COMMUNITY ECOLOGY 3(1): 19-29, 2002
1585-8553 © Akadémiai Kiadó, Budapest

Optimal classification to describe environmental change: pictures from the exposition

P. E. R. Dale[1] and M. B. Dale

Australian School of Environmental Studies, Griffith University, Nathan, Queensland 4111, Australia.
Fax: +61 7 3875 6717, Email: p.dale@mailbox.gu.edu.au and m.dale@mailbox.gu.edu.au.
[1] Corresponding author.

Keywords: Clustering, Discrete state, Dynamics, Impact assessment, Minimum message length (MML), Runnelling, Transition matrices.

Abstract: In this paper we examine the impact of runnelling on the vegetation of a salt marsh. Runnelling is a form of habitat modification used for mosquito control in Australia. Defining the states of the system through unsupervised clustering of vegetation records using the minimum message length principle, 11 states (or classes) were identified. The runnelled sites have a greater diversity of states present than the unrunnelled ones. The states at each time for each site were then used to develop transition matrices. From these, two different pathways were identified, indicating the patterns of change. The method of showing changes relied on pictures that represent average species size and density. Both the two main pathways of change started with the dominant grass (*Sporobolus*). One led to a reduction in *Sporobolous* and ended in bare ground; the other included changes involving variation in the size and density of a mix of *Sporobolus* and *Sarcocornia*. The effects can be interpreted in terms of the increased access of seawater to the marsh resulting in an extension of the lower marsh. We note, however, that this methodology does not distinguish between changes of state within a single process and changes associated with a change in the actual processes operating.

Abbreviation: MML - minimum message length.

Introduction

One of the major problems for decision makers and managers of the environment is that of effectively assessing impacts and the changes they induce. Science can provide a wide variety of models which attempt to predict and explain ecosystem processes. However, presenting the results in an understandable and convincing form is problematic. This paper takes a bifocal approach and attempts to combine cutting edge scientific methodology for assessing the state of the environment with clarity of exposition of results, so that the outcomes are easy to understand. To do this, we focus on a particular example which has relevance to an area of public health concern: managing intertidal saltmarshes both to minimise risk of mosquito borne disease and to conserve ecological value (P.E.R. Dale 2001). The method involves modifying the marsh by runnelling to inhibit mosquito breeding. Runnelling was developed in Australia in the mid 1980's and is a form of habitat modification designed to interfere as little as possible with the ecosystem (Hulsman et al. 1989, Dale et al. 1993). Intertidal saltmarshes are protected under environmental and fisheries legislation. Indeed, under the Integrated Planning Act 1997, all development in Queensland has to 'seek to achieve' ecological sustainability, and any modification to the intertidal zone is subject to permits from government agencies. These agencies need a good knowledge base upon which to make decisions. This is one focus of this paper. The other focus is to explore a method of assessing the state of the environment and how it is changing.

Earlier work on related topics includes a study of succession using first order transition matrices (and implied Markov processes) reported in Williams et al. (1969). Other studies involving Markov processes include that of Orlóci et al. (1993) which sought to establish the usefulness of such models, and Wildi and Schütz (2000) although they are more concerned with arranging fragmentary observations into what is presumed to be a single sequence. Recently, Anand and Orlóci (1995, 1996) have examined a notion of complexity in ecology using information theory and chaos theory. These studies consider the changes in complexity during the course of succession but do not address the estimation of Kolmogorov complexity nor its potential use in operationalising Occam's razor, which is, in part, the topic addressed here. Results supporting Anand and Orlóci can be found in Dale (2000),

while other uses of Minimum Message Length assessment of models in vegetation studies can be found in Dale (2001a,b) and Dale et al. (2001).

Data and methods

General

Our procedure follows that of Williams et al. (1969). We assume that the system can be described as progressing through a series of discrete types or states. Identifying types of vegetation can then be used to assess the state of the environment and if observations are repeated over a period of time the sequence of types will indicate changes. The starting point is to classify the sample plots. That is, all plots at all times are treated as independent observations and assigned to classes[1]. An unsupervised classification can be used to identify the necessary states or types and the assignment to types of any individual sample may be fuzzy. The classes represent the more or less discrete types (or states) of the environment over the time period.

Now we can look at the actual location in time (and space) of each sample plot and see if it has changed class, and if so, in what way. We can answer questions such as 'do the treatment (modified) plots fall into classes distinct from the control (unmodified) plots?' This is not the same as asking if there are statistically significant differences between treated and control as in standard ANOVA procedures. It is rather looking at the behaviour of sample plots over time and, to some extent, represents the dynamism of the environment (see also Dale and Hulsman 1988). More formal methods for making such an appraisal exist but we shall consider these elsewhere.

Study area and data

The study site is on Coomera Island (S27° 51', E153° 33'), to the north of the Gold Coast, Queensland and close to areas of rapid population growth. It is mainly vegetated with Marine Couch (*Sporobolus virginicus* (L. Kunth) and of *Sarcocornia quinqueflora* (Bunge ex Ung.-Stern) with the Grey Mangrove (*Avicennia marina* (Forsk)) along the inlet which floods the marsh. It is also an area of major mosquito breeding. The problem species, *Ochlerotatus vigilax* (Skuse)[2], is a vector of alphaviruses such as Ross River virus and Barmah Forest virus.

