COMMUNITY ECOLOGY 2(2): 231-247, 2001
1585-8553 © Akadémiai Kiadó, Budapest

Minimum message length clustering: an explication and some applications to vegetation data

M. B. Dale[1], L. Salmina[2] and L. Mucina[3]

[1] *Australian School of Environmental Studies, Griffith University, Nathan, Qld 4111 Australia. Tel.: 07 33714414, Fax: 07 38705681, Email: m.dale@mailbox.gu.edu.au*
[2] *Department of Botany and Ecology, University of Latvia, 4 Kronvalda Blvd, LV1586 Riga, Latvia*
[3] *School of Life Sciences, University of the North. Qwa-Qwa Campus, Private Bag X13, 9866 Phuthaditjhaba, South Africa*

Keywords: Fuzzy clustering, MML principle, Qualitative data, Quantitative data, Snob program.

Abstract: In this paper, we examine the application of a particular approach to induction, the minimum message length principle and illustrate some of the problems that can be addressed through its use. The MML principle seeks to identify an optimal model within some specified parameterised class of models and for this paper we have chosen to concentrate on a single model class, that of mixture separation or fuzzy clustering. The first section presents, in outline, an MML methodology for fuzzy clustering. We then present some applications, including the nature of the within-cluster model, examination of the univocality of results for different groups of species and the effectiveness of presence data compared to purely quantitative data. Finally, we examine some possibilities of extending MML methodology to include within-class correlation of species, the existence of dependence between observed samples and the comparison of different classes of models.

Abbrevations: MML - Minimum Message Length, MDL - Minimum Description Length

Introduction

There are various means by which models can be assessed for inductive purposes (Dale submitted). Commonly these include such items as the balancing of complexity against simplicity to reduce the effects of overfitting, and balancing the coarseness of parameter estimates against precision of fit to avoid the use of overly precise estimates. In the present paper we want to concentrate on the minimum message length principle of Wallace & Freeman (1987) and Wallace & Dowe (2000), and to illustrate some of the problems which can be addressed through its use. This principle is very simple in concept., From the class of models we are considering, we choose the model that permits us to encode both data AND model in the most concise manner. The coded message is composed of two parts. The first provides estimates of the unknown parameters of the model using a code optimal for the given prior distribution, in the sense of shortest message length. The second then describes the data using a code that would be optimal were the estimates correct. It is not necessary to construct the coded message since we need only calculate the length of the relevant messages to make comparisons between models.

Any patterns derived from a set of data must be dependent on the sampling scheme and the size and shape of the plots selected. Random sampling is not required by the induction method although the results will reflect any selection bias. Increasing the number of samples used will not remove such bias but should lead, asymptotically, to convergence on the optimal solution for the given data. Similarly, finding structure in data presupposes that we have made observations on some set of attributes. Ecologically, any inferences to be drawn are necessarily a function of categories, such as species or growth forms, and properties, such as presence or biomass, used to describe the samples. MML does not change these dependencies on scale and description. We shall assume that the data are in the form of a rectangular matrix showing the abundance of species in sites. The MML literature uses the term ‘things’ for the items to be clustered, and ‘attributes’ for the descriptors used and this practice will be adopted here.

Minimum message length clustering

How many clusters

Clustering involves choosing models from a class of models where the number of clusters is the single parameter to be estimated. The null hypothesis is simply that there is one cluster only; no special status for the null hypothesis is needed. Present methods for estimating the number of clusters leave much to be desired. For example, for TWINSPAN (Hill et al. 1975) the only strict criterion applied is that clusters must not be too small! Ignoring such size limits, the number of clusters must be an integer power of 2, a most curious restriction to impose. Combined with the breadth-first search used, this can result in large heterogeneous clusters remaining undivided while small clusters are broken into minute pieces. Users must rely on subjective evaluation of the meaning of the clusters (cf. Austin 1970) except when the data have special properties. Thus, the Sandland & Young (1979) test relies on replication of samples, while the Krishna-Iyer (1949) test presupposes the acceptance of spatial coherence as a desirable characteristic of the results.

Various suggestions have been made for determining the number of clusters to be retained (see Dale, 1987, Gordon 1994) but, because the calculated statistics are extremal values, normal significance levels are inappropriate. More recently data-based tests have been suggested with Pillar (1996) using bootstrapping and Boik (1987) using permutation tests, but still some caution is necessary. Hayes (1996) has noted that permutation methods are NOT always distribution-free. They may not require Gaussian distributions but do assume that any distributions being compared are the same. Cross-validation is another possibility, although this is known to be equivalent in the limit to using Akaike's (1978) coefficient.

The MML principle supplies a means of estimating the correct number of clusters. As the name suggests this is just the number of clusters which minimises the length of message needed to describe the data adequately. MML can be regarded as a means of estimating parameters so as to maximise the posterior probability of the estimate (Wallace and Dowe 2000). It uses prior probabilities but, unlike the standard Bayesian maximal a posteriori theory, it optimises a probability not a probability density and is invariant under 1-1 re-parameterisation. Instead of maximising the posterior probability directly, MML makes use of a conversion to message length; an event of probability p corresponds to a message length $\lambda = -\log_2(p)$. The procedure is then to minimise the overall message length. This is equivalent to finding the length of the shortest program which, together with noise, will generate the data (Chaitin 1966). MML can also be regarded as one way of estimating the Kolmogorov complexity (Kolmogorov, 1965) and of balancing complexity of a model against its adequacy of fit.

Thus, if we seek to estimate the mean and variance of a Gaussian distribution, which is equivalent to a 1-cluster solution for numeric variables with a Gaussian within cluster distribution, we find that, for the mean, the estimator is identical with the maximum likelihood estimator; $\bar{x} = \Sigma_i (x_i) / N$.

If we let $s^2 = \Sigma_i (x_I - \bar{x})^2$, the maximum likelihood estimator for the variance, when the mean has also to be estimated, is s^2 / N, which is known to be biased. The MML estimator is $s^2 / (N\text{-}1)$. Similar estimators can be obtained for discrete multistate variable probabilities, for the rate parameter of Poisson variables and for the mean and concentration parameters of circular von Mises distribution (Wallace and Dowe 2000). In mixture modelling, the number of parameters grows with the data and the number of clusters. Maximum likelihood can become inconsistent (or very inefficient) with such problems whereas the MML estimates remain consistent.

The message

As noted above, the message length to be minimised has two components. In the hypothesis component of the message there are 3 parts which are:

- *The number of component clusters. A priori*, all numbers up to some constant are assumed equally likely.
- *The relative abundance of each component.* This provides labels for the clusters that can be used to indicate the assignment of things. These are encoded as a multinomial distribution.
- *The cluster description.* The distribution parameters appropriate to each attribute for each component. Each parameter, except for multistate attributes, is considered to be specified to a precision of the order of its expected estimation error, so well-measured components have parameters specified with greater precision. This is significant since ordered category data is often employed in ecological studies, and such data are very coarsely measured.

The *fit-to-data component* is calculated as follows. For each thing we identify the cluster(s) to which it assigned, together with the relative probability of that assignment. Although the original program for Snob (Boulton and Wallace 1970) crisply assigned things to clusters, and hence was a segmentation procedure, it has since been de-

termined that fuzzy, probabilistic, assignment in fact generally leads to shorter message lengths (Wallace 1990). In practice, each thing is allocated (either pseudo-randomly or randomly) to a single class chosen from the posterior distribution. The associated uncertainty can then be used to provide information on the succeeding thing. With random assignment, an extension of this procedure can be used to sample the number of clusters in proportion to its posterior, and given sufficient time the estimation process will converge (Richardson and Green 1997).

The first 3 parts specify our hypothesis, the 4th part provides a possible encoding cost for our data. A more complex hypothesis may produce better data encoding but the extra cost incurred to code the hypothesis may exceed any gain in coding cost because of better fit. Thus, the optimal choice involves a balance between hypothesis complexity and fit. In addition, adjustments are made to balance the cost of high precision of expression of parameters to the quality of approximation of the data.

The mechanics of MML

The MML principle proposes that we measure the quality of a model by determining the minimal message length needed to transmit the data using an optimal code. Changes to the model are evaluated through changes in this message length. To measure the length of the message we have to combine the several parts. There may be other components of a full message but these are of constant length so we can ignore them when comparing model solutions. For any probability p, the message length required is $-\log_2(p)$ bits, so we have, for model H and data D, Message Length = $-\log_2(p(H)) - \log_2(p(D|H))$.

