available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/actoec

Original article

Segmentation and clustering as complementary sources of information

Michael B. Dale[a], *Lloyd Allison*[b], *Patricia E.R. Dale*[a,*]

[a]*Australian School of Environmental Studies, Griffith University, Nathan, Qld. 4111, Australia*
[b]*School of Computer Science and Software Engineering, Monash University, Vic. 3800, Australia*

ARTICLE INFO

Article history:
Received 21 July 2006
Accepted 20 September 2006
Published online 23 January 2007

Keywords:
Clustering
Segmentation
Scale

ABSTRACT

This paper examines the effects of using a segmentation method to identify change-points or edges in vegetation. It identifies coherence (spatial or temporal) in place of unconstrained clustering. The segmentation method involves change-point detection along a sequence of observations so that each cluster formed is composed of adjacent samples; this is a form of constrained clustering. The protocol identifies one or more models, one for each section identified, and the quality of each is assessed using a minimum message length criterion, which provides a rational basis for selecting an appropriate model. Although the segmentation is less efficient than clustering, it does provide other information because it incorporates textural similarity as well as homogeneity. In addition it can be useful in determining various scales of variation that may apply to the data, providing a general method of small-scale pattern analysis.

1. Introduction

Terms in italic type (at their first occurrence) are explained in the glossary.

Unsupervised clustering has been a useful technique for analysing vegetation data since its first use by Goodall (1953). The development of minimum message length (MML) clustering (Wallace and Boulton, 1968; Boulton and Wallace, 1970; Wallace and Dowe, 2000; Wallace, 2005) provided a method that estimated the number of clusters and their properties consistently using a *fuzzy* clustering procedure. Both *hierarchical* and *non-hierarchical* clustering can be obtained using the minimum message length (MML) procedure. This can accommodate several kinds of data (nominal, continuous and angular), sequences of such values and copes with missing values. By employing fuzzy clustering we can also obtain consistent estimators of cluster parameters whereas *crisp* clusters do not guarantee consistency.

In some circumstances, however, an unrestricted clustering is not demanded. Instead, a partition that retains spatial or temporal coherency is desired. Such circumstances include but are not limited to:

Mapping. The graphical presentation of spatial variation is difficult if the entities to be mapped represent complex, highly textured areas with fuzzy assignment to classes. Additionally, there will be further complications because of the existence of multi-scale patterns. The delimitation of

* *Corresponding author*. Tel.: +61 7 3735 7136; fax: +61 7 3735 6717.
E-mail addresses: m.dale@griffith.edu.au (M.B. Dale), lloyd.allison@infotech.monash.edu.au (L. Allison), p.dale@griffith.edu.au (P.E.R. Dale).
1146-609X/$ – see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.actao.2006.09.002

spatially coherent areas with crisp boundaries will usually be desirable.

Process and context changes. In temporal and spatial series some processes may be restricted to particular contiguous sections. Alternatively, the environmental context may itself be changing abruptly resulting in the series showing markedly different properties in different coherent sections as well as smoother gradations.

Scale variation. Vegetation processes are scale dependent and interactions between scales is a major component of pattern formation. Many methods for identifying interesting scales, for example those derived from the work of Greig-Smith (1961) assume that the component patches are of similar size. Since environmental gradients can be of varying steepness it is unlikely that patches of similar sizes actually occur, although some environmental processes such as cryoturbation could generate them. Mostly, such patches are directly generated by vegetation processes (see Boerlijst and Hogeweg, 1991). A method for identifying spatially or temporally coherent segments would be useful in identifying such fragments without any assumptions concerning their size.

1.1. *Pattern and process*

Historically, the methods used to identify patch structure in vegetation, such as those of Greig-Smith (1961), have assumed that there is a periodic pattern of equal-sized clusters. By employing minimum message length criteria we can avoid such assumptions. Segmentation can involve fragments of varying length. Periodicity can be identified by establishing common models for (non-adjacent) fragments if these are not too costly; that is they do not increase the message length by very much. In addition, any hierarchical structure may also be investigated, though we have not done so in this paper.

In order to determine coherent fragments we need to segment a series, or an array, of observations. In *segmentation* we seek to break a series into a series of segments within which common processes may be assumed to be operating. The (internal) ends of any segment are called *change-points* or *edges*. Various methods have been proposed for doing this (see Dale, 1994), but the development of model selection criteria such as MML has provided means of estimating the number of change-points whereas many previous methods assumed some fixed number (e.g. Emad-Eldin and BuHamra, 1996) or used human assessment. A large number of proposals have been made for edge detection in image processing and for detecting changes in time series. However, Baxter and Oliver (1996) have suggested MML methodologies that seem to provide a preferable basis (see also Oliver et al., 1998; Oliver and Forbes, 1997; Fitzgibbon et al., 2000; Hanlon and Forbes, 2002).

Segmentation is one form of *constrained clustering*; in a simple form this can mean no more than permitting only adjacent samples to join the same cluster. Methods that have been proposed for this task include those of Ferligoj and Batagelj (1983), Dale and Dale (1994) and more generally Critchlow (1985). As with all clustering, a key requirement is to estimate the number of clusters or segments. Here we use the minimal message length principle combining the description of a model for the data with the likelihood of the data assuming the model is correct. It is possible to incorporate prior knowledge concerning the model but none was available here.

