dale et al. (2002a) found a first order process sufficient to describe temporal processes in their salt-marsh vegetation.

Dale et al. (2002b; see also Li et al. 2002) have examined the use of dynamic models within clusters. In their example, a single cluster solution proved adequate, although a 2-cluster solution was obtained when only a subset of the data were used and this corresponded reasonable well with the concepts of 'high' and 'low' marsh communities. They fitted hidden Markov models (Fig. 1a) using all the variables, but more appealing options are available. For example, it might be useful to introduce environmental variables to act as drivers for the hidden Markov processes, so-called input-output models (Fig. 1b). But to better model the continuum approach, we might proceed as follows.

Assume that we build a hidden Markov model for each species separately. These models can be linked (Fig. 1c) or coupled (Fig. 1d) so that one species modifies the responses of other species, at some later time. In principle, we can ask if the addition of coupling, which is an added complexity, is adequately rewarded by improvements in fit, and also whether there is evidence for different degrees of coupling in different clusters. The continuum theory argues for independence of species so the existence of coupling would be evidence against it. Existence of clusters with differing coupling coefficients would strengthen evidence for the community theory. The difficulty here is computational, for fitting coupled Markov processes is heavily data dependent and computationally demanding (e.g., Davis et al. 2002). The results obtained often depend on the effectiveness of the training method. A first attempt (Dale and Davis, unpublished) using coupled HMMs with the salt marsh data suggested that coupling was not beneficial. We are presently investigating this further as well as the possibility that non-Markovian processes might be preferable. Still more complex models allowing for processes operating at several different time scales can also be considered, but the computational problems are then still more acute.

Choosing between the continuum and community unit theory is not, then, a question of *a priori* predilection. Instead, if the models are specified sufficiently, it is possible to test formally and identify the preferable model for any particular dataset. Whether such a decision is worth the sampling effort required is another matter.

Acknowledgments: To Madhur Anand, Laco Mucina, David Goodall, and Rasmus Erjnæs, my thanks for permission to use their data. They are not responsible for any misuse or misinterpretations. To R. Davis, my thanks for assistance with coupled hidden Markov models.

References

Anand, M. and Orlóci, L. 1997. Chaotic dynamics in a multispecies community. *Ecological and Environmental Statistics* 4: 337-344.

Attias, H. 1999. Independent factor analysis. *Neural Computation* 11: 803-851.

Bar-Yam, Y. 2002. Sum rule for multiscale representations of kinematically described systems. *Advances in Complex Systems* 5: 409-431.

Barron, A. R. and Conover, T. M. 1991. Minimum complexity density estimation. *I. E. E. E. Trans. Inform. Theory* 37: 1034-1054.

Boerlijst, M. C. 2000 Spirals and spots: novel evolutionary phenomena through spatial self-structuring. In: U. Dieckmann, R. Law and H. Metz (eds.), *The Geometry of Ecological Interactions: Simplifying Spatial Complexity*, Cambridge University Press, Cambridge, pp. 171-182.

2 Such variation might better be derived from a hidden Markov process, where the observed values of the features are a result of a random generation from some distribution associated with each of the states of the latent process. Adding noise is also a means of avoiding local optima in search procedures.

Boulton, D. M. and Wallace, C. S. 1970. A program for numerical classification. *Comput. J.* 13: 63-69.

Boulton, D. M. and Wallace, C. S. 1973. An information measure for hierarchic classification. *Comput. J.* 16: 254-261.

Brokaw, N. and Busing, R. T. 2000. Niche versus chance in tree diversity in forest gaps. *TREE* 15: 183-188.

Bruun, H. H. and Erjnæs, R. 2000. Classification of dry grassland vegetation in Denmark. *J. Veg. Sci.* 11: 585-596.

Dale, M. B. 1994. Straightening the horseshoe: a Riemannian resolution? *Coenoses* 9: 43-53.

Dale, M. B. 2000. On plexus representation of dissimilarities. *Community Ecology* 1: 43-56.

Dale, M. B. 2001. Minimal message length clustering, environmental heterogeneity and the variable Poisson model. *Community Ecology* 2: 171-180.

Dale, M. B. 2002. Models, measures and messages: an essay on the role of induction. *Community Ecology* 3: 191-204.

Dale, M. B. and Anderson, D. J. 1973. Inosculate analysis of vegetation data. *Austral. J. Bot.* 21: 253-276.