To control the mosquito, a small part of the marsh (0.5 ha) was runnelled in November 1985. Runnels up to 0.30 m deep and 0.90 m wide were constructed to link isolated pools where the mosquitoes breed to the tidal source, allowing increased predator access. The method has worked to reduce mosquito populations; previous impact assessment indicates relatively little impact (Dale et al. 1993, 1996). Data used here were collected from 30 sample sites at quarterly intervals from November 1985 to November 1999. These included sites near to runnels (treatment) and those near isolated and unrunnelled pools (control). The vegetation variables measured included the size and density of *Sporobolus* and *Sarcocornia* in permanent quadrats[3]. The environmental variables included characteristics of the water table, substrate and distance from the tidal flooding front (tidal edge of the marsh). In all there were 1680 observations (each site at each time).

A few values are missing for some environmental variables because of marsh flooding at the time of collection and in one case because of collector accident. Fortunately the analytic method adopted is robust to such omissions.

Clustering

The clustering procedure adopted in this paper is a Minimum Message Length (MML) encoding method, fully described in Wallace and Dowe (2000; see also Appendix 1). The method uses a neo-Bayesian approach to separate mixtures of possibly overlapping distributions and can also be related to information theory concepts such as Kolmogorov's (1965) complexity. Basically we are seeking to optimally encode or describe the data so that we can transmit it efficiently using the code. If the data are random then we can do no better than directly encode it on a one-to-one basis; this would result in a single cluster encompassing the entire data. On the other hand, if there is pattern in the data, a code can be chosen so as to reduce the length of the message. In the present case the pattern is represented by the existence of clusters.

Assumptions and nature of the clustering method

The method separates mixtures of distributions and several choices are available depending on the nature of the data collected. Here we have assumed that the mixtures are of Gaussian distributions. We further assume that there is no within-cluster correlation between attributes. Edwards and Dowe (1998) have modified the program to incorporate a single axis of variation within clus-

1 This will be further discussed later.

2 Until recently known as *Aedes vigilax* (Skuse).

3 Density: number in 10x10 cm permanent plots; Size: height to youngest node for *Sporobolus*; length of succulent shoots for *Sarcocornia*.

ters, but this has some unfortunate consequences for consistency of estimation and has not been used here. Both these assumptions underlie other clustering methods. The effect of the assumption will be a possible increase in the number of clusters detected.

The procedure further assumes that there is no spatial or other correlation between sample sites, which is not necessarily true for the present data. Again the assumption is common to most clustering methods. Procedures for coping with such dependency are known, (Wallace 1998, Edgoose and Allison 1999) and we hope to investigate the impact of such dependency in a future study.

Finally the program utilises the precision of the measurements to dichotomise continuous variables. This means that attributes measured using a coarse measure contribute less information than those more precisely measured.

It is not necessary to actually identify the code, so long as we can determine the length of the message; shorter messages are more efficient for our purposes and indicate cluster or class structure.

Philosophy

MML can be regarded as implementing a form of Occam's Razor, in which simplicity is balanced against complexity in explaining phenomena. The former is represented by the message length needed to encode the cluster descriptions, the latter by the likelihood of the data conditional on the model. A trade-off can be made for, as we increase the number of clusters, the fit to the data improves and reduces the message length, but the complexity of cluster description increases and that increases the message length. The measure of precision is also a trade-off between extra message length for very precise statements and possible bias for very coarse statements. Maximum message length assigns every individual (e.g., sample plot) to one or more classes probabilistically. For those who feel that models should be assessed by their predictive quality, Wallace (1996) has examined the relationship of induction and prediction, concluding that "MML minimises the degree to which future data will surprise us'. This would seem to be a reasonable goal.

The message length is closely related to the posteriori probability of the model (or theory). The selection of the number of clusters relies on this probability and not on any subjective decisions or the application of rules of thumb of dubious worth. It is possible that the number of clusters is estimated to be 1, which represents the null hypothesis of no clustering. Differences in (log) probability between different models represent the odds ratio in favour of the model with the shorter message length. All measurements are in nits and a difference of about 10 nits indicates odds of over 1000:1 in favour of the shorter message.

The program

The program Snob (Boulton and Wallace 1970) estimates the message length by combining three components. These are:

- a measure of precision of measurement for any numeric or angular data;
- a measure dependent on the prior probability of a particular model, calculated from the number of clusters and the various parameters of attributes for each cluster; and
- a measure concerned with the probability of the data, conditional on the model being true.

The program allows both things and attributes to be masked. This means they are excluded from the cluster formation steps. However, they are not excluded from the assessment of the clusters, so that the things are assigned to clusters and the attribute parameters are assessed for significance.

The program outputs a variety of information, including the message length for the 1-cluster solution and for the n-cluster solution it finds as optimal. The difference between these 2 values represents redundancy, i.e., the amount of structure captured by the clustering; the larger this difference, the better. The output also includes identification of all attributes whose mean and standard deviation within a particular cluster are significantly different from those of the overall population. Finally, it assigns every sample site at each time to one or more clusters and indicates the relative probability of the thing belonging to each cluster. Thus, there is an inherent possibility of fuzziness in the assignment of units to clusters. As Wallace and Dowe (2000) explain, this fuzziness can be used to reduce the message length and it also permits the estimates of cluster parameters to be consistent whereas most commonly employed clustering methods only allow inconsistent estimates.

Applied analyses

In our analyses the things to be clustered are descriptions of sample plots at each of 56 specific times over a period of 14 years, with quarterly observations. In principle, we could use the entire temporal history of each plot as a single attribute, but there is presently no available MML procedure for such data. The descriptions we have

are of 2 kinds. First the density and size of the two principal plants *Sporobolus* and *Sarcocornia*. These are numeric variables and we used them to generate the clusters. Second, we have various measures of environmental attributes, both numeric and multistate, including salinity, pH and water content of the substrate, water table depth and salinity, density of crab holes, and distance from tidal source. These were masked, so that they did not contribute to the identification of the clusters. However, to aid in interpretation, information on the possible significance of these attributes to clusters is output by the program. For numeric attributes, the within cluster distribution may be either Gaussian or Poisson, but for all these data analyses we have chosen the Gaussian.