Assessing the significance of spatial coherence was first discussed by Krishna-Iyer (1949). However, it must be recognised that the fragments formed by segmentation will differ from those sought in unconstrained clustering. Clusters are usually defined to be homogeneous collections of things, although the definition of homogeneity can be complex.

In contrast coherent segments may be expected to be mixtures since they can encompass various trends and (textural) mixtures of things, so long as these remain coherent. With segments, the critical property is that the same processes appear to operate throughout the segment. Homogeneity, per se, has no special significance compared with common trends and common texture.

In this paper we shall examine the results obtained by segmenting a spatial series of observations, termed 'things' and compare these with the results of an unsupervised and unconstrained clustering. For comparison we use the minimal message length principle to provide a common measure of quality.

1.2. *Alternative methods*

1.2.1. *Segmentation*

Dale (1994) has previously examined problems of determining boundaries, edges or change-points and, more generally, non-stationarities and non-linearities in phytosociological data. One observation was that such data are often extremely noisy. Smoothing is necessary and this can make edges difficult to detect and enhance apparent continuity. In addition, that study was largely concerned with detecting single change-points, and rather ignored the possibility that several might occur. All the methods mentioned below, except those of Dale (1984, 1986, 1988a,b), have been applied to the data used here.

Dale (1984, 1986, 1988a,b) investigated the spacing and intermingling of species boundaries on an environmental gradient. The method is restricted to binary data, requires a linear relationship between spatial gradient and environmental gradient and relies on determining the boundaries of the distribution. This last may be problematic, as shown by Timone et al. (1993). To be fair, this method was intended to examine whether changes for different species showed coincidence and not as a general change-point test. Viswanatthan et al. (1999) provide a more general MML-based method for the binary problem.

The MML procedure estimates the number of change-points directly, using either numeric angular or binary data or mixture of these. It provides consistent estimates of cluster or segment parameters and it may be used to compare models of different complexity. With the Gaussian model used here, it identifies changes in either mean or variance or both. It can be applied where serial correlation is expected and even to provide a test for both existence and extent of such correlation.

MML can be applied to two-dimensional data such as images. A recent survey of other methods for this application can be found in Skarbeck and Koschan (1994). Two-dimensional segmentation methods sometimes use models

such as Markov fields but also may rely on linking edges to form regions, which implies a context dependency. In any case, since our data are one-dimensional these possibilities are not considered further here.

For quantitative data, the detection of change-points is usually based on applying some test of statistical difference across a potential boundary. Mostly these rely on differences in mean (Srivastava and Worsley, 1986) or median (Pettitt, 1979, 1980, 1981), although tests are available which employ difference in variance, change in some regression function (Kim and Siegmund, 1989) or in (hidden) Markov models (Kehagias, 2002). This last provides a method that can cope with serial correlation between observations.

A few tests can distinguish more generally between distributions. Thus the Anderson–Darling test (Stephens, 1974), based on the Kolmogorov–Smirnov statistic, is used to test if a sample of data comes from a specific distribution. Currently, tables of critical values are available for the normal, lognormal, exponential, Weibull, extreme value type I and logistic distributions. Barnett and Eisen (1982) provide another test for distribution change for a single change-point.

Indeed, a major problem with these methods is that they are designed to identify only one change-point; an exception is that of Emad-Eldin and BuHamra (1996) which looks for two change-points. The other methods can be applied recursively, but this modifies significance levels since it involves multiple testing. It also introduces the question of when to stop looking for more. Ferligoj and Batagelj (1983) used constrained clustering leaving the user to decide where to stop. Lombard (1988) and Losch and Cruz (2002) both provide procedures for application to multiple change-points but rely on human examination to determine the final number. Yao (1988) uses Schwarz (1978) criterion, which is an alternative to MML but one generally found to be less powerful for model comparison.

1.2.2. *Clustering*

With unsupervised clustering, the possibility of alternatives is much more simply dealt with for segmentation. No clustering method, other than MML, provides estimates of the number of clusters, consistent estimates of the cluster parameters, the possibility of incorporating serial correlation, the use of numeric, binary and angular data and appropriate treatment of missing values.

2. Materials and methods

2.1. *Data and analyses*

The analysis was applied to spatial sequences of vegetation (Gitay and Agnew, 1989). The primary data were recorded on a transect of 113 contiguous samples, each 4×4 cm, from a dune slack in the Ynyslas National Nature Reserve, West Wales, UK. In each sample the combined above- and below-ground biomass for all perennial species was measured. In all, 12 species were recorded but three were very rare and have been ignored here. Thus all analyses used nine species. Some data on mineral nutrient concentrations was also collected but has not been used here.

An unsupervised clustering was made using the SNOB program (Boulton and Wallace, 1970). The segmentation algorithm was implemented using a library for machine-learning (Allison, 2003, 2005) which was written in the functional programming language Haskell (Peyton Jones, 2003).[1] We assumed Gaussian within-segment distributions, ignoring any correlation between attributes. These correlations had been checked and were small (Table 1). They are unlikely to change the message length by more than 10 bits and such a change is unlikely to change the results much. Other distributions than Gaussian could be used if desired, such as multi-state, angular and Poisson. Each analysis provides an estimate of the number of change-points and the mean and variance parameters of each segment so delimited.