Minimising the message length is equivalent to maximising $Pr(H).Pr(D|H)$ which by Bayes rule equals $Pr(D).Pr(H|D)$ and, since $Pr(D)$ is independent of H, MML maximises the Bayesian posterior probability $Pr(H|D)$. More precisely, assuming a locally flat prior and a quadratic likelihood function, we have

$$E(\text{Message Length}(y, \theta)) = -\log(h(\theta)) - \log(f(y, \theta)) + \text{precision terms},$$

where $h(\theta)$ is the assumed known prior density on θ and $f(y|\theta)$ is the likelihood of y given θ. This expression remains an approximation and there may be modifications necessary in other applications of the principle, for example, in factor analysis.

The expressions used for determining the optimal encoding depend on the distribution. Thus to transmit K values from an M-state multistate distribution, assuming a uniform prior over the possible combinations for the frequency of observed variables is given by

$$\text{MessLen(H\&D)} \approx (M-1)/2\ (\ln(K/12)+1) - \ln(M-1)! - \sum_{m=1}^{M}(n[m]+1/2)\ln(p[m]),$$

where $n[m]$ is the number of values in state m and $p[m]$ is the probability stated for state m estimated as

$$p[m]=(n[m]+1/2)/(K+M/2).$$

To transmit K values from a normal distribution where the values are continuous real and stated to a specified accuracy ε, with mean μ and standard deviation σ and a global distribution μ_p, with a uniform prior in the range $\mu_p, \pm 2\ \sigma_p$ and standard deviation σ_p with $\ln(\sigma_p)$ having a uniform prior in the range $\ln(\varepsilon)$ to $\ln \sigma_p\ (2\pi)^{0.5}$ is

$$\text{ML(H\&D)} \tilde{} -\ln(4\sqrt{\{K/12\}}\ .\ \sigma_p/\sigma)) - \ln\{\ln(\sqrt{(2\pi)}\ \sigma_p/\varepsilon)\ \sqrt{((K-1)/6)} - K\ln(\sigma\sqrt{(2\pi)}/\varepsilon + 1/2) + 1/2.$$

Similar expressions can be derived for a Poisson distribution, with parameter α, and for the circular von Mises distribution with mean μ and concentration κ (see Wallace and Dowe 2000).

The Snob program

The MML clustering algorithm is implemented in the Snob program, first presented by Boulton and Wallace (1970, 1973) and in updated form by Wallace and Dowe (2000). The program is written in FORTRAN and is available for not-for-profit academic research use from http://www.cssw.monash.edu.au/~dld/Snob.html. Documentation is available at the same site.

The input to Snob starts with a description of the attributes which indicates their inclusion or exclusion and their type. Excluded attributes are not used during the clustering procedure itself but their significance to the final clusters is evaluated. Attributes may be any 1 of 4 types - multistate, numeric-Gaussian, numeric-Poisson or Circular (von Mises). Each type requires some additional information; for multistates, the number of states, for numeric and angular, the precision of recording and for angular, whether the recording is in degrees or radians. The data themselves then follow. Each thing is given a unique reference number and a list of attribute values; a reference value of zero terminates the input. If the reference number is negative the thing is not used to form clusters although it will be notionally assigned. Missing values are permitted in the data (a - sign suffices), and these are assumed to be at random. The program also assumes no correlation

of attributes within clusters, and that the things are independent samples. Similar assumptions are found in almost all other clustering methods and we shall discuss ways of relaxing them later.

The program structures the things into clusters, the number of which it estimates. Each thing is, conceptually, assigned a relative probability of belonging to every cluster although output is suppressed if this probability is $p \leq 0.01$. For each cluster, the attributes are examined to see if their parameters within that cluster differ significantly from the corresponding parameters of the whole population and thus whether they are contributing to the differentiation of the cluster. As far as assessing significance is concerned, the message length is actually calculated in nits (using $\log_e$ in place of $\log_2$), and a change of 10 nits represents considerable significance; it indicate odds of more than 22000:1 in favour of the model with the shorter length!

The message information is composed from the components by summation of the message lengths associated with each of them. The overall accuracy of the estimate of the message length is approximately 1 nit.

The algorithm involves reallocation of things between clusters using an EM algorithm, as in *k*-means clustering, and splitting and merging of clusters, all evaluated by their effects on the message length. Splitting is investigated by maintaining subclusters within all sufficiently large clusters and investigating if using these produces a reduction in message length. Merging involves trials of all pairs of clusters and accepting a merge if the message length is reduced. To identify potentially useful actions, both merging and splitting initially assume that only things in the cluster(s) being examined will be affected by the changes. Once some candidates have been found using this approximation, a full evaluation is made. Initialisation is usually by specification of some arbitrary number of clusters with things randomly assigned. Alternatively, the user may supply an initial configuration.

Unfortunately, the implemented algorithm does not guarantee a globally optimal solution, in part because it does not sample the number of clusters from the appropriate *a posteriori* distribution. The reallocation procedure may also encounter local optima and several starting assignments are usually employed. Recently, as noted earlier, Richardson and Green (1997) have developed appropriate procedures (the Reverse Jump Monte Carlo Model Composition procedures) which do guarantee optimality given sufficient time; how long that time is remains subject to further study.

The program will not find very small clusters, those with fewer than 4 members, but they can often be detected by closer examination of the results. Within a cluster, the message lengths associated with things provide a means of identifying outlying members and these are *prima facie* candidates for small clusters.

Data and analyses

The analyses presented here are exemplary rather than substantive for it is hoped to examine the questions in more depth later. However, they do provide some evidence of the flexibility of the methodology in answering various questions of interest to a user and in establishing the impact of user choices.

Slovak data

It would be desirable for the within-cluster variation itself to be subject to estimation from some class of possible distributions. While this facility is not present in the program, Snob does provide for two different numeric attribute distributions within clusters, Gaussian and Poisson. If we have frequency data we can elect to use either. We have used data from Slovakian calcareous grasslands (L. M.) to examine within-cluster distribution. These data consist of 22 samples containing 46 species whose performance was recorded using a cover-abundance scale. The samples were originally thought to be arranged along a gradient, although Dale (2000a) has suggested that this is not so. We have numbered both things and attributes in order of the supposed gradient.

Two analyses were performed using these data, one assuming a Gaussian within-cluster distribution, the other a Poisson distribution. This is possible because the ordered category data can be interpreted either as a number or as a non-negative integer frequency counted out of 9, the maximal code value. The question of interest is which interpretation provides the most effective coding. We have assumed equal prior probabilities for Gaussian and Poisson models.

A second experiment was made to assess the effectiveness of embedding in spaces with different metrics. Given a dataset A, it is possible to transform it to another dataset B such that a Euclidean distance in A is equivalent to some other metric in B, such as a chi-square metric, a chord distance or a Hellinger distance (Legendre and Gallagher 2001)[1] . In order to make the analyses comparable,

1 Hellinger distance $= 1 - \sum_x \sqrt{p_k(x)}\,\sqrt{p_l(x)}$ for probability distributions p_k and p_l.

the Gaussian analysis was repeated using the same precision as these alternative distances. By identifying the metric providing the greatest reduction in message length, it is possible to decide which is most suited for capturing vegetation dissimilarity structure.

Latvian data

The second set of data pertains to a Latvian bog (L.S). The data consist of 48 stands of 1 m^2 from a single site, described by the % cover of 32 species, with the stands selectively chosen to provide a characterisation of the particular vegetation type. An analysis using the 'peeling' technique of Hoffman and Jain (1987)[2] indicates that these data might be close to a multivariate normal distribution. There is weak evidence (Dale unpublished), using methods developed by Hubert and Arabie (1994), for the existence of two gradients or perhaps a circular sequence.

These data are used in two different comparisons. First, Watanabe (1969) has proved in his 'ugly duckling' theorem that the notion of similarity relies on selecting a limited set of features. With increasing numbers of features, the similarity of all pairs of objects is asymptotically a constant! In most multivariate analyses of vegetation data all species are included, yet this presupposes that they are coherent (univocal) in the message they are sending. If they are incoherent, then some structure is likely to be hidden, even if the entire analysis is not vitiated by ambiguity. With these bog data, we have two major components present in the vegetation, comprising the vascular and the non-vascular species. With other vegetation types we might argue that the loss of information from non-vascular species is tolerable because such species largely reflect very local conditions, but the non-vascular component is clearly of considerable significance in bogs. There is, then, some interest in discovering if the two components agree in their structuring. The MML principle provides a convenient method of assessing the total variability using the one-class lengths. It also allows us to evaluate the structure present in the two components. We can further examine how far the two results are related by examining the assignment of things to clusters.