The process is illustrated in Figure 1. First the complete data set was clustered, using the MML method described above. The number of classes was determined by the smallest message length ("cost'), each class was described in terms of its classifying variables (the plants) and by any of the environmental variables which were significantly associated with the class. Each class was then labelled with a descriptive name, although the original class numbers have been retained in the text for brevity. The pattern of classes in time was plotted as a graph for the data as a whole as well as for the treatment and control data. Next we created a matrix for all sites, and separately for control and treatment sites showing, for each class, how many times it was followed by any other class - that is a transition matrix. Thus, if a class remained the same throughout it would have all its observations recorded in the one cell. This was used to generate a general model of change. The classes were represented symbolically to provide a picture of their character. For typical sequences of change (episodes) the classes were plotted for sites selected to demonstrate a visually effective demonstration of changes over time.

Results

Clusters and their temporal patterns

The 11-class clustering proved to be the optimal one in terms of message length. The majority of sample plots

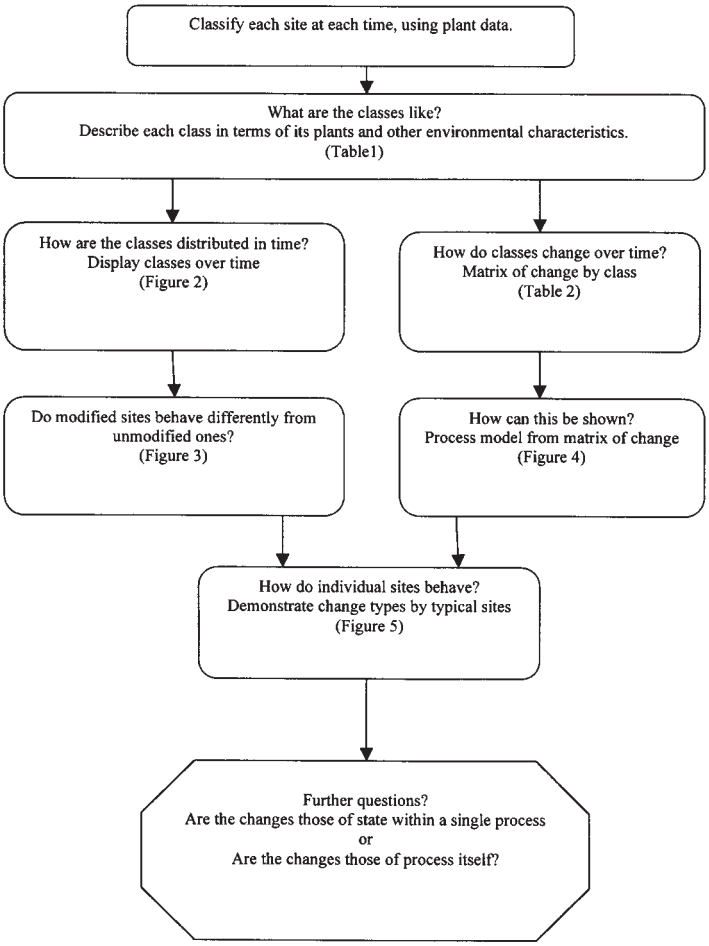

Figure 1. Flow chart of process for identifying change.

Table 1. Summary of the class descriptions.

Class # N=1680	Class Name Sp – *Sporobolus* Sa - *Sarcocornia*	Sp #	Sp height	Sa #	Sa size	Water salinity	Soil Water g/g	Soil salinity ppt	Soil pH	Dist.	Real Dist. (from tidal source)	# crab holes
1 n=685	Tall dense Sp	106.14	102.86	0	0		.61				90.69	
2 n=63	Tall Sp/Sparse Sa	80.67	101.9	8.55	51.18							0.05
3 n=98	Dense Sp/Med density Sa	79.59	82.97	32.26	37.82	33.12				2.09		0.01
4 n=41	Medium Sp/Small Sa	36.63	84.26	54.97	120.67							
5 n=155	Sparse Sp/ Small Sa	22.3	86.15	36.52	36.17				6.57			0.07
6 n=115	Very sparse Sp/Med Sa	3.52	56.16	41.96	40.72							
7 n=63	Shot Sp/ Dense Sa	9.97	43.51	75.05	50.67	35.51	.55	48.42			77.99	0.05
8 n=108	Short sparse Sp	9.85	-	0	0		.62	29.63				0.29
9 n=203	Small medium density Sa	0	0	422.83	41.92		.58					
10 n=89	Medium very dense Sa	0	0	75.2	51.55		.56					0.03
11 n=60	Bare ground	0	0	0	0	24.26	.64	26.38			68.92	0.57

were unambiguously assigned to a single class. The 1-class length (all the data ignoring any pattern) was 42357, whereas the 11-class length was only 28158. The redundancy or structure captured is the difference between these numbers, in this case 14199. The greater the size of this number the more likely it is that the clusters did not occur by chance. In the present case, this represents an odds ratio in favour of the 11 cluster solution, compared to one cluster solution of e^{14199}:1 which is a VERY large number. We therefore accept first that clusters exist in these data and second, that 11 is a reasonable estimate of the number of such clusters. Note that this does NOT mean that some other structure, such as ordination axes, might not result in a still shorter message length and hence be preferable to our cluster solution. In principle, it would be possible to make such a test, but we have not done so at this time.