2.2. *Calculating the coding cost*

If the data are numbered $d[0],d[1],\ldots,d[n-1]$, we consider a directed acyclic graph with edges $e[i,j]$, $-1 \le i < j \le n-1$, where edge $e[i,j]$ corresponds to a segment of data $d[i+1],\ldots,d[j]$. There are $n \times (n-1)/2$ edges in total. The cost of edge $e[i,j]$ is the cost of stating the model of the segment and the data in the segment. This cost involves coding the positive integer $(j-i)$, the parameters of the segment distribution, and the segment of data given those parameters (cf. Wallace, 2005). The segmentation problem amounts to finding a shortest path from vertex $d[-1]$ to $d[n-1]$. This can be solved by Dijkstra (1959) shortest-path algorithm. The required prior probabilities are based on uniform prior on μ and $\frac{1}{\sigma}$ prior on σ, within the given limits.

Various edge detection methods were applied, mostly to individual species. Only single edges were sought. Changes in both mean and variance were examined. In addition, by fitting a regression on location in series it was possible to use a more complex model.

3. Results

A preliminary question is to determine if segmentation or clustering is warranted or whether a single segment or class is sufficient. Details of the message lengths are given in Table 2. In general, the smaller the message length the better is the model. It is obvious from Table 2 that both segmentation and clustering make considerable improvements on the one-class model, and further that clustering is preferable to segmentation since it results in a further reduction in message length. Obviously, the constraints implied by the maintenance of coherence by the segmentation have an additional, and quite considerable, cost. The odds in favour of a shorter message length are given by $e^{-\mathrm{diff}}$ where diff is the difference in message lengths. For example, the message length for the one-cluster solution is 5946.1, for the five-cluster solution is

[1] See Allison (2003, 2005) for reasons for using a functional language, especially for creating prototype programs quickly. This results from the high level of the language and the easy re-use of existing code.

Table 1 – Interspecific correlations

Species	*Agrostis stolonifera*	*Amblystegium serpens*	*Carex arenaria*	*Carex flacca*	*Eleocharis uniglumis*	*Hydrocotyle vulgaris*	*Juncus articulatus*	*Pressia quadrata*	*Ranunculus bulbosis*
Agrostis stolonifera	1	−0.03	0.21	0.25	0.11	0.10	0.02	−0.08	0.01
Amblystegium serpens		1	−0.15	−0.13	0.33	0.0	0.34	0.02	−0.07
Carex arenaria			1	−0.01	−0.03	0.05	−0.04	0.07	0.36
Carex flacca				1	−0.27	0.05	−0.38	−0.22	0.36
Eleocharis uniglumis					1	0.08	0.11	−0.08	−0.07
Hydrocotyle vulgaris						1	−0.02	−0.06	−0.02
Juncus articulatus							1	0.05	0.06
Pressia quadrata								1	−0.09
Ranunculus bulbosis									1

4562.2 and the difference is 1383.9. Thus the odds of the one-cluster solution compared with the five-cluster solution are given by $e^{-1383.9}$:1, a very strong support for the five-cluster solution.

3.1. *Segmentation*

The segmentation analysis identified four change-points and thus five segments of contiguous samples. These segments are, in terms of position of the things in the sample series: samples 1–14, 15–38, 39–52, 53–77, 78–113. The means and variances for the nine species in these segments, and for the total population, are presented in Table 3.

All segments show changes in abundance of the common species. However *Amblystegium serpens* and *Eleocharis uniglumis* are restricted to the final segment while *Ranunculus bulbosus* appears in segments one and three but not two, four or five. Means for *Carex flacca* show an arched distribution, first rising and then falling and the means for *Juncus articulatus* have the reverse pattern, falling then rising (see Table 3). However, see Dale (2005) for difficulties in using mean values. No symmetry is assumed nor any 'bell-shape' for the arch response.

3.2. *Clustering*

The SNOB clustering program (Boulton and Wallace, 1970) also identified five components (labelled 5, 6, 7, 14 and 15), although these are not spatially contiguous (Table 4). Very little fuzziness was apparent in the assignment of things to clusters so this potential source of problems is unlikely to be of serious concern. The means and standard deviations for the nine species for these clusters are presented in Table 5.

Table 2 – Message length statistics for various cluster and segment analyses

Data analysis	Message length	Difference from the one cluster analysis
One cluster or segment	5946.1	
Segmentation (five segments)	5109.0	837.1
Clustering (five clusters)	4562.2	1383.9

Cluster 6 is dominated entirely by *Carex flacca* while cluster 5 has more *Carex flacca* than cluster 6 and also has *Ranunculus bulbosus*. Cluster 7 has the same mean for *Carex flacca* as cluster 6, together with *Preissia quadrata*. It is close to cluster 6, since in those few cases where the assignment is fuzzy for things assigned to cluster 6, cluster 7 is the alternative, although the probability is never very high ($p \leq 0.13$).