Dale, M. B., Dale, P. E. R. and Edgoose, T. 2002a. Markov models for incorporating temporal dependence. *Acta Oecologica* 23: 261-269.

Dale, M. B., Dale, P. E. R., Li, C. and Biswas, G. 2002b. Assessing impacts of small perturbations using a model-based approach. *Ecol. Modell.* 156: 185-199.

Dale, M. B., Salmina, L. and Mucina, L. 2001. Minimum message length clustering: an explication and some applications to vegetation data. *Community Ecology* 2: 231-247.

Dale, P. E. R. and Dale, M. B. 2002. Optimal classification to describe environmental change: pictures from the exposition. *Community Ecology* 3: 19-30.

Davis, R. I. A., Lovell, B. C. and Caelli, T. 2002. Improved estimation of hidden Markov model parameters from multiple observation sequences. In: R. Kasturi, D. Laurendeau and C. Suen (eds.), *Proc. Internatl. Conf. Pattern Recognition*, August 11-14 II, Quebec City, Canada. pp. 168-171.

Desrochers, R. E. and Anand, M. 2003. The use of taxonomic diversity indices in the assessment of perturbed community recovery. In: *Proc. 4th Internatl. Conf. Ecosystems and Sustainable Development*, June 4-6, 2003, Siena, Italy. WIT Press, Southampton.

Domingos P. 1999. The role of Occam's Razor in knowledge discovery. *Data Mining and Knowledge Discovery* 3: 409-425.

Edwards, R. T. and Dowe, D. 1998. Single factor analysis in MML mixture modelling. *Lecture Notes in Artificial Intelligence* 1394, Springer Verlag, Berlin, pp. 96-109.

Erjnćs, R. and Bruun, H. H. 2000. Gradient analysis of dry grassland vegetation in Denmark. *J. Vèg. Sci.* 11: 573-584.

Fisher, D. H. 1992. Pessimistic and optimistic induction. TR CS-92-12 Dept. Comput. Sci., Vanderbilt Univ.

Gamberger, D. and Lavra, N. 1997. Conditions for Occam's razor applicability and noise elimination. In: *Proc. 9th European Conf. Machine Learning*. Springer Verlag. pp. 108-123.

Gilbert, N. and Wells. T. C. E. 1966. Analysis of quadrat data. *J. Ecol.* 54: 675-686.

Gillison, A. N. and Brewer, K. R. W. 1985. The use of gradient directed transects or gradsects in natural resource surveys. *J. Environ. Manage.* 20: 103-127.

Goodall, D. W. 1953. Objective methods for the classification of vegetation: the use of positive interspecific correlation. *Austral. J. Bot.* 1: 39-63.

Hájek, P. and Havránek, T. 1977. On generation of inductive hypotheses. *International. J. Man-Mach. Stud.* 9: 415-438.

Hanson, R. Stutz, J. and Cheeseman, P. 1991. Bayesian Classification with Correlation and Inheritance. In: *Proc. 12th International Joint Conference on Artificial Intelligence*. Sydney, Australia. August 24-30. Morgan Kaufmann, San Francisco. pp. 692-698.

Hogeweg, P. 2002. Computing an organism: on the interface between informatic and dynamic processes. *BioSystems* 64: 97-109.

Ihm, P and van Groenewoud, H. 1975. A multivariate ordering of vegetation data based on Gaussian type gradient response curves. *J. Ecol.* 63: 767-777.

Jelinski, D. E. and Wu, J-G. 1996. The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology* 11: 129-140.

Kiers, H. A. L. 1994. SIMPLIMAX: oblique rotation to an optimal target with simple structure. *Psychometrika* 59: 567-579.

Kreinovich, V. and Kunin, I. A. 2003. Kolmogorov complexity and chaotic phenomena. *Internatl. J. Engineering Science* 41: 483-493.

Legendre, P. and Gallagher, E. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* 270: 271-280.

Li, C., Biswas, G., Dale, M. B. and Dale, P. E. R. 2002. Matryoshka: A HMM based temporal data clustering methodology for modelling system dynamics. *Intelligent Data Analysis Journal* (in press)

Lippe, E., de Smidt, J. and Glenn-Lewin, D. 1985. Markov models and succession: a test from a heathland in the Netherlands. *J. Ecol.* 73: 775-791.