Second, we examine the effects of logical correlation between presence and abundance. Briefly we cannot measure how much a species is absent; some noughts are 'noughtier'/naughtier than others! Williams and Dale (1962) suggested that the data should be partitioned into presence/absence and abundance-when-present, with abundance-when-absent regarded as a missing value. There are three possible analyses using such partitioned data: including both the components, using the presence/absence component alone and using the numeric conditional on presence component alone. The last will usually contain many missing values.

Throughout it must be remembered that the datasets used here are very small. This means that the estimate of cluster number may well be in error unless the clusters are sharply separated. With more data we can support more clusters if they are present. We have an analysis of rainforest data using nearest neighbour sampling to define the sample plots (Williams et al. 1969), which estimated close to 100 clusters to be present! We also have a dataset with approximately the same number of things (1000+) where the number of clusters is estimated at 11. More data may give the opportunity for more clusters to be found but opportunity is not necessity.

Results

Overall the estimation of number of clusters seems to work quite well. Any tendency to overestimate the number of clusters is not shown by the present data probably because of the small sample sizes. However, message lengths associated with other cluster numbers were generally markedly larger and thus significantly different. In other analyses the overall minimum message length has occasionally been difficult to find, but with these data convergence was rapid and consistent from various starting configurations. The Slovak data always have the things uniquely assigned to a cluster, and generally the Latvian data also show little fuzziness, except in one analysis noted later.

Slovak data

From Table 1, it is clear that the Poisson model is a more succinct model of the data even with a single class. It also captures slightly more of the cluster structure though the difference is small. (20% compared with 18% for the Gaussian solution). While the difference may look numerically small, being approximately 211 nits it represents an odds-ratio of some ca. 10^{95} in favour of the Poisson solution! The Poisson alternative may not be truly optimal. Robinson (1954) suggested that cover might follow a β-distribution, but it is clearly more effective than the Gaussian in capturing the data structure at least for these data.

2 This is one of a number of largely nonparametric procedures based on the notion of data depth. See Liu et al. (1999) for more details on such methods.

Table 1. Slovak data Gaussian v. Poisson assumptions analysis: general comparison of class message lengths for clusters . 1-cluster length supplies the null hypothesis cost. Difference is the difference between 1-cluster and *n*-cluster costs which represents the reduction in redundancy.

Data	1-cluster Length	Number of clusters	n-cluster Length	Difference	Difference % 1-cluster
Poisson	1524.3	4	1213.3	311.0	20.4
Gaussian	1756.2	3	1424.7	331.5	18.9
High-Precision Gaussian	13488.8	4	9913.8	3575.0	26.5
Chord	10611.4	4	7951.0	2660.4	25.1
χ^2	7845.3	4	5786.0	2059.3	26.3
Hellinger	19776.3	4	8646.6	11129.7	56.28

Table 2. Slovak data: comparison of assignments of samples to clusters.

Sample	Gaussian	Poisson
1	4	5
2	4	5
3	4	5
4	4	5
5	4	5
6	4	5
7	4	6
8	4	5
9	5	6
10	5	3
11	5	3
12	5	6
13	5	3
14	5	4
15	3	4
16	3	4
17	3	3
18	3	4
18	3	4
29	3	4
21	3	4
22	3	3

If things and attributes are truly ordered along a gradient we might expect clear blocks of attributes associated with each group and consecutive blocks of things. The assignment of the things (Table 2) shows just such a picture for the Gaussian result with three disjoint groups in the order 4, 5, 3. In contrast, the Poisson result shows intermingling with no group in a single block. Since the existence of the gradient is doubtful, the Poisson result is perhaps more realistic. However, it may simply be a better model for rare species of which there are many in these data.

The levels of significance of the attributes are not very high, a function of the small number of things analysed. Of the higher significance levels, most are concerned with species absence or at least having a mean lower than that of the population; there are about twice as many negative indicators and the higher significance levels are largely associated with them. In the Gaussian solution (Table 3), a clear pattern of low level positives separates groups 3 and 4, though group 5 shows some overlap with 3. The Poisson solution (Table 4) is a little more complex and there is some intermingling, but not overmuch. From the attribute patterns, groups 3 and 4 seem disjoint from 5 and 6 suggesting that a two-group solution might have some merit if a hierarchical solution were used.

For patch-size determination, the Poisson rate parameter for individual species can be substituted for cluster labels (Table 5). The Slovak data have no particular spatial arrangement, and adjacency here reflects the supposed gradient; for illustrative purposes we assume equal spacing. Figure 1 shows the result for *Helianthemum nummularium,* a species not strongly associated with the cluster structure as can be seen from the entry in Table 5. The general trend towards an increasing rate towards the end of the presumed gradient is apparent, but this is not a monotone trend.

Figure 2 shows gradient with the rates for all species superposed. There is a considerable variation in pattern between species considering the small number of groups and short length, but overall a preponderance of relatively sharp changes. While a pattern distinguishing the start, middle and end of the sequence is commonly present, none of these segments is without interruption as might be expected if a continuous gradient was present. Local heterogeneity is displayed at scales from a single sample up to 6 samples. The all-species figure emphasises the points of change. With two exceptions the maximum rate for each species is restricted to a single cluster, which is what would be expected if the area was composed of relatively homogeneous patches. The two exceptional species, *Draba lasiocarpa* and *Genista pilosa* define larger patches, with the former reduced at the end of the sequence, and the latter at the start.

The Poisson solution captures more structure and indicates deviations from a simple gradient, which suggests that it is in fact the preferable choice as the MML principle indicates. If the assumed gradient is accepted then the Gaussian solution might be preferable. Certainly it provides a clean separation into 3 groups.

Table 3. Slovak data: Gaussian attribute significance. Species not significantly different in any group not shown. Under cluster columns are listed the probability of difference from the population, for 20%, 10%, 5%, 1%, 0.1% and 0.01% probability levels. The +/- column identifies if the mean is more or less than the population value.

Species	Clus3	3+/-	Clus4	4+/-	Clus5	5+/-
1 Jovibarba hirta ssp. glabrescens	.01	-	20	+		
2 Sedum album	.01	-	10	+		
3 Draba lasiocarpa	10	-				
4 Allium montanum	.01	-	20	+		
5 Hieracium bauhinii	5	-	20	+		
7 Dianthus lumnitzeri	5	-	20	+		
9 Globularia aphyllanthes	.1	-	20	+	1	-
10 Linum tenuifolium	1	-	20	+	5	-
11 Carex caryphyllea	.1	-	10	+	1	-
12 Scabiosa canescens	1	-	20	+	10	-
13 Leontodon incanus	.1	-				
14 Potentilla cinerea	10	-	20	+	20	-
15 Teucrium montanum	10	-	20	+	20	-
18 Juniperus communis	20	-				
20 Festuca pallens			20	+		
22 Plantago media			.01	-		
23 Trinia glauca			.1	-		
24 Helianthemum nummularium			1	-	20	+
25 Hieracium macranthum			1	-		
27 Polygala amara ssp. brachyptera	20	+	10	-		
28 Rosa spp.	20	+	20	-	20	-
29 Genista pilosa	20	+	1	-		
30 Linum catharticum	20	+	1	-		
31 Anthyllis vulneraria ssp. polyphylla	20	+	10	-	20	-
32 Taraxacum Sect. Erythrosperma	20	+				
33 Taraxacum Sect. Ruderalia	20	+	10	-	20	-
34 Plantago lanceolata	20	+				
35 Brachypodium pinnatum	20	+	10	-		
36 Centaurea stoebe	5	+	.01	-		
37 Biscutella laevigata ssp. austriaca	5	+	.01	-	20	-
38 Teucrium chamaedrys	20	+	.01			
40 Galium austriacum	20	+	10	-	.1	-
41 Sesleria albicans	20	+	.1	-		
43 Hippocrepis comosa	20	+			20	-
44 Seseli annuum	10	+	20	-	5	+
45 Koeleria macrantha	20	+	.01	-	5	+
46 Campanula moravica	20	+	.01	-		