Clusters 14 and 15 have reduced *Carex flacca* but with *Amblystegium serpens*, *Preissia quadrata* and *Eleocharis uniglumis* in varying amounts; 14 has more *Preissia quadrata*, 15 has more *Amblystegium serpens* and *Eleocharis uniglumis*. Where there are fuzzy assignments for cluster 14 the alternative is always cluster 15, but again these are few in number and the alternative has a relatively low probability.

3.3. *Segment–cluster relationships*

Examining this cluster information (Table 6) in the context of the segments, the first four segments are clearly distinguished from segment 5, the latter having additional species and lacking others present in the four others. Segments 1–4 appear much more similar than any of them are to segment 5, so it is of interest to determine how well models for individual segments fit other segments. This is shown in Table 7, which shows the message length for the fit of each model of a segment to segments other than to itself. The 'best' fitting segment, i.e. the 'most alike segment is highlighted. Thus the model for segment 1 fits itself best, as would be expected, but it is also a reasonable fit for segment 3 and likewise for segments 2 and its fit to segment 4. Segment 5 does not fit at all well to any of the other segments. The diagonal elements in Table 7 represent models fitting the segment from which they were defined. Normalising these values to obtain the K-L distances between segments (Table 8), the remoteness of segment 5 as a model for the remainder of the sequence is made very clear. The relationships are, of course, asymmetric.

It is clear that the segment pairs (1,3) and (2,4) are similar. Even allowing for asymmetry, the distances between these pairs are the smallest in their respective columns; furthermore, the pair (2,4) is clearly closer than is (1,3). Models for segments 2 and 4 fit reasonably to segment 5 data, but the segment 5 model is a bad fit for all other segments; it has its best fit with segment 4. Note that the pair members (2,4) have similar distances from segment 5, as do the pair members (1,3). This suggests that the results might be further improved by

Table 3 – Mean biomass (g) and (variances) for population and segments

Species	Population	Segment 1: samples 1–14	Segment 2: samples 15–38	Segment 3: samples 39–52	Segment 4: sample 53–77	Segment 5: sample 78–113
Agrostis stolonifera	21.10 (11.98)	10.61 (5.77)	27.11 (13.36)	30.53 (14.17)	17.33 (10.45)	20.19 (7.97)
Amblystegium serpens	0.28 (0.86)	0 (0)	0 (0)	0 (0)	0 (0)	0.89 (1.37)
Carex arenaria	61.01 (34.44)	49.84 (24.35)	96.85 (36.80)	43.01 (23.08)	46.14 (21.83)	59.14 (30.77)
Carex flacca	72.13 (48.72)	45.28 (26.31)	77.00 (49.43)	147.81 (44.05)	85.69 (31.96)	39.18 (21.87)
Eleocharis uniglumis	1.84 (5.62)	0 (0)	0 (0)	0 (0)	0 (0)	5.93 (8.89)
Hydrocotyle vulgaris	13.27 (7.49)	15.74 (9.62)	10.89 (7.81)	14.47 (4.96)	14.01 (7.18)	12.87 (7.27)
Juncus articulatus	23.76 (22.78)	30.69 (28.76)	21.76 (20.40)	10.52 (13.11)	13.69 (18.22)	35.14 (22.27)
Preissia quadrata	0.62 (1.53)	0 (0)	0.20 (0.66)	0 (0)	0.22 (0.54)	1.70 (2.32)
Ranunculus bulbosus	0.99 (4.47)	2.22 (5.95)	0 (0)	5.74 (10.23)	0 (0)	0 (0)

fitting a common model to the members of each pair, which could reduce the message length still further, at the expense of losing spatial connectivity. An *a posteriori* search for common models to disjoint segments could give insight into repetitive structure in the data. In the present case, the first four segments seem to represent alternating patches within the same larger community, whereas the final segment is a more distinct patch and might represent an element of a different mosaic.

3.4. *Scale effects*

The lengths of the segments are: segment 1 with 14 samples, segment 2 with 24, segment 3 with 14, segment 4 with 25 and segment 5 with 36. Note that the first and last segments are open-ended so their true size is unknown. Segment 5 is thus at least as large as the sum of segments 3 and 4, and probably segments 1 and 2 as well. Segment 5 is distant from all the others, indicating pattern at a different scale. Whether this disjunction represents a higher level separation of patches or some more significant change in processes operating requires further study.

Note, too, the alternating pattern formed by segments 1–4. Segments 1 and 3 are similar and alternate with segments 2 and 4. This suggests that the pairs of segments 1 with 2, and 3 with 4, are components of mosaic patterns at a similar scale to that of segment 5. (Their similarity is captured in the K-L distances.)

3.5. *Edge detection methods*

Almost all methods identified a change-point close to position at sample 77 in individual species analyses and in a multivariate study (Table 9). The actual locations range from 69 to 81, with the multivariate result indicating position at sample 77. The exceptional case is for *Agrostis stolonifera* which suggests two other locations.

A second change was sometimes identified around position at sample 31, though much less consistently. Interestingly, this change is most commonly identified by the test for change in parameters for regression on position. The test for two change-points, in contrast, does not supply much support for this position. However, a change in regression parameters for regressions on location in series is the strongest support.

All these results are based on differences in means across the change-point. Change-points in variance were also found in four cases. This suggests that relying on means only is insufficient.