Table 1 summarises the class descriptions and associated relationships with other environmental factors, not used to classify. The types ranged from tall dense *Sporobolus*, through mixtures of *Sporobolus* and *Sarcocornia* to *Sarcocornia* alone. Tall dense *Sporobolus* occupies the slightly elevated areas at some distance from the lower marsh edge, whereas *Sarcocornia* tends to be found in the lower parts of the marsh. The very sparse *Sarcocornia* (class 8) and the bare ground (class 11) have significant crab activity, wet substrate and lower salinities.

Figure 2 shows the pattern of the classes over time for all the sites. At the start of the experiment in November 1985, six of the 11 classes were present. These ranged from the tall dense *Sporobolus* to the dense *Sarcocornia* with a small amount of *Sporobolus* (Table 1). At the last reported observation only 3 of the starting classes remained in the system. Gone were the tall relatively dense *Sporobolus* with moderate amounts of *Sarcocornia* (3) and the two classes of relatively dense *Sarcocornia* with some *Sporobolus* (6 and 7). Classes 8, 9, 10 and 11 had developed over the course of the experiment. One class had only very few and very small *Sporobolus* plants. Two contained *Sarcocornia* and no *Sporobolus* and one contained no plants at all. So the method shows a gross change over time of diminished vegetation cover.

Looking at the results in more detail there appear to be several types of classes based on their temporal behaviour. First, there are the persistent classes which are always or almost always present at the marsh. Always present is the tall dense *Sporobolus* class (1). Mostly present were classes 2, 5 and 6. These are mixtures of *Sporobolus* and *Sarcocornia* of at least moderate size and density. Second, there are the classes which were there at the start but which became extinct. These were classes 3 and 7. They contain tall and relatively dense *Sporobolus* (3) or medium and sparse *Sporobolus* but with dense *Sarcocornia* (7). Third are classes which developed over the

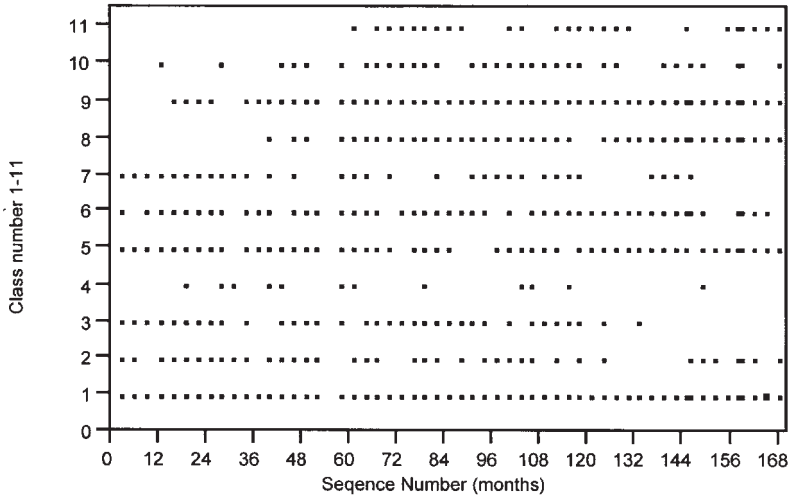

Figure 2. Overall pattern of change for all sites (dots indicate presence of the class at the specific time).

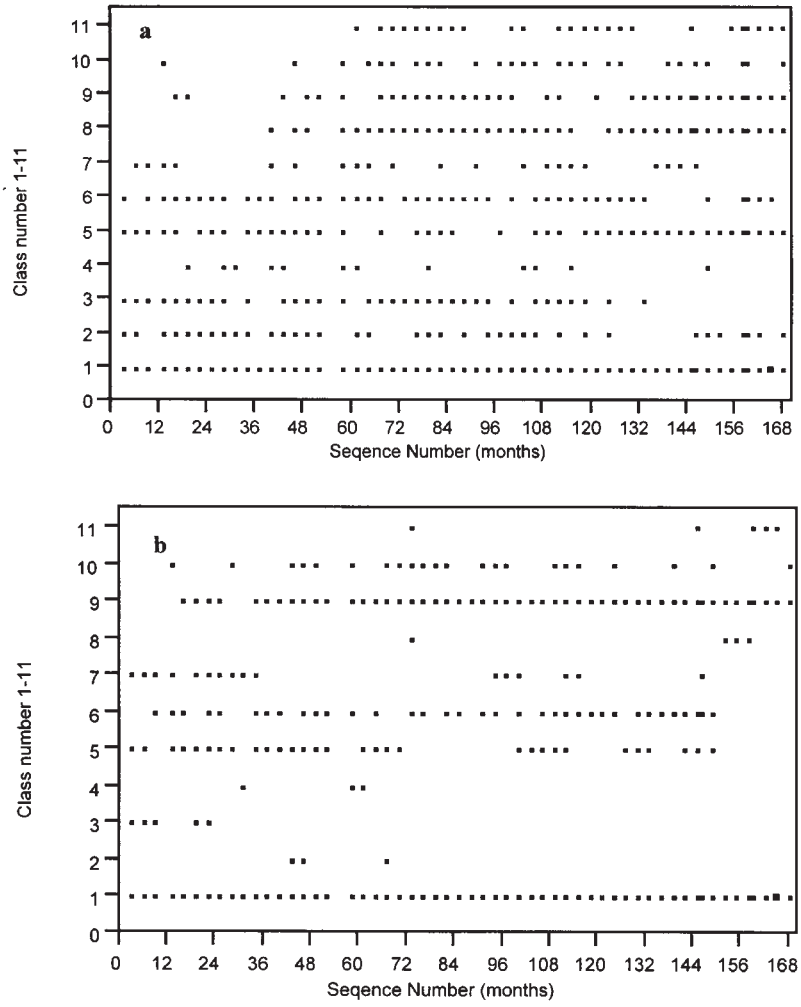

Figure 3. Change pattern for **a**. Runnel (treatment) sites and **b**. Pool (control) sites (dots indicate presence of the class at the specific time).

course of the experiment and persisted. These were classes 8 and 9 both with short sparse plants (*Sporobolus* in class 8; *Sarcocornia* in class 9). Fourth are classes which developed but were periodic in their presence. These were classes 4, 10 and 11. Class 4 had relatively large *Sporobolus* and *Sarcocornia*, but of medium density. Class 10 is of dense *Sarcocornia* and 11 is bare mud. Class 11 did not appear until the end of year 5.