Table 4. Slovak Data: Poisson Attribute Significance. Entries as in Table 4

Species	Clus3	+/-	Clus 4	+/-	Clus 5	+/-	Clus 6	+/-
1 Jovibarba hirta ssp. glabrescens			.1	-	20	+		
2 Sedum album			.1	-	20	+		
3 Draba lasiocarpa			20	-				
4 Allium montanum			.01	-	1	+		
5 Hieracium bauhinii			.1	-	20	+		
7 Dianthus lumnitzeri			20	-	20	+		
9 Globularia aphyllanthes	10	+	5	-	1	+	20	-
10 Linum tenuifolium	20	-	10	-	5	+	20	-
11 Carex caryphyllea	10	-	5	-	5	+		
12 Scabiosa canescens	20	-	20	-	5	+	20	-
13 Leontodon incanus	5	-	1	-	20	+	20	+
14 Potentilla cinerea	20	-	20	-	20	+	20	-
15 Teucrium montanum	20	-	20	-	20	+	20	-
17 Euphrasia stricta					20	-	20	+
18 Juniperus communis			20	-			20	+
19 Seseli elatum			20	-	20	-	20	-
20 Festuca pallens	20	-	20	-	20	+	20	-
21 Sanguisorba minor	20	-	20	-	20	-		
22 Trinia glauca	20	+			20	-		
23 Trinia glauca			20	+	5	-	20	-
24 Helianthemum nummularium			20	+	10	-		
25 Hieracium macranthum	20	-	20	+	20	-		
26 Carex humilis	20	-			20	-	20	-
27 Polygala amara ssp. brachyptera	20	-	20	+	20	-	20	-
28 Rosa spp.	20	-	20	+	20	-	20	-
29 Genista pilosa					10	-		
30 Linum catharticum			20	+	5	-	20	-
31 Anthyllis vulneraria ssp. polyphylla	20	+	20	-	20	-	20	-
32 Taraxacum Sect. Erythrosperma	20	+	20	-	20	-	20	-
33 Taraxacum Sect. Ruderalia	20	+	20	-	20	-	20	-
34 Plantago lanceolata	20	+	20	-	20	-	20	-
35 Brachypodium pinnatum			5	+	5	-	20	-
36 Centaurea stoebe			1	+	.01	-	5	-
37 Biscutella laevigata ssp. austriaca			.01	+	.01	-	5	-
38 Teucrium chamaedrys			1	+	.1	-	5	-
40 Galium austriacum	20	+	20	+	10	-	10	-
41 Sesleria albicans			5	+	5	-	20	-
43 Hippocrepis comosa					20	-		
44 Seseli annuum			20	+	.01	-		
45 Koeleria macrantha	. 10	+	20	+	.01	-		

Table 5. Slovak data: Poisson rates. Maximum rate in a cluster in **bold.**

Species	Cluster 3	Cluster 4	Cluster 5	Cluster 6
1 Jovibarba hirta ssp. glabrescens	0.1743	0.0073	**0.3128**	0.1743
2 Sedum album	0.1541	0.0072	**0.2662**	0.1541
3 Draba lasiocarpa	**0.2299**	0.1262	**0.2299**	**0.2299**
4 Allium montanum	0.2652	0.0075	**0.5167**	0.2652
5 Hieracium bauhinii	0.3106	0.0679	**0.5059**	0.3106
7 Dianthus lumnitzeri	0.3308	0.1590	**0.4923**	0.3308
9 Globularia aphyllanthes	0.0087	0.0066	**0.2178**	0.0126
10 Linum tenuifolium	0.0080	0.0062	**0.1559**	0.0113
11 Carex caryphyllea	0.0089	0.0067	**0.2221**	0.0885
12 Scabiosa canescens	0.0076	0.0060	**0.1254**	0.0105
13 Leontodon incanus	0.0092	0.0069	0.1996	**0.2046**
14 Potentilla cinerea	0.0067	0.0054	**0.0805**	0.0087
15 Teucrium montanum	0.0062	0.0051	**0.0660**	0.0080
17 Euphrasia stricta	0.1289	0.1289	0.0636	**0.2733**
18 Juniperus communis	0.0281	0.0051	0.0281	**0.0719**
19 Seseli elatum	**0.0038**	0.0015	0.0015	0.0017
20 Festuca pallens	0.0031	0.0028	**0.0138**	0.0034
21 Sanguisorba minor	0.0016	0.0015	0.0015	**0.0038**
22 Plantago media	**0.2192**	0.1339	0.0071	0.1339
23 Trinia glauca	0.0734	**0.1501**	0.0065	0.0123
24 Helianthemum nummularium	0.0583	**0.0936**	0.0062	0.0583
25 Hieracium macranthum	0.0070	**0.0729**	0.0056	0.0382
26 Carex humilis	0.0016	**0.0038**	0.0015	0.0017
27 Polygala amara ssp. brachyptera	0.0067	**0.0805**	0.0054	0.0087
28 Rosa spp.	0.0042	**0.0253**	0.0036	0.0049
29 Genista pilosa	**0.0583**	0.0583	0.0062	**0.0583**
30 Linum catharticum	0.0835	**0.1667**	0.0067	0.0128
31 Anthyllis vulneraria ssp. pol	**0.0624**	0.0047	0.0047	0.0071
32 Taraxacum Sect. Erythrosperma	**0.0153**	0.0028	0.0028	0.0034
33 Taraxacum Sect. Ruderalia	**0.0624**	0.0047	0.0047	0.0071
34 Plantago lanceolata	**0.0153**	0.0028	0.0028	0.0034
35 Brachypodium pinnatum	0.1743	**0.3564**	0.0364	0.0763
36 Centaurea stoebe	0.2046	**0.4198**	0.0074	0.0157
37 Biscutella laevigata ssp. austriaca	0.1794	**0.4593**	0.0073	0.0153
38 Teucrium chamaedrys	0.1592	**0.3680**	0.0072	0.0150
40 Galium austriacum	**0.2406**	0.2217	0.0357	0.0147
41 Sesleria albicans	0.0835	**0.1934**	0.0067	0.0128
43 Hippocrepis comosa	0.4268	**0.4268**	0.2678	0.4268
44 Seseli annuum	0.6591	**0.9222**	0.1937	0.6591
45 Koeleria macrantha	**0.7067**	0.5734	0.0076	0.4167
46 Campanula moravica	0.3762	**0.4541**	0.0074	0.0162

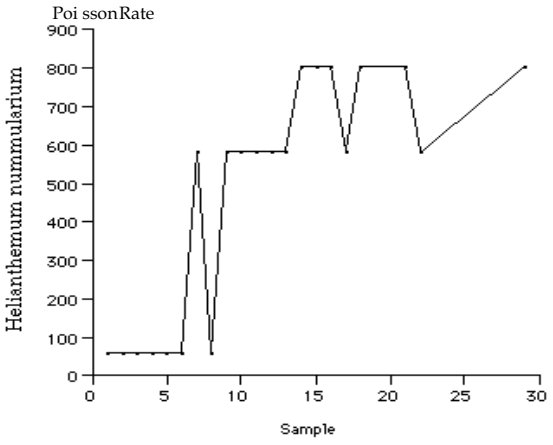

Figure 1. Slovak data: Poisson rates for *Helianthemum nummularium.* Rates are in order of the presumed gradient show a general upward trend with finer scale interruptions.

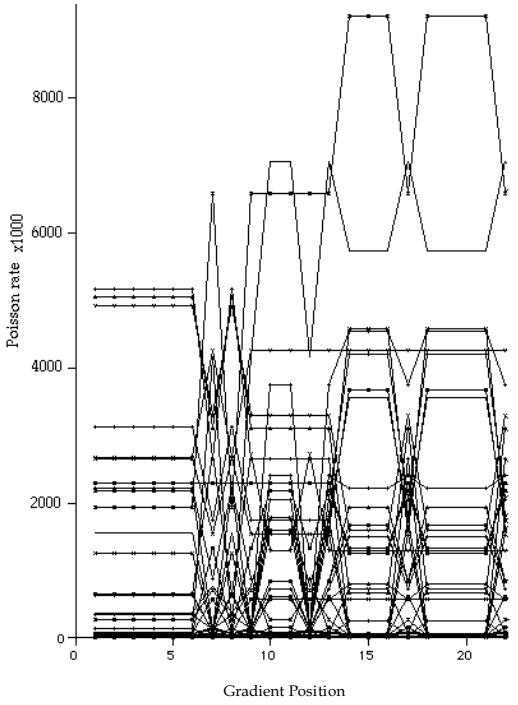

Figure 2. Slovak data: Poisson rates all species in order of the presumed gradient.