Table 4 – Locations of cluster members along series

Sample no.		
Start position	End position	Cluster
1	2	5
3	13	6
14	14	5
15	26	6
27	28	7
29	31	6
32	32	7
33	38	6
39	41	5
42	51	6
52	52	5
53	53	6
54	56	7
57	65	6
66	66	7
67	69	6
70	70	7
71	75	6
76	76	7
77	78	6
79	83	14
84	84	15
85	85	14
86	87	7
88	88	14
89	90	7
91	91	14
92	92	7
93	93	14
94	94	15
95	97	7
98	102	15
103	103	6
103	110	15
111	111	7
112	112	14
113	113	15

Table 5 – Species means (standard deviations) for unsupervised clusters

Species	Cluster				
	5	6	7	14	15
Agrostis stolonifera	No significant variation from population and not contributing to the clustering				
Amblystegium serpens	0 (0)	0 (0)	0 (0)	0.3 (0.9)	1.9 (1.7)
Carex arenaria	No significant variation from population and not contributing to the clustering				
Carex flacca	111.2 (75.2)	72.1 (48.7)	72.1 (48.7)	50.5 (16.2)	25.2 (13.2)
Eleocharis uniglumis	0 (0)	0 (0)	0 (0)	1.8 (5.6)	11.7 (11.0)
Hydrocotyle vulgaris	No significant variation from population and not contributing to the clustering				
Juncus articulatus	No significant variation from population and not contributing to the clustering				
Preissia quadrata	0 (0)	0 (0)	0.6 (1.5)	3.1 (3.4)	0.9 (0.6)
Ranunculus bulbosus	13.9 (10.7)	0 (0)	0 (0)	0 (0)	0 (0)

4. Discussion

Mapping was not an aim of these analyses, so we will not discuss this topic, except to say that segmentation clearly provides sensible boundaries not available using unconstrained clustering. It is also likely that in the length of transect available, about 4 m, it will be difficult to identify if processes operating have changed markedly. The distinction between the first 77 things and the remaining 36 could indicate that such a change has occurred but it can equally well be interpreted in terms of a patch mosaic. Certainly the original authors of the data regarded it as representative of a 'homogeneous' community (H. Gitay, personal communication).

4.1. *Ecological interpretation*

With the total length of the transect being only 11.30 m, much of the variation investigated here lies in the range of morphologically induced pattern. The results, as noted previously, suggest patches forming an alternating pattern before a major change. The major disjunction of the final 36 plots could perhaps be related to an environmental change. Data were originally collected on the abundance of various soil minerals but these do not conform to the change-point. Given a longer transect, and appropriate data, the segmentation would permit testing of environmental correlates to change-points. But it would be difficult to determine whether changes in soil minerals were the cause, and not the effect, of vegetation changes. By using tests that particular variables have an ordered response along a series of segments would suggest a continuous response to the environmental variable, but this could also be addressed by using a different model for within segment variation involving a regression on the external variable.

Although the segmentation results provide a less parsimonious description of the data than the clustering results, they also provide somewhat different information. The spatial coherence constraint imposed in segmentation provides emphasis on spatial scales and patterns of mosaics, instead of vegetation homogeneity alone, an emphasis that may be of considerable interest to management if the patterns found suggest vegetation processes are responsible (cf. Boerlijst and Hogeweg, 1991). Combining segmentation and clustering enhances the value of data and further advances the arguments for using 'gradsects' (Gillison and Brewer, 1985) as a framework for data collection.

Contribution to the study of scale is a major component of the value of segmentation. All segments were definable by specific mixtures of clusters. However, some segments have the same cluster elements. Thus the pair of segments 1 and 3, and again the pair 2 and 4, are similar and combine to suggest an alternation pattern along the sequence. Segment 5 is rather different and appears to occupy greater length. Interestingly, it would seem that each of the pairs (1,2) and (3,4)

Table 6 – Comparison of clusters and segments

Cluster	Segment				
	1	2	3	4	5
5	X		X		
6	X	X	X	X	x
7		X		X	X
14					X
15					X

X indicates a major contribution of the cluster to the segment. Cluster 6 has a single occurrence in segment 5 indicated by x.

Table 7 – Fitting segments with models for other segments

Fitted segment	Model segment 1	Model segment 2	Model segment 3	Model segment 4	Model segment 5
1	566.91	1766.87	**809.60**	1541.22	216015.02
2	27044.65	953.28	91710.19	**1040.21**	**202031.17**
3	**673.09**	1696.08	573.24	1533.49	216202.48
4	27028.05	**1050.81**	91710.44	991.18	202189.95
5	27143.99	1263.96	92012.85	1303.86	1665.78

Column minima (excluding diagonal elements) for comparison between models shown in **bold** type.

are fragments of more or less the same size as segment 5, although this may be a chance effect given the small sample of fragments. In the present case, where the size of the basic things is of the same order as that of plant clones, this may simply reflect the size of the plants themselves. Table 8 shows that indeed segments 1, 3 and 2, 4 are most similar, while section 5 is markedly different. Unlike the traditional techniques for pattern analysis derived from Grieg-Smith (1961), there is no assumption of periodicity, and the patches can be, and are, of different sizes.