Having established the nature of change in the system as a whole, the next question is 'how do the runnelled sites behave compared to the unrunnelled ones (pools)?'

Table 2. Matrix of change for all sites.

Change from class	Change to class											
Class	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	TOTAL
1	*603*	16	6	1	4	0	0	26	0	0	2	658
2	**16**	*22*	**14**	2	7	1	0	0	1	1	0	64
3	3	10	*49*	6	15	0	1	0	0	0	0	84
4	0	3	5	*4*	**15**	5	3	0	0	0	0	35
5	7	10	4	**14**	*87*	**21**	9	7	2	1	0	162
6	0	0	1	4	15	*61*	15	2	**28**	5	3	134
7	0	0	0	1	**15**	**16**	*23*	0	1	1	0	57
8	20	2	0	1	2	3	0	*76*	0	0	**17**	121
9	0	0	0	1	0	**21**	1	1	*148*	**35**	**8**	215
10	0	0	0	1	0	3	2	0	**36**	*24*	1	67
11	0	0	0	0	0	1	0	**14**	5	4	*29*	53
TOTAL	649	63	79	35	160	132	54	126	221	71	60	1650

Figure 3 shows the distribution of the classes over time again, but separates treatment from control.

The most striking difference between the runnelled and unrunnelled sites is in the greater variety of classes or states exhibited by the runnelled sites at any one time, although both consistently contain the persistent *Sporobolus* class (1). Two of the 3 persistent classes (2 and 6) disappear from the unrunnelled sites. Class 2 (tall dense *Sporobolus* with some large *Sarcocornia*) does not persist beyond the end of year 2; Class 6 persists for over 12 years, but has been missing for the last 2 years. It has periodically gone from the unrunnelled sites but usually reappears within a year. Class 8 (with just a small amount of very short *Sporobolus*) first appears in the runnelled sites after 3 years. It does not appear in the unrunnelled sites until briefly in year 5, and later in year 13.

Transition matrices

Table 2 shows the matrix of change for all sites. There was no obvious difference between treatment and control sites in this regard; a formal test was not applied but is available, see Kullback et al. (1962). From the table a sequence model can be constructed using the common sequences (Figure 4). This illustrates the fate of classes. It appears that a site in the dominant class tall dense *Sporobolus* (1), if it changes at all, may change in two major directions. On one hand it can lose size and density of *Sporobolus* and become bare ground (class sequence 1-8-11). This appears to be associated with wetness (indicated by significant associations with high soil water). In this state it may oscillate between bare ground, short *Sporobolus* and tall dense *Sporobolus* as shown by the site in Figure 5a. If it follows the other path it may change by *Sarcocornia* colonising the site (Figure 5b and c). If this happens then *Sporobolus* may become shorter, less dense and the site may have only *Sarcocornia* (Classes 9 and 10). Figure 5b shows part of the sequence for a site that tends to occupy the upper part of the longer sequence in Figure 4. Figure 5c illustrates a site that tends to occupy the lower part of the longer sequence in Figure 4. It is rare for a site to get out of the class 1 to 8 loop, but it does happen occasionally. However, such sequences do not usually get to the bare ground stage. From Table 3, it can be seen that bare ground (11) generally remains as bare ground.

What is this all telling us? The changes, particularly in the runnelled sites, appear to be associated with decreasing vegetation and especially of decreasing *Sporobolus*. Although Class 1 is a persistent one it is becoming less dominant and accounting for a declining proportion of the total vegetation at the sites. This is so for runnelled and unrunnelled sites and perhaps reflects overall changes in the general area; during part of the period the area was subject to very low rainfall. The only class which has increased its proportion of the overall vegetation is class 9 (low density small *Sarcocornia*, typical of the low marsh). Otherwise the only other increasing contributors are found as Class 8 mainly in the runnelled sites though increasing slightly in the unrunnelled sites, too. Class 8 is one with very sparse and very short *Sporobolus* and may be a class indicative of the decline of a *Sporobolus* dominated class.

Discussion

The classes are consistent with those obtained by Dale et al. (1986) for the larger marsh based on aerial survey and classifying spectral reflectance recorded on large scale colour infrared aerial photographs. There, 8 classes could be discerned based on the size and density of the two dominant species. These ranged from monospecific stands of tall dense *Sporobolus* through mixtures of dominant *Sporobolus* with *Sarcocornia*, to dominant *Sarco-*

cornia with *Sporobolus* and thence to monospecific stands of *Sarcocornia* (Dale et al 1986). However, in the present study, there was an additional class of bare ground or mud that was not identified in the earlier research. This, and the pattern of classes over time, indicate that bare ground is a newcomer to the area. The bare ground also was associated with larger numbers of crab holes and this is consistent with Chapman et al. (1998) who reported increased crab numbers at the Coomera site in the runnelled area, compared to the unmodified marsh.