Neal, R. M. 1998. Markov chain sampling methods for Dirichlet process mixture models. Tech. Rep. 9815, Department of Statistics, Univ. Toronto.

Neil, J. R., Wallace, C. S. and Korb, K. B. 1999. Bayesian networks with non-interacting causes. Tech. Rep. 1999/28, Dept. Computer Science, Monash University, Melbourne.

Openshaw, S. 1984. The modifiable areal unit problem. CATMOG 38. GeoBooks, Norwich, England.

Orlóci, L., Anand, M. and He, X. S. 1993. Markov chain: a realistic model for temporal coenosere? *Biométrie-Praximétrie* 33: 7-26.

Pagie, L. and Hogeweg, P. 1999. Colicin diversity: a result of eco-evolutionary dynamics. *J. Theoret. Biol.* 196: 251-261.

Posse, C. 1995 Projection pursuit exploratory data analysis. *Computat. Statist. Data Anal.* 20: 669-687.

Ramsey, J. B. and Yuan, H-J. 1990 The statistical properties of dimension calculations using small data sets. *Nonlinearity* 3: 155-176.

Rietkerk, M., Boerlijst, M. C., van Langevelde, F., HilleRisLambers, D. , van der Koppel, J., Kumar, L. Prins, H. H. T. and de Roos, A. M. 2002. Self-organization of vegetation in arid ecosystems. *Amer. Natur.* 160: 524-530.

Rissanen, J. J. 1978. Modelling by shortest data description. *Automatika* 14: 465-471.

Rissanen, J. J. 1987. Stochastic complexity. *J. Royal Statist. Soc. B* 49: 223-239

Rissanen, J. J. 1996. Fisher information and stochastic complexity. *I. E. E. E. Trans. Information Theory* 42: 40-47.

Shalizi, C. R., and Crutchfield, J. P. 1999. Computational mechanics: Pattern and prediction, structure and simplicity. Sante Fe Institute Working Paper 99-07-044.

Shipley, B. and Keddy, P. A. 1987. The individualistic and community-unit concepts as falsifiable hypotheses. *Vegetatio* 69: 47-55.

Stone, J. V. and Porrill, J. 1998. Independent component analysis and Projection Pursuit: a tutorial introduction. Available as file ica_tutorial2.tex from www.shef.ac.uk/psychology/stone

Trunk, G. V. 1976. Statistical estimation of the intrinsic dimensionality of data collections. *Inform. Control.* 12: 508-525.

Tucker, B. C. and Anand, M. 2003. The use of matrix models to detect natural and pollution-induced forest gradients. *Community Ecology* 4: 89-100.

Uebersax, J. S. and Grove, W. M. 1993. A latent trait finite mixture model for the analysis of rating agreement. *Biometrics* 49: 823-835.

Wallace, C. S. 1995. Multiple factor analysis by MML estimation. Tech. Rep. 95/218, Dept Computer Science, Monash University, Clayton, Victoria 3168, Australia. 21 pp.

Wallace, C. S. 1996. MML Inference of predictive trees, graphs and nets. In: Gammerman, A. (ed.), *Computational Learning and Probabilistic Reasoning*, John Wiley, London. pp. 43-66.

Wallace, C. S. 1998. Intrinsic classification of spatially-correlated data. *Comput. J.* 41: 602-611.

Wallace, C. S. and Dowe, D. L. 2000. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing* 10: 73-83.

Wallace, C. S. and Freeman, P. R. 1992. Single-factor analysis by minimal message length estimation. *J. Roy. Statist. Soc.* B 54:195-209.

Wallace, C. S. and Georgeff, M. P. 1983. A general objective for inductive inference. Tech. Rep. 32, Dept. Computer Science, Monash University, 3168 Australia.

Wallace, C. S., Korb, K. B. and Dai, H. 1996. Causal discovery via MML. Tech. Rep. 96/254 Dept. Computer Science, Monash University, Clayton, Victoria 3168, Australia.

Webb, G. I. 1996. Further experimental evidence against the utility of Occam's Razor. *J. Artif. Intell. Res.* 4: 387-417.

Wisheu, I. and Keddy, P. A. 1992. Competition and centrifugal organisation of plant communities: theory and tests. *J. Veg. Sci.* 3: 147-156.

Young, P., Parkinson, S. and Lees, M. 1996. Simplicity out of complexity in environmental modelling: Occam's razor revisited. *J. Appl. Statist.* 234: 165-210.