The overall results suggest that the first 77 things show a repeating pattern of segments (1,3,2,4) where 1 and 3 and 2 and 4 are very similar. The final segment (5) represents a mosaic element at a different scale. The size of segment 5 is at least as large as either of the pairs (1,2) and (3,4) in the initial stages of the series (38, 41 compared with 36). The small size of the sample (only three segments are bounded at both ends) prohibits any definitive result, but the results are suggestive of the existence of several scales, one nested within one part of the other.

4.2. *Extending the segmentation procedure*

We need to consider possible extensions of the segmentation procedure used here in the following areas: shifts in variance and other properties, independence of samples and Markov models, two-dimensional segmentation and image segmentation and assessment of the magnitude of changes between segments.

4.2.1. *Shifts in variance and other properties*

While in most cases the shifts between segments are changes of mean values, MML can also distinguish between segments differing in variance. Thus the end of the first segment can also be identified with a change in variance for the species *Agrostis stolonifera* (confirmed by using the Talwar and Gentle (1981) test). Three other species, *Carex arenaria* (sample 41), *Juncus articulatus* (sample 77) *Ranunculus bulbosus* (sample 6) also individually show changes in variance at the locations indicated. Such a change in *Juncus articulatus* would be subsumed in the major change at or about that position, but the other two species have lost significance.

If segments are to reflect common textures it is necessary that they capture changes in variance (and possibly skewness). If the aim is also to identify common processes then similar within-segment trends also have to be identified. Such trends can range from simple linear trends to more complex monotone and umbrella responses and it is expected that MML or similar methods will be needed to determine the 'most supportable' model by balancing model complexity against fit to data.

4.2.2. *Independence of samples and Markov models*

It is very likely that a series of observations, such as the transect used here, will have correlation between adjacent samples. In such cases alternative models may be more appropriate than segmentation. Edgoose and Allison (1999) proposed a procedure using first order Markov models for this situation and Dale et al. (2002a) used this in examining variation through time, and also examined possible higher order Markov processes. Li et al. (2001, 2002) and Dale et al. (2002b) examined clustering of Markov models using a BIC criterion. This involves segmentation if a change in Markov model is identified at some point in the series of observations. It would also be possible to examine other time series (ARIMA) models within segments.

4.2.3. *Two-dimensional segments and image segmentation*

The series here is one-dimensional, but the application of segmentation to two-dimensional data has also been considered. Skarbeck and Koschan (1994) provide a review of methods

Table 8 – K-L distance between segments

Fitted segment	Model from segment 1	Model from segment 2	Model from segment 3	Model from segment 4	Model from segment 5
1	0	58.11	**16.88**	39.29	15310.7
2	1103.24	0	3797.37	**2.04**	8348.56
3	**7.58**	53.06	0	39.29	15443.0
4	1058.45	**3.90**	3645.49	0	**8020.97**
5	738.25	8.63	2539.99	8.69	0

Column minima for comparisons between models shown in **bold** type.

Table 9 – Location of change-points using various edge-detection methods

Species	Location (sample no.) of change-point					
	Emad-Eldin and BuHamra (1996): edge 1	Emad-Eldin and BuHamra (1996): edge 2	Kim and Siegmund (1989): change in regression on position in series	Quartile test: Barnett and Eisen (1982): mean shift	Quartile test: Barnett and Eisen (1982): variance shift	Srivastava and Worsley (1986): multivariate test (i.e. species are not identified)
Agrostis stolonifera	22	51	22		13	
Amblystegium serpens	39	78	31	76		
Carex arenaria	41	78	31		41	
Carex flacca	27	77	31	25		
Eleocharis uniglumis	40	81	31	74		77
Hydrocotyle vulgaris	43	69	15			
Juncus articulatus	35	77	31		77	
Preissia quadrata	40	78	31	24		
Ranunculus bulbosus	42	77	31		6	

Emad-Elmin test identifies two change-points in mean. Kim–Siegmund test identifies any change in a regression of species abundance against position in sequence. Barnett–Eisen test identifies changes in mean and changes in variance. Srivastava–Worsley is a test for multivariate mean change, although it is sensitive to variance changes as well. All other tests are univariate.

applied to image segmentation. Wallace (1998) looked at clustering such data while incorporating local spatial correlation using MML and Markov processes. A major problem here is providing sufficient data. A 25 × 25 grid requires 625 samples yet is a relatively small size for inductive analysis. In some cases a three-dimensional data set may be needed, where a temporal dimension is also included, but the spatial and temporal dimensions will then have different properties; spatial dimensions are symmetric, temporal ones are directional. Again the data requirements would be large, especially for the time dimension; financing recording over long periods is a difficult task.

Finding segments in vegetation data may prove difficult because, unlike images, there may be lines of sharp change but these need not connect to form discrete patches. Some preliminary experiments by M. Dale, using edge detection methods with vegetation data indicates that this is likely to occur.

4.2.4. *Assessing the 'magnitude' of a change*

An important adjunct to segmentation is the provision of some clustering of the segments themselves, because the same segment, more or less, may recur at different positions. As in our example, this will suggest that multiple scales are effective and that some patterns are nested within others, or form mosaics within larger structures. Wallace and Dale (2005) examined a fully hierarchical MML clustering approach and hierarchies may also be constructed using the Kullback–Liebler divergence (or its generalisation Bregman divergence; see Banerjee et al., 2004). The method used here provides an asymmetric measure but otherwise seems satisfactory.