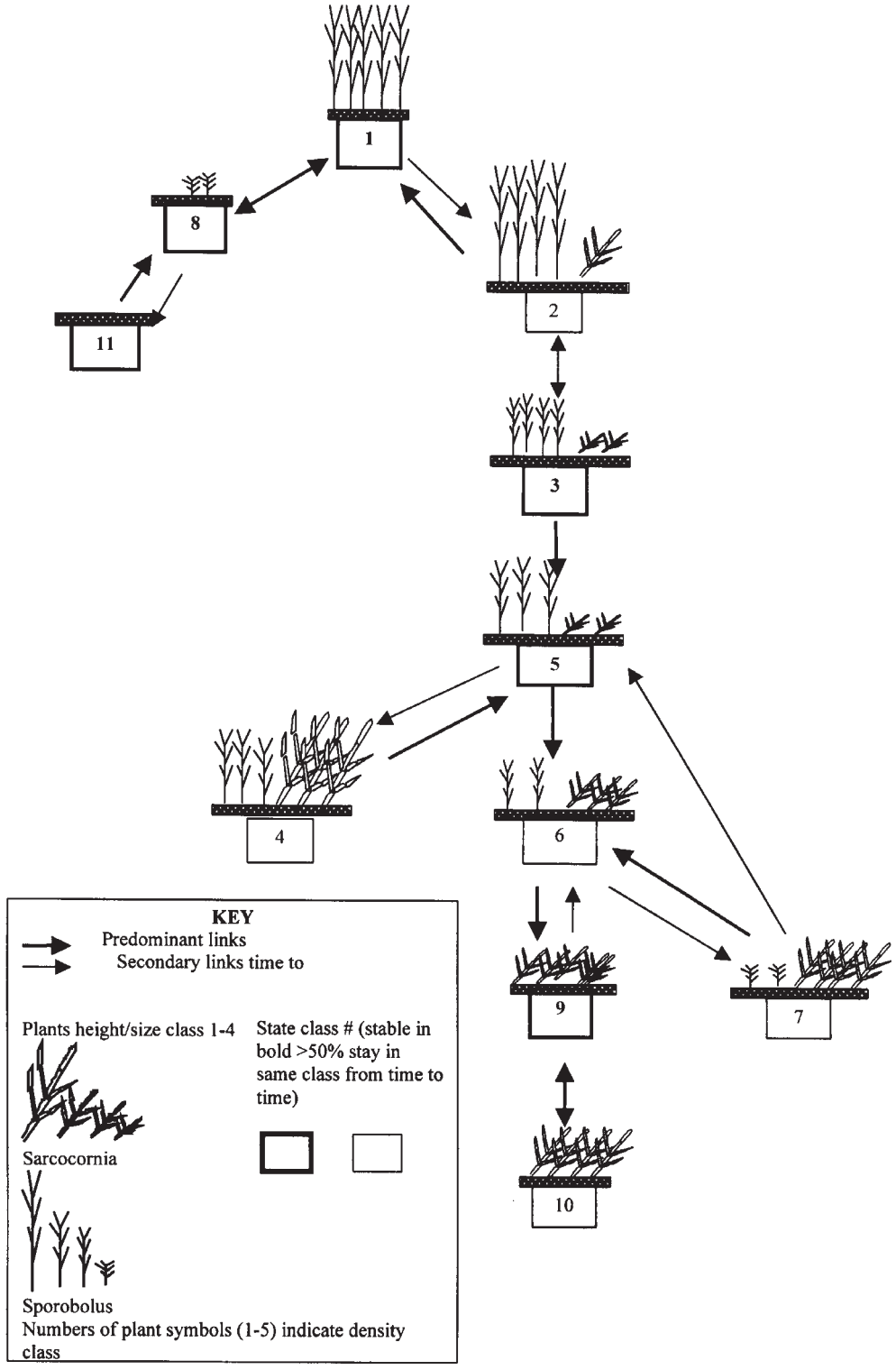

Figure 4. Diagrammatic sequence model.

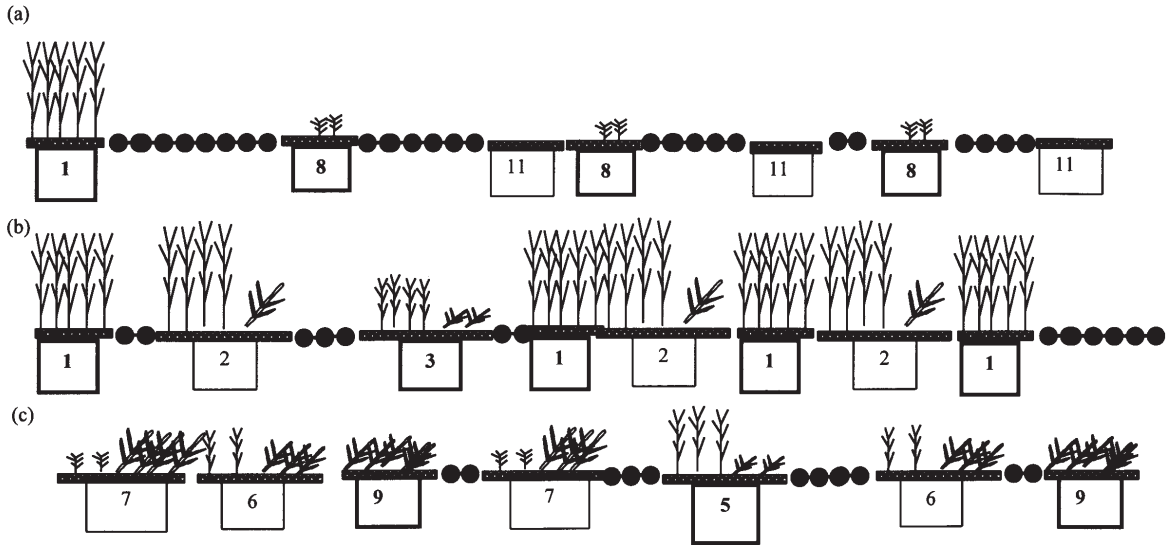

Figure 5. Exemplary episodes of changes. Dots represent continuation of the preceding type. **a.** from tall dense *Sporobolus* to bare ground. **b.** *Sporobolus* (dominant)-*Sarcocornia* mixtures. **c.** *Sarcocornia* (dominant)-*Sporobolus* mixtures. Key as for Figure 4.