Table 6. Entries are probability and R values for predicting row clusters from column clusters. n.s. = nonsignificant.

Y \ X	Euclidean proper precision	Euclidean high precision	Chord Distance	χ^2 Distance	Hellinger Distance
Euclidean proper precision	-	0.001 0.55	0.001 0.55	0.004 0.47	0.18 0.25
Euclidean high precision	0.001 0.55	-	0.00… 1.0	0.004 0.47	0.03 0.37
Chord Distance	0.001 0.44	0.00… 1.0	-	0.004 0.47	0.03 0.37
χ^2 Distance	0.04 0.39	0.004 0.46	0.004 0.46	-	0.07 0.33
Hellinger Distance	0.18 0.20	0.03 0.38	0.03 0.38	0.07 0.32	-

Table 7. Latvian data: general comparison of class message lengths coherency analysis. Entries as in Table 2.

Data	1-cluster solution	Number of Clusters	n-cluster solution	Difference	Difference % 1-cluster
Full Data	2563.1	6	2100.6	462.5	18.04
Vascular Species only	976.4	4	794.4	180.0	18.44
Nonvascular species only	1586.7	4	1348.8	237.9	14.99

The general results are presented in Table 6. What is remarkable is that the Hellinger distance provides a much greater structuring than ANY other solution. The high precision Gaussian and chord solutions resulted in identical clusters being formed. By analysing the contingency tables between clusters the metrics can be organised by their relationship with the Gaussian analysis. The order of similarity was Gaussian → High Precision/ Chord → χ^2 → Hellinger. The Hellinger result is almost independent of Gaussian result but is correlated with the χ^2 result.

This would seem to suggest that the Gaussian/Euclidean solution is not particularly good at capturing structure within the vegetation space. The Hellinger distance outperforms all the others, including the χ^2 metric by a factor of 2. However, some caution is required. Inspecting the significant attributes it becomes clear that the rarest species are contributing most to the Hellinger cluster distinction - in a phytosociological idiom we are finding 'faithful' species rather than 'constant' species. The sensitivity of the χ^2 metric to rarities is known from correspondence analysis where rare species are usually *severely* downweighted to avoid them dominating the results. It seems the Hellinger distance is even more sensitive. Such a propensity is also present in Goodall's probabilistic dissimilarity coefficient (Goodall and Feoli 1988). We may well wish that undue weight not be assigned to the rare species and opt for another solution. In such a case we might also reject the χ^2 metric.

Table 6 also shows the effects of precision. Because Snob is sensitive to the precision assigned to the numeric attributes, the two Gaussian solutions are not identical; indeed the high precision analysis seems more effective. This is illusory since the actual precision is certainly not the 10^{-5} value ascribed in the high precision analysis. This will also have inflated the results for the other analyses and aggravated the effects of rarity. It therefore appears that the Poisson solution still offers an acceptable clustering.

Latvian data

Table 7 shows the general characteristics of the three analyses. The vascular data are clearly less variable and more strongly structured than the nonvascular. Both find 4 of the 6 groups found by the full analysis. Table 8 shows the actual assignments. Although a formal test of significance is vitiated, likelihood ratio tests show all the analyses are 'highly significantly' related. Through a correspondence analysis of the intergroup contingency tables (Figure 3ab), the approximate mappings between groups could be recognised (Table 9). These suggest that the vascular and nonvascular results both distinguish some of the clusters found in the full analysis, but confuse others. The nonvascular group 5 has no obvious cognates in the full

Table 8. Latvian data coherency analysis: assignments of things to clusters.

Sample Number	Full Data	Vascular Species only	Non-vascular Species only
1	4	8	4
2	4	9	4
3	4	9	4
4	6	8	4
5	7	9	4
6	6	8	4
7	3	8	4
8	4	8	4
9	3	11	4
10	4	9	4
11	10	3	9
12	10	3	4
13	10	3	9
14	3	3	3
15	10	3	9
16	4	3	9
17	10	3	9
18	10	3	9
19	10	3	9
20	10	3	9
21	6	3	9
22	6	3	5
23	6	11	4
24	10	8	5
25	6	8	4
26	6	8	4
27	6	8	4
28	6	8	9
29	7	9	3
30	7	3	3
31	7	9	4
32	3	9	4
33	7	9	3
34	7	3	3
35	7	9	3
36	7	3	3
37	7	9	4
38	7	3	3
39	11	11	9
40	11	11	9
41	11	11	5
42	11	11	9
43	11	11	9
44	7	11	4
45	3	9	4
46	11	11	4
47	3	11	4
48	3	11	5

Table 9. Group equivalence for coherency analysis groups.

Full Data	Vascular species only	Non-vascular species only
3	?	4
4	9	4
6	8	4
7	9	3
10	11	9
11	11	9
?	?	5

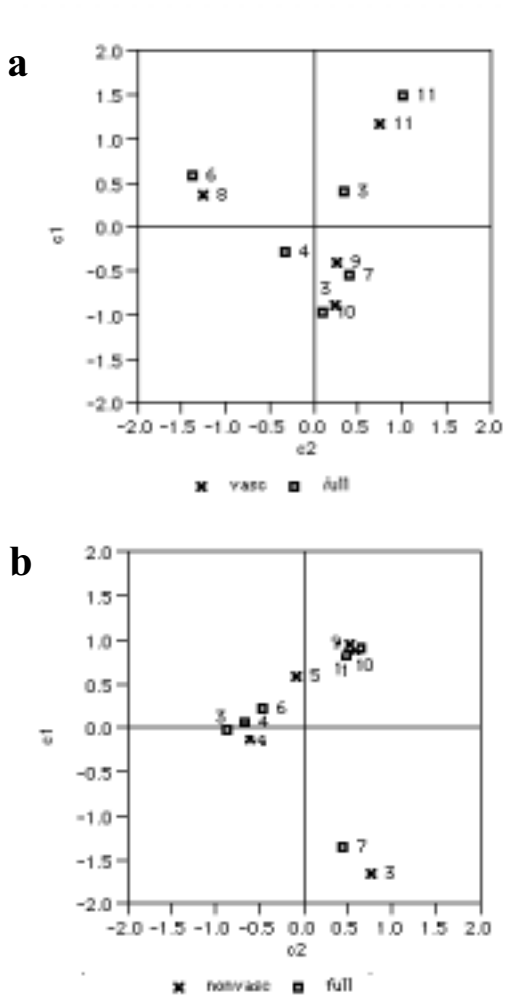

Figure 3. Latvian data correspondence analysis of contingency tables of group assignments, **a**: full x vascular with eigenvalues for c1=0.78 and c2=0.75 respectively. **b**: full x nonvascular with eigenvalues for c1=0.81 and c2=0.60.

analysis, while nonvascular group 9 represents full groups 10 and 11, and nonvascular group 4 merges full groups 3, 4 and 6. The vascular analysis shows 4 groups cognate with full groups, but 2 full groups, 3 and 4, have no representation in the vascular groups.

Table 10 presents the species significantly related at the 1% level or greater. As in the previous analysis, most of the significant species relationships are negative. In two analyses, a group is defined entirely by negative relationships. The same species tend to recur in each analysis, with only 2 species restricted to a single analysis. Both the reduced-species analyses find clusters that are also significant for species of the other kind and overall appear to have slightly more positive indicators.

Table 11 presents the general information. Both the full partitioned data and the presence/absence analysis in-

Table 10. Latvian data: coherency analysis attributes significant at the 1% Level. Nonvascular species in [].