5. Conclusions

The results obtained here indicate that segmentation of an observed series can contribute to determining patches of variable size. By calculating between-segment distances repetitive patterns can be identified, although finding common models for multiple segments might result in an improved model. However, segmentation results in crisp boundaries, whereas clustering without constraint permits fuzzy assignment and consistent estimation of cluster parameters. As Dale (2005) indicates, ecological data generally requires fuzzy clusters. Thus the use of clustering and segmentation are complementary, both providing information not present in the other.

Acknowledgements

We thank Dr Habiba Gitay for permission to use her data. We also thank the two anonymous referees for their suggestions for clarification and improvement of this paper.

REFERENCES

Allison, L., 2003. Types and classes of machine learning and data mining. Australian Computing Science Conference ACSC-2003, pp. 207–215.

Allison, L., 2005. Models for machine learning and data mining in functional programming. J. Function. Program. 15, 15–32.

Banerjee, A., Dhillon, I., Ghosh, J., Merugu, S., Modha, D., 2004. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. Technical Report UTCS TR04-24, UT, Austin, 2004.

Barnett, A., Eisen, E., 1982. A quartile test for differences in distribution. J. Am. Statist. Assoc. 77, 47–51.

Baxter, R.A., Oliver, J.J., 1996. The kindest cut: minimum message length segmentation. In: Arikawa, S., Sharma, A.K. (Eds.), Algorithmic Learning Theory. Lecture Notes in Artificial Intelligence 1160. Springer-Verlag, Berlin, pp. 83–90.

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

Boulton, D.M., Wallace, C.S., 1970. A program for numerical classification. Comput. J. 13, 63–69.
Critchlow, D.E., 1985. Metric Methods for Analyzing Partially Ranked Data. Lecture Notes in Statistics 34, Springer-Verlag, Berlin.
Dale, M.B., 1994. Walking a fine line: Bourne and Bound. Abstr. Bot. 17, 11–28.
Dale, M.B., 2005. On gradients and response curves. Commun. Ecol. 6, 155–166.
Dale, M.B., Dale, P.T., 1994. Multiple dissimilarity matrices: sources and classification. Coenoses 9, 9–13.
Dale, M.B., Dale, P.E.R., Li, C., Biswas, G., 2002b. Assessing impacts of small perturbations using a model-based approach. Ecol. Model. 156, 185–199.
Dale, M.B., Dale, P.E.R., Edgoose, T., 2002a. Markov models for incorporating temporal dependence. Acta Oecol. 23, 261–269.
Dale, M.R.T., 1984. The contiguity of upslope and downslope boundaries of species in a zoned community. Oikos 42, 92–96.
Dale, M.R.T., 1986. Overlap and spacing of species' ranges on an environmental gradient. Oikos 47, 303–308.
Dale, M.R.T., 1988a. The distribution of variables measuring gap and overlap in sheaves of line segments: measured by events. Util. Math. 33, 163–172.
Dale, M.R.T., 1988b. The spacing and intermingling of species boundaries on an environmental gradient. Oikos 53, 351–356.
Dijkstra, E.W., 1959. A note on two problems in connexion with graphs. Numer. Math. 1, 269–271.
Edgoose, T., Allison, L., 1999. MML Markov classification of sequential data. Statist. Comput. 9, 269–278.
Emad-Eldin, A., BuHamra, S.S., 1996. Rank test for two change-points. Comput. Statist. Data Anal. 22, 363–372.
Ferligoj, A., Batagelj, V., 1983. Some types of clustering with relational constraints. Psychometrika 48, 541–552.
Fitzgibbon, L.J., Allison, L., Dowe, D., 2000. Minimum message length grouping of ordered data. In: Arimura, H., Jain, S., Sharma, A. (Eds.), Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, pp. 56–70.
Gillison, A.N., Brewer, K.R.W., 1985. The use of gradient directed transects or gradsects in natural resource surveys. J. Environ. Manage. 20, 103–127.
Gitay, H., Agnew, A.D.Q., 1989. Plant community structure, connectance, niche limitation and species guilds within a dune slack grassland. Vegetatio 83, 241–248.
Goodall, D.W., 1953. Objective methods in the classification of vegetation I. The use of positive interspecific correlation. Aust. J. Bot. 1, 39–63.
Greig-Smith, P., 1961. Data on pattern within plant communities. I. The analysis of pattern. J. Ecol. 49, 695–702.
Hanlon, B., Forbes, C., 2002. Model selection criteria for segmented time series from a Bayesian approach to information compression. Working paper 8/2002, Department of Econometrics and Statistics, Monash University, Melbourne, Australia.
Kehagias, A., 2002. Hidden Markov model segmentation of hydrological and environmental time series. http://citeseer.ist.psu.edu/kehagias02hidden.html.
Kim, H.-J., Siegmund, D., 1989. The likelihood ratio test for a change-point in simple linear regression. Biometrika 76, 409–423.
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.
Li, C., Biswas, G., Dale, M.B., Dale, P.E.R., 2001. Building models of ecological dynamics using HMM based temporal data clustering. In: Advances in Intelligent Data Analysis, 4th International Conference on Intelligent Data Analysis, Lecture Notes in Computer Science Series, Vol. 2189. Springer-Verlag, Berlin, pp. 53–62.
Li, C., Biswas, G., Dale, M.B., Dale, P.E.R., 2002. Matryoshka: a HMM-based temporal data clustering methodology for modelling system dynamics. Intell. Data Anal. 6, 281–308.
Lombard, F., 1988. Detecting change-points by Fourier analysis. Technometrics 30, 305–310.
Losch, R.H., Cruz, F.R.B., 2002. Applying the product partition model to the identification of multiple change-points. Adv. Complex Syst. 5, 371–387.
Oliver, J.J., Baxter, R.A., Wallace, C.S., 1998. Minimum message length segmentation. Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-98), pp. 222–233.
Oliver, J.J., Forbes, C.S., 1997. Bayesian approaches to segmenting a simple time series, TR 97/336, Department of Computer Science Technical Reports, Department of Computer Science, Monash University, Melbourne, 20 pp.
Pettitt, A.N., 1979. A nonparametric approach to the change-point problem. Appl. Statist. 28, 126–135.
Pettitt, A.N., 1980. Some results on estimating a change-point using non-parametric type statistics. J. Statist. Comput. Simul. 11, 261–272.
Pettitt, A.N., 1981. Posterior probabilities for a change-point using ranks. Biometrika 68, 443–450.
Peyton Jones, S. (Ed.), 2003. Haskell 98 Language and Libraries, the Revised Report. Cambridge University Press, Cambridge.
Schwarz, G., 1978. Estimating dimension of a model. Ann. Stat. 6, 461–464.
Skarbeck, W., Koschan, A., 1994. Colour image segmentation – a survey. Technical Report 94-32 Fachbereich 13 Informatik, Technische Universität, Berlin.
Srivastava, M., Worsley, K.J., 1986. Likelihood ratio tests for a change in a multivariate normal mean. J. Am. Statist. Assoc. 81, 199–204.
Stephens, M.A., 1974. EDF statistics for goodness of fit and some comparisons. J. Am. Statist. Assoc. 69, 730–737.
Talwar, P.P., Gentle, J.E., 1981. Detection of a scale shift in a random sequence at an unknown time point. Appl. Statist. 30, 301–304.
Timone, K.P., La Roi, G.H., Dale, M.R.T., 1993. Subarctic forest-tundra vegetation gradients: the sigmoid wave hypothesis. J. Veg. Sci. 4, 387–394.
Viswanatthan, M., Wallace, C.S., Dowe, D.L., Korb, K.B., 1999. Finding cutpoints in noisy binary sequences – a revised empirical evaluation. In: Foo, N. (Ed.), Lecture Notes in Artificial Intelligence 1747. Springer Verlag, Berlin, pp. 405–416.
Wallace, C.S., 1998. Intrinsic classification of spatially-correlated data. Comput. J. 41, 602–611.
Wallace, C.S., 2005. Statistical and Inductive Inference by Minimum Message Length. Information Science and Statistics. Springer-Verlag, Berlin.
Wallace, C.S., Boulton, D., 1968. An information measure for classification. Comput. J. 11, 185–194.
Wallace, C.S., Dale, M.B., 2005. Hierarchical clusters of vegetation types. Commun. Ecol. 6, 57–74.
Wallace, C.S., Dowe, D.L., 2000. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. Statist. Comput. 10, 73–83.
Yao, Y.C., 1988. Estimating the number of change-points by Schwarz's criterion. Statist. Probab. Lett. 6, 181–187.