The intermittent classes may represent a dynamic state which particular sites go through as the vegetation diminishes. The decline in the grass *Sporobolus*, generally dominant in the higher areas (Adam et al. 1988), and the increase in *Sarcocornia*, especially for runnelled sites, indicates an environment which may be acquiring low marsh characteristics at the expense of the high marsh. Indeed, Adam et al. (1988) reported that *Sarcocornia* is found near to mangroves (low marsh). The results are consistent with the nature of the modification that brings tidal water on to the marsh more often than before modification and flushes it more frequently. There are other indications that the marsh may be taking on low marsh characteristics, such as the establishment of *Avicennia marina* mangroves in slow flowing runnels and connected pools and in the apparent increased use of the habitat by crabs (*Australoplex tridentata*), which are more usually found in low marsh/mangrove habitats.

The greater variety of classes in the runnelled sites compared to the unrunnelled one could represent a greater habitat diversity in the former. Does runnelling increase habitat diversity? There are some circumstantial and unpublished data to support this! Alternatively, it could be related to an ongoing change whereby some at least of the classes represent transitions from one state to another. Is this precluded by the fixed discrete nature of the classes with invariant boundaries? It is interesting that Liebovitch (1995) has suggested possibilities of defining and using states with labile boundaries.

Conclusion

The methods we have employed provide a theoretical framework for identifying and investigating changes. This is the role of science in the project. To render the results interpretable at a management level, we have represented the change model in pictorial form whilst maintaining a close relationship between the pictures and the class characteristics as identified by the exposition. We have also shown that changes have occurred in the marsh vegetation, though not all are related to the runnelling activities.

But changes can be of two kinds. First, there may be a single suite of processes operating and the impact means only that some sites shift between states of that suite. Second, it may be that the process suite itself has been changed so that there exists more than one process suite operating over the time period being examined. In the second case, the impact is clearly more severe, for we have introduced new processes.

In effect, for this research we have assumed that the system is driven by a single set of processes and that our activities did not change this. This is assumed by most studies of change, including those using BACI (before-after, control-intervention) methods. But it is obvious that an action that modifies the processes in operation in an ecosystem is likely to have more significant impacts than one which simply changes the state of the system. We are

presently exploring alternative methods that permit the identification of this situation.

Acknowledgments. We thank the many student field assistants who have helped to collect the field data. We thank the Gold Coast City Council for providing boat transport to the study site for the whole time period. Financial support has come from the Gold Coast City Council and the Mosquito and Arbovirus Research Committee as well as from Queensland State Health.

References

Adam, P., N. C. Wilson and B. Huntley. 1988. The phytosociology of coastal saltmarsh vegetation in New South Wales. *Wetlands (Australia)* 7: 35-84.

Anand, M. and L. Orlóci. 1995. On the notion of system complexity and its definitions in ecology. *Western Journal of Graduate Research*. 5: 71-83.

Anand, M. and L. Orlóci. 1996. Complexity in plant communities: the notion and quantification. *J. theoret. Biol.* 179: 179-186.

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

Chapman, H., P. E. R. Dale and B. H. Kay. 1998. A method for assessing the effects of runnelling on salt-marsh grapsid crab populations. *J. Amer. Mosquito Assoc.* 14: 61-68.

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

Dale, M. B. 2001a. Functional synonyms and environmental homologues: an empirical approach to guild delimitation. *Community Ecol.* 2:67-79.

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

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

Dale, P. E. R., K. Hulsman and A. L. Chandica. 1986. Classification of reflectance on colour infra red aerial photographs and subtropical salt marsh vegetation types. *International J. Remote Sensing* 7: 1783-1788.

Dale, P. E. R. 2001. Wetlands of conservation significance: mosquito borne disease and its control. *Arbovirus Research in Australia* 8: 102-108.

Dale, P. E. R. and K. Hulsman. 1988. To identify impacts in variable systems using anomalous changes: a salt marsh example. *Vegetatio* 75: 27-35.

Dale, P. E. R., P. T. Dale, K. Hulsman and B. H. Kay. 1993. Runnelling to control saltmarsh mosquitoes: long-term efficacy and environmental impacts. *J. Amer. Mosquito Control Assoc.* 9: 174-181.

Dale, P. E. R., A. L. Chandica and M. Evans. 1996. Using image subtraction and classification to evaluate change in subtropical intertidal wetlands. *International J. Remote Sensing* 17: 703-719.

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 Art. Intell 1394 Springer, pp. 96-109.

Hulsman, K., P. E. R. Dale and B. H. Kay. 1989. The runnelling method of habitat modification: an environment focussed tool for salt marsh management. *J. Amer. Mosquito Control Assoc.* 5: 226-234.

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

Kullback, S., M. Kupperman and H. H. Ku. 1962. Test for contingency tables and Markov chains. *Technometrics* 4: 573-608.