Species	F3	F4	F6	F7	F10	F11	V7	V8	V9	V11	N3	N4	N5	N9
Calluna vulgaris				-							-			
Eriophorum vaginatum				-				+	-					+
Andromeda polifolia								-						
Rhyncospora alba			-			-		-		-	+		-	
Oxycoccus palustris	+													
Drosera rotundifolia			-	+			+	-	+		+			
Pinus sylvestris										+				
[Cladina sylvatica]		+		-	-		-				-			
[Spagnum tenellum]			-	+	-	-	+				+			-
[Milia anomala]	+													
[Cephalozia connivens]			-											
[Sphagnum magellanicum]		-							-					
[Dicranum affine]	-		-	-	-	+	-			+	-	-	+	-
Tricophorum cespitosum			-		-		-	-						-
Drosera anglica			-				+	-		-				
Rubus chamaemorus		-	-	-		+	-	-		+				
Empetrum nigrum			-	-				-	-					
[Sphagnum flexuosum]	-		-	-	+		+	-	-	-	-	-		+
[Cladopodiella fluitans]	-									-	+	-		
[Sphagnum cuspidatum]	-		-	-	+		+		-	-	-	-		+
[Sphagnum fuscum]	+			-	-						-			
[Calypogeia sphagnicola]					+								+	
[Sphagnum angustifolium]						+				+		-	+	
[Cladina rangiferina]							-	-		+		-	+	
[Aulocomnium palustre]	-					+				+		-		
[Pleurozium schreberii]													+	

Table 11. Latvian data partition of information: general comparison of class message lengths

Data	1-cluster	Number of clusters	n-cluster	Difference	Difference % 1-cluster
Full Partitioned Data	1334.6	1	1334.6	0	0
Presence/Absence	686.1	1	686.1	0	0
Numeric excluding presence	728.1	2	646.9	81.2	11.15
'Effective Entropy'			1.6		

dicate a single cluster only, while presence/absence contributes less than half the full information. The purely numeric data do manage to identify 2 clusters though the overall gain is relatively small. The assignments to groups for the numeric data are shown in Table 12, where it is obvious that there is considerable overlap. More than half the things are assigned partially, and 10 things are decidedly ambiguous with relative probabilities of 65 to 35 or worse! Clearly, treating all absences as missing values muddies the result considerably. The 1-class cost for the full data is 1756.2 while for the numeric with missing values the cost is 728.1, a difference of over 1000 nits! Even with the addition of the presence/absence information, we still have a loss of over 400 nits (full = 1756.2, presence/absence + numeric = 1333.0). Again, we have a reminder that much of the pattern in vegetation data is related to patterns of absence, an example of Babad and Hoffer's (1984) argument that even no data has value!

The presence/absence and numeric data should be orthogonal and we can test this by comparing the sum of the message lengths for the two distinct analyses with that for the combined data for the *n*-class solution. The value is very small, and lies within the errors of estimation of the program. This is reinforced by the attribute significance data where only 3 species show differences, *Sphagnum tenellum*, *Rubus chamaemorus* and *Sphagnum flexuosum*. All three are significantly negative in group 13, while the two *Sphagnum* species are weakly ($p < 20\%$) positive for group 16. One other species, *Empetrum nigrum* shows weak ($p < 20\%$) negative presence/absence differences for group 16. Three species show very weak association in terms of presence information, although not, of course, otherwise active.

Table 12. Latvian data: assignments for numeric data (excluding presence/absence information). Length = message length to encode the data for that thing assuming is belongs to its primary cluster. Rel. Prob. is the realtive probability of the thing belonging to a cluster.

Sample	Length	Primary Cluster	Relative Prob. x100	Secondary Cluster	Relative Prob. X100
1	9.9	1	60	2	40
2	4.4	1	65	2	35
3	8.0	1	93	2	7
4	2.9	1	65	2	35
5	6.5	1	65	2	35
6	9.2	1	65	2	35
7	11.3	1	88	2	12
8	10.7	1	88	2	12
9	9.0	1	93	2	7
10	15.4	1	91	2	9
11	15.7	2	99		
12	9.7	1	99		
13	14.4	1	75	2	25
14	10.3	1	65	2	35
15	13.5	1	95	2	5
16	12.3	2	99		
17	9.8	2	99		
18	14.1	2	99		
19	15.4	2	99		
20	17.1	2	99		
21	9.4	1	65	2	35
22	10.5	1	98	2	2
23	8.2	1	65	2	35
24	7.5	1	98	2	2
25	10.8	1	98	2	2
26	7.9	1	98	2	2
27	5.9	1	65	2	35
28	6.6	1	65	2	35
29	14.2	2	99		
30	13.3	1	88	2	12
31	15.0	1	90	2	10
32	13.5	1	99		
33	16.1	1	99		
34	12.3	1	99		
35	14.6	1	99		
36	10.2	1	99		
37	13.8	1	99		
38	15.6	1	99		
39	10.4	1	92	2	8
40	13.5	1	95	2	5
41	18.3	2	99		
42	15.0	2	99		
43	**29.2**	2	99		
44	13.5	1	99		
45	13.3	1	98	2	2
46	17.4	2	99		
47	16.3	2	99		
48	14.1	1	98	2	2

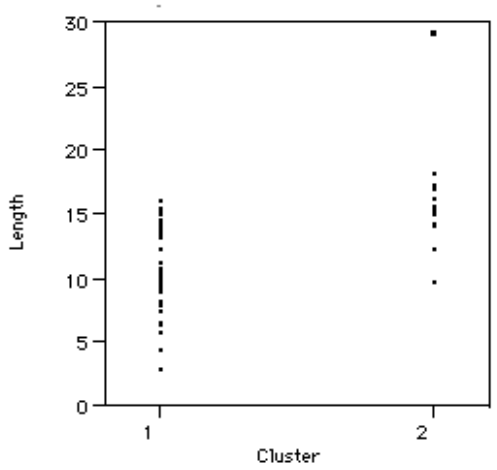

Figure 4. Latvian data: numeric data independent of presence/absence effects. Distribution of 'thing' message length estimates for 2 clusters. Note the outlier (stand 43).

Figure 4 shows the lengths required for coding the things using the 2 numeric clusters. The interesting point here is the obvious outlier (in fact stand 43) to group 2. This is an example of how the message length of a thing can indicate a probable outlier.

Discussion

Substantive

The limited results presented in these examples should be regarded as suggestive of certain properties of vegetation data and their structure. They also indicate some of the possibilities of using the MML principle.

It seems that, for the Slovak data, a Poisson solution is preferable to a Gaussian one, although it may not be optimal since percent cover might have some other distribution than Poisson. The generality of this conclusion re-

mains doubtful, and further examination of other datasets is desirable.

It also seems that models using metrics that emphasise rare species identify more structure than those that do not. Whether this is useful depends somewhat on the analyst's prior beliefs about the nature of vegetation; conservationists might argue for the importance of rarity, other vegetation managers might be less enthused. Dale (1994) suggested that a Riemannian space might be more appropriate but we have not yet examined such a possibility although an approximation might be obtained by using step-across methods (Bradfield and Kenkel 1987).

The possibility of using the estimates of the Poisson parameters to provide a variable Poisson model (Stevens 1937) for the species is interesting and will be examined further elsewhere (Dale 2001). The result, treated at face value, does suggest that spatially vegetation can be regarded as a series of relatively small patches. In the present case, these do NOT have a monotone relationship with the presumed gradient, but by increasing the sample plot area, small variations could be smoothed away. If you have a prior belief that vegetation responds smoothly to environmental gradients, this may be an acceptable procedure. This will be discussed later when the possibility of comparing classes of models, for example mixture clustering and axial ordinations, is considered.

The Poisson patches need not be environmentally determined, instead forming as a consequence of the various processes of growth and regeneration (see e.g, Boerlijst and Hogeweg 1991, Dale and Hogeweg 1998, Dale 1999). Such autopoietic patterns can modify the selection pressures operating on the plants (Savill et al. 1997). We might seek to test this by examining the changing positions of patches through time, since environmentally determined patterns should be more static spatially.

The Latvian data analyses suggest that partition of data into qualitative and quantitative portions is not required, with the combination and the presence data both accepting a single cluster. It should be remembered that these data were selected to be representative and the absence of presence/absence information may be a consequence of the quality of this selection. The analysis using only known quantities gives a solution with considerable ambiguity and with few species associated with the disjunction. It also emphasises the importance of absence information in vegetation structure.

The species coherency analysis suggests that vascular and nonvascular species are partially replicating the full analysis but that each ignores some aspects, and the nonvascular even finds a novel disjunction. The species are not completely coherent in their representation of the vegetation structure. The results illustrate 2 different ways of diverging from the full data structure. The vascular data, more structured than the nonvascular, manage to identify 4 of the full groups relatively cleanly, but lose 2 altogether; this represents a partial recovery of structure. The nonvascular species seem to have a more idiosyncratic view, identifying at least one group not clearly found in the full analysis and merging several others; this represents a refocussed recovery, like astigmatic blurring. But despite their differences, the subset analyses do identify some shared structure with species of one kind showing differentiation within the clustering formed by the other.