Glossary

Clustering: This is a procedure for subdividing the things forming the data into subgroups. Each subgroup will be characterised by similarity of its members and by disjunction from other clusters

Unsupervised clustering: Unsupervised clustering employs a single dataset and seeks clusters within it. In contrast, supervised clustering also uses an *a priori* definition of the

classes as a dependent variable, with the others representing the independent variables.

Crisp and fuzzy clustering: Most clustering methods provide crisp clusters, with every thing assigned uniquely to a single cluster. With fuzzy clustering the assignment of things to clusters is partial and any thing may have some degree of membership in several clusters. In addition to any substantive reason for accepting fuzziness in assignment, its use allows consistent estimates of cluster parameters to be obtained. Crisp clustering does not guarantee consistency. With the present data, the unsupervised clustering results have only a single thing with high probability of belonging to two clusters.

Non-hierarchical and hierarchical clustering: Non-hierarchical methods of clustering partition the data into groups without suggesting any form of linkage between them. Hierarchical methods demand that clusters are nested with 'higher' level clusters including 'lower' level clusters. In nbsp;most such methods, all clusters are arranged in a single hierarchy, although this is not a necessary condition.

Constrained clustering: In unconstrained clustering any thing may be placed in a cluster with any other. In constrained clustering two things may either be necessarily separated or necessarily placed in the same cluster using some external criterion

Segmentation, edges and change-points: If the data form a sequence of observation then it is possible to constrain the clustering so that only adjacent things are permitted in the same cluster or segment. Between each pair of consecutive segments there exists a change-point. This marks the boundary of the segments.

Coherent segments and fragments: These are names used in this paper to indicate consecutive sections of a sequence, whether formed by segmentation or clustering.