Liebovitch, L. S. 1995. Ion channel kinetics. In: P. M. Iannaccane and M. K. Khokha (eds.), *Fractal Geometry in Biological Systems: an analytical approach*. CRC Press, London. pp. 31-56.

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

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

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

Wallace, C. S. and D. L. Dowe. 1993. MML estimation of the von Mises concentration parameter. Technical Report TR 93/193. Dept. Computer Science, Monash University, Clayton 3168, Victoria, Australia.

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. Royal Statist. Soc.* B 49:240-252.

Wildi, O. and M. Schütz. 2000: Reconstruction of a long-term recovery process from pasture to forest. *Community Ecol.* 1: 25-32.

Williams, W. T., G. N. Lance, L. J. Webb, J. G. Tracey and M. B. Dale. 1969. Studies in the numerical analysis of complex rain forest communities III. The analysis of successional data. *J. Ecol.* 57: 515-53.

Appendix

The Snob program is available for non-profit use (in the form of FORTRAN 77 source code from http://www.csse.monash.edu.au/~dld/Snob.html) and a detailed description of the algorithm and various heuristics used is in print (Wallace and Dowe 2000) The following is therefore an abbreviated description only. Note that minimising message length is equivalent to maximising the posterior probability. A message of length m bits is equivalent to a probability of $p=2^{-m}$. MML estimation leads to consistent estimates which are efficient in that they converge rapidly to any true underlying parameter value. The estimates are also invariant under 1-to-1 parameter transforms.

To evaluate the quality of any clustering we have to calculate the associated message length. This involves two components, one based on the extant clusters, the other on the fit of the data assuming correct assignment to these clusters. The message must include a statement of the number of clusters. For each cluster and each attribute we also calculate the message length needed to encode the necessary parameters of the attribute for that cluster. Each

type of attribute is associated with different parameters, so the exact expressions vary with type. Thus multistate attributes require specification of probabilities of occurrence of all states, Gaussian attributes need mean and standard deviation, Poisson attributes need only the Poisson parameter, while angular attributes have mean and concentration, and can be specified in degrees or radians.

As an example, the minimal message length needed to state both the parameter estimates and encoding things in the light of these estimates is given by

$$(M\text{-}1)/2 \log(N/12+1) - \log(M\text{-}1)! - S_m (n_m + 0.5) \log p_m$$

where there are M states and n_m is the number of things in state m and $N=n_1 + n_2 \ldots + n_M$, and $p_m = (n_m+0.5)/(N+M/2)$. The difference between the minimum message length estimates of p_m and the maximum likelihood estimate is due to the Fisher information term in the former, which for the case of $M=2$ is given by $N/(p_1(1\text{-}p_1))$. When calculating the message length associated with fit various corrections are made, for example, for the multistate case where $M>N$. Missing values may be specially coded and such coding does not affect the minimisation of message length

For numeric attributes, the parameter values are only expressed to some necessary precision determined by the program. Over-precise specification of parameters increases the message length while too coarse a precision will reduce the quality of fit, so some balance must be reached. This can be quantified using lattice constants for optimally tessellating Voronoi regions (Wallace and Dowe 1993).

Having thus encoded the parameters for all attributes, we can calculate the fit of the things to the clusters to which they are assigned. This is obviously based on the probability of observing such a thing given the specific cluster parameters for each attribute. Wallace and Dowe also show that by introducing probabilistic assignment the message length can actually be reduced while the estimates of parameters become consistent. However, rather than using partial assignment the program effectively assigns a thing to a class probabilistically which, on average, gives approximately the same result. The approximation is due to the use of a quadratic Taylor series expansion rather than what Wallace and Freeman (1987) call ‘strict MML’

The basic search algorithm used is an EM routine, also known as a Pickard iteration and similar to the well-known k-means algorithm. Overall, it seeks to move things between clusters to try and reduce the overall message length. The program will utilise some user-fixed number of re-assignment iterations (‘adjustments’) unless the message length does not change for 30 iterations. Control of the program is interactive although the user can set up a control file to perform a specific series of operations. The user can also save configurations and recover them.

Besides the adjustment operations, the program also employs splitting and merging heuristics. For all large enough clusters, the program retains 2 sub-clusters, randomly initialised, and it is these that are actually used in the re-allocation scheme. The program can therefore determine if splitting any cluster into its subclusters will reduce the message length still further. Occasionally these sub-clusters may be remade to avoid local minima. Very small clusters are suppressed and their members allocated elsewhere. However, any outliers in a cluster can be identified because their contribution to the fit component of the message length is overly large. To investigate merging of clusters, the program attempts to fuse pairs of clusters. To save time, it first assesses such merges on the assumption that no things will get moved to other clusters as a result of the merge. Only if the merge looks promising is a full evaluation attempted.

The heuristic search does not guarantee a global optimum although, in theory and with sufficient data, this might be attained (in fact there is a slight bias in sampling from the posterior distribution of the number of clusters and although methods for avoiding this are known they are not presently included in the program). To overcome local convergence, the program can be initialised using random partitions into a user-specified number of clusters or from a user-specified partition of some selection (not necessarily all) of the things. Note that the number of clusters used to initialise the procedure need not, and generally will not, remain constant once the analysis commences. In practice, we have started from 2 positions, one with very few clusters, the other with very many. The results from these usually provide close enough bounds on an approximately ‘correct’ number of clusters and several random starts can then be initiated in this region.

The program provides reports on classes, attributes and things. Although it is possible to mask attributes and things so they do not contribute to the clustering itself, information concerning them will be present in the reports. For attributes, the program further assesses whether the cluster parameters can be regarded as significantly different from those of the entire population. For things, a message length is reported so that outliers can be easily identified.