It is possible that the size of the sample plots masks some variation, especially for the nonvascular species that do not adopt a phalanx strategy. The choice of the area of sample plots is always difficult whenever plants of markedly different sizes or life forms are present. Lux (2000, see also Lux and Bemmerlein-Lux 1998) has proposed using different descriptors to normalise differences in temporal scales of variation but it is more difficult to see how such a change would apply to spatial scale. Such questions of scale will be examined elsewhere.

Methodological

There remain several questions that will be briefly addressed here. These are:

- Intra-cluster correlation between attributes;
- Choice between hierarchical clusters and partition;
- Choice between crisp and fuzzy clusters;
- Dependencies between things.

Intra-cluster correlation between attributes

Pattern in ecological data is usually reflected in inter-attribute correlation and clustering is a way of modelling inter-attribute correlations. Mixtures of uncorrelated attributes can model any distribution at the cost of introducing extra clusters if the correlation does not arise due to actual homogeneous unknown subpopulations. In the real world, we might expect correlation structure of a continuous kind and this can be modelled by hypothesising the existence of a continuous attribute, not measured, which represents the common factor affecting the attributes that were measured. This can be generalised for more complex correlation patterns by introducing more continuous variables. MML methods that incorporate intra-cluster correlation have been partially developed by Wallace and Free-

man (1992), Wallace (1995, 1998) and Edwards and Dowe (1998) although these are segmentation methods.

Hierarchy or partition

Many clustering methods presently employed in ecology are hierarchical. Goodall and Feoli (1988) extend this a little by allowing some individual samples to remain unassigned within nested classes; that is, given a cluster A then it may contain other subclusters, B or C, but not all members of cluster A will be assigned to these. A hierarchical MML segmentation procedure has been described (Boulton and Wallace 1973, 1975). There seems to be an intimate relationship between attribute subsets and such things as nesting and overlap. Nesting would then be only one pattern that might be observed. More complex interrelationships between feature subsets than a simple nesting might be capable of resolution using multiple factor within cluster variation and allowing axes used to describe any within-cluster variation to be shared with axes in other clusters.

There may, for example, be several orthogonal structures present; for example, in kin relationships we can distinguish generation (grandparent - parent, parent-child) and sex (father-son, mother-daughter). Uncle-nephew, aunt-niece, and both can be present at the same time. Which one you find in a hierarchical search will depend on slight variations in the sample. However, a reticulate representation (a plexus) might represent both simultaneously. The additive clustering method of Arabie and Carroll (1980) is one means of finding and presenting such structures and Wallace's (1995) suggestion provides another.

Crisp segmentation or fuzzy clustering

Clusters may be defined such that any thing is assigned to a single cluster only (segmentation) or allowed to contribute to several clusters (fuzzy). Most clustering methods used in ecology do not permit an assessment of degree of belonging of things to clusters. There do exist fuzzy solutions of several different kinds (Dale 1988) including some where a "degree of belonging" to a group is provided (Bezdek 1974) and these have sometimes, though rarely, been applied to vegetation data (Yarranton et al. 1972). However, these suffer because the amount of fuzziness can be varied through the user choice of an exponent. Thus, the degree of fuzziness is not estimated from the data alone.

It is known that crisp methods can be inconsistent in their definition of clusters parameters if the true clusters in fact overlap (Figure 5). Assigning things to clusters in a fuzzy manner avoids the inconsistency, and Wallace and Freeman (1987) have shown that this can be used to further reduce the message length. Since difference in message length is related to odds ratios for the associated models, this means that fuzzy assignment cannot be less acceptable than crisp assignment. Ganesalingam. and

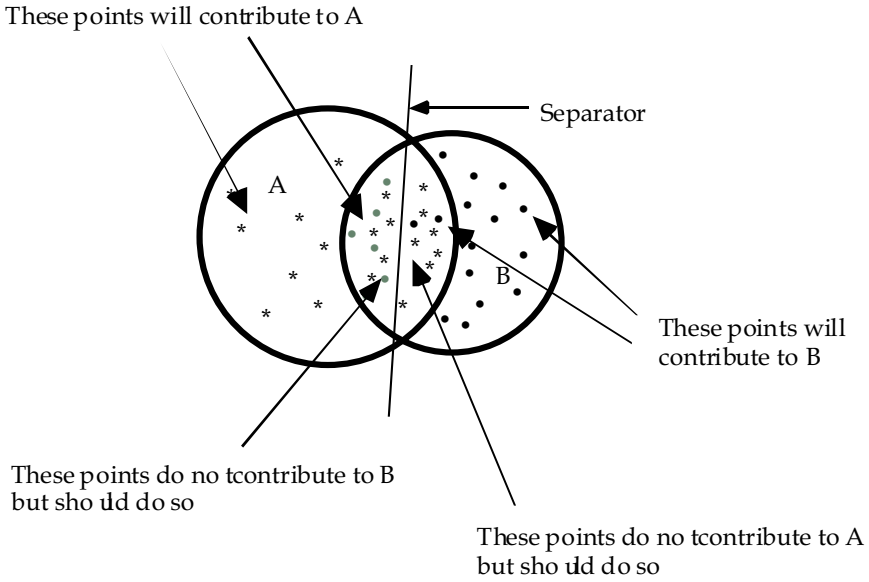

Figure 5. Segmentation and mixture separation: inconsistency resulting from overlap. The boundary marks the line of segmentation. Points lying on the 'wrong' side of the boundary are indicated. The parameters for clusters A and B will be incorrectly estimated if these points are excluded.

McLachlan (1980) show that mixture modelling is preferable to segmentation where cluster sizes are very disparate which is perhaps an additional advantage.

A further difficulty was raised by Chatfield (1995). He argues that traditionally the estimation of model parameters assumes that a model has a prespecified form, which disregards possible model uncertainty. This implicitly assumes the existences of some 'true' model that may be a fiction. Model uncertainty is a fact of life and likely to be more serious than other sources of uncertainty. However this question is beyond the scope of the present discussion.

There may be good reason for requiring a segmentation (Oliver and Forbes 1997), that is for demanding a sharp boundary between classes with no overlap. A typical application would be identifying elements of structure in a photographic image. MML can be used for segmentation (Viswanathan et al. 1999) and seems to be more effective than other procedures, including cross-validation. In any case, fuzzy clustering seems more realistic for vegetation, where ecotones and ecoclines are ubiquitous (van der Maarel 1990) and sample plots can easily include several vegetation types because of spatial or temporal overlap. MML also provides a relative probability of belonging to any cluster for each observation and the message length for encoding. This last can be used to identify things with very high costs, and therefore possible outliers.

Wallace (1998) shows that using fuzzy assignment can shorten the message length in an MML clustering. The trick is to use the uncertainty to obtain advance information on the next thing. In fact a thing is assigned probabilistically based on the *a posteriori* probabilities of the various clusters, which has the same effect as partial assignment. A random assignment is a form of Gibbs sampling of plausible classifications.

Dependency: temporal and spatial

An assumption common to almost all methods of clustering presently in use in ecology is that the things to be clustered are independent samples. At the same time it is common to use grid sampling patterns and transects where spatial dependency is a strong possibility, or to examine observations made sequentially in time where again dependency is very likely. MML procedures which incorporate models of various forms of dependency have been proposed by Edgoose and Allison (1999) and Wallace (1998) and the former comment specifically that the results obtained by ignoring dependence, as in Dale (2000b), and those where it is incorporated in the model can be expected to differ considerably, though this is not always the case. Li and Biswas (1999, 2000) have also proposed another clustering procedure based on Hidden Markov Models for temporal dependency.

Comparing classes of models

The final question concerns the possibility of comparison between different *classes* of model has to be considered. We need, for example, to be able to assess whether a model based on fuzzy clusters is preferable to one based on axes for a particular dataset, or whether a hierarchical clustering is to be preferred to some other scheme of cluster interrelationships. MML and more especially its relative the Minimum Description Length principle (Rissanen 1983, 1995) provide means of choosing between classes of models.

Falsification approaches do not seem to permit such a comparison and indeed Shipley and Keddy (1987) have argued that some questions, such as the choice between clustering and axial representations, are unanswerable because falsification cannot be applied. Unfortunately, this proposition has become part of the present interpretative community (see Carley and Palmquist 1992), and some questions have, in consequence, been excluded for consideration for being undecidable, unnecessary and/or unscientific.

Yet Shipley and Keddy's proposition is incorrect! It is possible to compare non-hierarchical and hierarchical cluster models or, more generally, specific cluster and axial (ordination) models. Rissanen's (1983, 1995) Minimum Description Length was developed for precisely that purpose! The minimal message must now contain information on the classes of models and the assignment of prior probabilities for particular classes might cause difficulties, but the principle remains intact. The difficulties that remain are ecological, such as choice of scale and choice of descriptors. Such problems will determine if making the comparison is useful but do not change the fact that it is possible. In comparing two classes of model, the user has to supply a prior probability for each so any preference for a continuum, say, can be captured in the priors.

There is no universal mechanism for capturing all structure in data, nor is there any reason to assume that a single form of structure is universally applicable to all vegetation data. Rather, we choose to look for structure that is useful to us. Given a specific dataset we can assess whether one structuring method is more effective than another, but the extension of this to a general principle is another matter.

Acknowledgments: Our thanks to Sanyi Bartha who provided some extremely useful and important comments on an earlier draft.

References

Akaike, H. 1978. A Bayesian analysis of the minimum AIC procedure. *Annals Inst. Statist. Mathematics* 30:9-14.

Arabie, P. and J. D.Carroll. 1980. MAPCLUS: a mathematical programming approach to fitting the ADCLUS model. *Psychometrika* 45: 211-235.

Austin, M. P. 1970. An applied ecological example of mixed data classification. In: R. S. Anderssen and M. R. Osborne (eds.), *Data Representation*. Univ. Queensland Press, Brisbane. pp. 113-117.

Babad, Y. M. and J. A. Hoffer. 1984. Even no data has value. *Commun. Assoc. Comput. Mach.* 27: 748-756.

Bezdek, J. C. 1974. Numerical taxonomy with fuzzy sets. *J. Math. Biol.* 1: 57-71.

Boerlijst, M. C. and P. Hogeweg. 1991. Spiral wave structure in prebiotic evolution: hypercycles stable against parasites. *Physica* D 48: 17-28.

Boik, R. J. 1987. The Fisher-Pitman permutation test: a non-robust alternative to the normal theory F-test when variances are heterogeneous. *Brit. J. Math. Statist. Psychol.* 40:26-42.

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

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

Boulton, D. M. and C. S. Wallace. 1975. An information measure for single-link classification. *Comput J.* 18: 236–238.

Bradfield, G. E. and N. C. Kenkel. 1987. Nonlinear ordination using flexible shortest path adjustment of ecological distance. *Ecology* 68: 750-753.

Carley, K. and M. Palmquist. 1992. Extracting, representing and analyzing mental models *Social Forces* 70: 601-636.

Chaitin, G. J. 1966. On the length of programs for computing finite sequences. *J. Assoc. Comput. Mach.* 13:547-549.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference *J. Royal Statistical Soc. Series A* 158: 419—466.

Dale, M. B. 1987. Knowing when to stop: cluster concept-concept cluster. *Coenoses* 3: 11-32.

Dale, M. B. 1988. Some fuzzy approaches to phytosociology: ideals and instances. *Folia Geobot. Phytotax.* 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. 2000a. On plexus representation of dissimilarities. *Community Ecol.* 1:43-56.

Dale, M. B. 2000b. Mt Glorious revisited: secondary succession in subtropical rainforest. *Community Ecol.* 1:181-193.

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

Dale, M. B. (submitted) Models, measures and messages: a role for induction.

Dale, M. B. and P. Hogeweg. 1998. The dynamics of diversity: a cellular automaton approach. *Coenoses* 13:3-15.

Edgoose, T. and L. Allison. 1999. MML Markov classification of sequential data. *Statistics and Computing* 9: 269-278.

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

Ganesalingam, S. and G. J. McLachlan. 1980. A comparison of the mixture and classification approaches to cluster analysis. *Commun. Statist. Theor. Meth. A*9: 923-933.

Goodall, D. W. and E. Feoli. 1988. Application of probabilistic methods in the analysis of phytosociological data. *Coenoses* 1: 1 -10.

Gordon, A. D. 1994. Identifying genuine clusters in a classification. *Comput. Statist. Data Analysis* 18: 561-581.

Hayes, A. F. 1996. Permutation test is not distribution free. *Psychol. Methods* 1: 184-198.

Hill, M. O., R. G. H. Bunce and M. W. Shaw. 1975. Indicator species analysis: a divisive polythetic method of classification and its application to a survey of native pinewoods in Scotland. *J. Ecol.* 63: 597-613.

Hoffman, R. L. and A. K. Jain. 1987. Sparse decomposition for exploratory pattern analysis. *I. E. E. E .Trans. Patt. Anal. Mach. Intell.* PAMI-9: 551-560.

Hubert, L. and P. Arabie. 1994. The analysis of proximity matrices through sums of matrices having (anti-)Robinson forms. *Brit. J. Math. Statist. Psychol.* 47:1-40.

Kolmogorov, A. N. 1965. Three approaches to the quantitative description of information. *Prob. Inform. Transmission* 1: 4-7 (translation).

Krishna-Iyer, P. V. 1949. The first and second moments of some probability distributions arising from points on a lattice and their application. *Biometrika* 36: 135-141.

Legendre, P. and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* 129: 271-280.

Li, C. and G. Biswas. 1999. Temporal pattern generation using hidden Markov model-based unsupervised classification. In: *Advances in Intelligent Data Analysis, Lecture Notes in Computer Science 1642*, Springer-Verlag, Berlin. pp. 245-256.

Li, C. and G. Biswas. 2000. Bayesian temporal data clustering using hidden Markov model representation. In: P. Langley (ed.), *Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA. pp. 543-550.

Liu, R. Y., J. M. Parelius and K. Singh. 1999. Multivariate analysis by data depth: descriptive statistics (with discussion). *Ann. Statist.* 27:783-885.

Lux, A. 2000. Die Dynamik der Kraut-Gras-Schicht in einem Mittel- und Niederwaldsystem. Untersuchungen im Gebiet des Kehrenbergs bei Bad Windsheim. *Dissertationes Botanicae* Vol. 333.

Lux, A. and F. A. Bemmerlein-Lux 1998. Two vegetation maps of the same island: floristic units versus structural units. *Appl. Veg. Sci.* 1: 201-210.

Oliver, J. J. and C. S. Forbes. 1997. Bayesian approaches to segmenting a simple time series. Tech. Rep. 97/336 Dept. Comput. Sci. Software Engineering, Monash University. Clayton, Victoria 3168, Australia..

Pillar, V. D. 1996. A randomization-based solution for vegetation classification and homogeneity testing. *Coenoses* 11: 29-36.

Richardson, S. and P. J. Green. 1997. On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc.* B 59: 731-792.

Rissanen, J. 1983. A universal prior for integers and estimation by minimum description length. *Annals of Statistics 11:416-431*.

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.

Robinson, P. A. 1954. The distribution of plant populations. *Ann. Bot.* 18: 35-45.

Sandland, R. L. and P. C. Young. 1979. Probabilistic tests and stopping rules associated with hierarchical classification techniques. *Aust. J. Ecol.* 4: 399-406.

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.

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

Stevens, W. L. 1937. Significance of grouping. *Ann. Eug. London*. 8:57-69.

Van der Maarel, E. 1990. Ecotones and ecoclines are different. *J. Veg. Sci.* 1:135-138.

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.

Wallace, C. S. 1990. Classification by minimum message length inference. In: G. Goos and J. Hartmanis (eds.), *Advances in Computing and Information – ICCI'90*, Springer-Verlag, Berlin. pp. 72-81.

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. 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. and P. R. Freeman. 1992. Single factor analysis by minimum message length estimation. *J. Roy. Statist. Soc. Ser. B* 54: 195-209.

Watanabe, S. 1969. *Knowing and Guessing*. Wiley, New York.

Williams, W. T. and M. B. Dale. 1962. Partitioned correlation matrices for heterogenous quantitative data. *Nature* 196: 502.

Williams, W. T., G.N. Lance, L.J. Webb, J.G. Tracey. and J.H. Connell. 1969. Studies in the numerical analysis of complex rainforest communities IV. A method for the elucidation of small scale pattern. *J. Ecol.* 57: 635-654.

Yarranton, G. A., W. J. Beasleigh, R. G. Morrison and M. I. Shafti. 1972. On the classification of phytosociological data into non-exclusive groups with a conjecture about determining the optimum number of groups in a classification. *Vegetatio* 24: 1-12.