

Bayesian clustering with AutoClass explicitly recognises uncertainties in landscape classification

J. Angus Webb, Nicholas R. Bond, Stephen R. Wealands, Ralph Mac Nally, Gerry P. Quinn, Peter A. Vesk and Michael R. Grace

J. A. Webb (angus.webb@unimelb.edu.au) and M. R. Grace, Water Studies Centre, Dept of Chemistry, Monash Univ. and the eWater CRC, Clayton, Victoria 3800, Australia. (Present address of J.A.W.: School of Social and Environmental Enquiry, The Univ. of Melbourne, Parkville 3010, Victoria, Australia.) – N. R. Bond, S. R. Wealands and G. P. Quinn, School of Biological Sciences, Monash Univ. and the eWater CRC, Clayton, Victoria 3800, Australia. (Present address of S.R.W.: School of Social and Environmental Enquiry, The Univ. of Melbourne, Parkville 3010, Victoria, Australia. Present address of G.P.Q.: School of Life and Environmental Sciences, Deakin Univ., Warrnambool, 3280, Victoria, Australia.) – R. Mac Nally and P. A. Vesk, Australian Centre for Biodiversity: Analysis, Policy, Management, Monash Univ. and the eWater CRC, Clayton, Victoria 3800, Australia. (Present address of P.A.V.: School of Botany, The Univ. of Melbourne, Parkville 3010, Victoria, Australia.)

Clustering of multivariate data is a commonly used technique in ecology, and many approaches to clustering are available. The results from a clustering algorithm are uncertain, but few clustering approaches explicitly acknowledge this uncertainty. One exception is Bayesian mixture modelling, which treats all results probabilistically, and allows comparison of multiple plausible classifications of the same data set. We used this method, implemented in the AutoClass program, to classify catchments (watersheds) in the Murray Darling Basin (MDB), Australia, based on their physiographic characteristics (e.g. slope, rainfall, lithology). The most likely classification found nine classes of catchments. Members of each class were aggregated geographically within the MDB. Rainfall and slope were the two most important variables that defined classes. The second-most likely classification was very similar to the first, but had one fewer class. Increasing the nominal uncertainty of continuous data resulted in a most likely classification with five classes, which were again aggregated geographically. Membership probabilities suggested that a small number of cases could be members of either of two classes. Such cases were located on the edges of groups of catchments that belonged to one class, with a group belonging to the second-most likely class adjacent. A comparison of the Bayesian approach to a distance-based deterministic method showed that the Bayesian mixture model produced solutions that were more spatially cohesive and intuitively appealing. The probabilistic presentation of results from the Bayesian classification allows richer interpretation, including decisions on how to treat cases that are intermediate between two or more classes, and whether to consider more than one classification. The explicit consideration and presentation of uncertainty makes this approach useful for ecological investigations, where both data and expectations are often highly uncertain.

Clustering (unsupervised classification) of multivariate data is common in ecological research. For each of a number of cases (e.g. sampling units, individuals, populations), a researcher has data on a number of attributes (i.e. for sampling units, the abundances of many species and measures of environmental parameters). The cases lie along continuous gradients in multivariate space, but it is a feature of human

psychology to group objects rather than to recognize this continuity (Lakoff 1987). Clustering algorithms can be used to sort the cases into a number of classes, the members of which are more similar to each other than they are to members of other classes. This simplifies interpretation of the data, but it must be remembered that this simplification is the imposition of a false “order” on the data. For real-world data sets,

there are no “true” classes, and it is up to the clustering algorithm, with guidance from the user, to find the most parsimonious groupings.

There are dozens of approaches for clustering multivariate data (see review by Jain et al. 1999). Different algorithms will often yield different classifications for a given set of data, and no single technique can be shown to perform better than all others for all problems (Jain et al. 1999). We believe that ecologists probably over-utilize readily accessible and traditional clustering techniques without necessarily considering which approach might best be suited to a particular problem. Limitations of clustering and uncertainty of outcomes are also rarely considered.

Clustering algorithms belong to two main approaches: hierarchical and partitional. Hierarchical approaches, which are comprised of the familiar agglomerative and divisive methods, are more flexible than partitional methods but are significantly more computationally expensive (Jain et al. 1999). The output from a hierarchical clustering is a dendrogram, which shows the relationships among cases, based on a nested series of partitions of the data (here, “partition” is used in the mathematical sense, where a partition of a set S is a collection of disjoint subsets that have S as their union [Rosen 2003]). Usually, the user decides which level of the partition to treat as the final classification of the data (Jain et al. 1999). Hierarchical approaches almost always produce “hard” classifications of the data, in which individual cases are assigned to a single class (but see Wang and Leou 1993 for a counterexample). Partitional methods divide the data set into a single partition, rather than the nested series produced by hierarchical methods. Partitional algorithms can be used to produce “fuzzy” classifications of the data, where cases are not assigned to an individual class, but membership values for all classes are calculated for each case. For an individual case, the highest membership value indicates its most likely class. As stated above, clustering imposes a false order upon data, and there are no true classes. The fuzzy approach to class membership recognises this, and is therefore more realistic than simple inclusion or exclusion. Moreover, there will almost always be cases that lie part way between two or more larger groups in multivariate space. A hard clustering algorithm will assign these cases to one of the two groups, failing to recognise the poor fit. A fuzzy algorithm avoids this problem by computing an appropriately low membership value for these “intermediate” cases.

Regardless of the method chosen, finding the single optimum classification of a data set becomes computationally intractable with large numbers of cases and/or attributes. For hierarchical algorithms, computational time and memory requirements increase as a polynomial function of data set size (Kurita 1991, Jain et al.

1999). Therefore, for large data sets, hierarchical methods may be impractical because of memory requirements and the time taken to reach a solution. Conversely, the time and memory requirements of partitional methods increase only linearly with data set size (Belbin 1987, Jain et al. 1999). This allows them to be used to analyse large data sets, although they will generally return sub-optimal solutions due to the impossibility of covering all of the search space. For the majority of partitional algorithms, the order in which the data are presented (Upal and Neufeld 1996) and the initial assignment of clusters from which the algorithm begins (Belbin 1987, Ter Braak et al. 2003) also affect the solution obtained. These problems are exacerbated when working with large data sets. We believe that most clustering software packages do little to inform the user of this type of uncertainty in clustering results. Rather, clusters are presented as a definitive representation of group structure in the data.

Here we utilise a fuzzy partitional method, Bayesian mixture modelling – implemented in the software package AutoClass (Hanson et al. 1991, Cheeseman and Stutz 1996) – to classify catchments (i.e. watersheds) within the Murray Darling Basin (MDB) of south-eastern Australia. Our choice of method was influenced by the nature of the problem. Primarily, we believed that although the landscape should form natural clusters, these clusters would not be crisp, with the continuous nature of environmental gradients within the basin leading to some cases being intermediate between main clusters. Better-known fuzzy clustering algorithms (e.g. fuzzy k-means; Bezdek et al. 1984) require the user to specify both the number of classes, and the degree of fuzziness for the clustering. These essentially arbitrary choices have major influences on the clustering result (Bolliger and Mladenoff 2005). AutoClass requires no such inputs, and calculates the number of classes and their memberships directly from the data. The membership values for individual cases are also probabilities, and as such are more easily manipulated than membership values based on distances or similarities. For example, it is straightforward to calculate joint or conditional probabilities of certain membership combinations of particular interest. We also required a method that allowed both continuous and categorical data to be included in the same clustering, and such clustering algorithms are uncommon (but see Belbin 1987, Kaufman and Rousseeuw 1990, Ter Braak et al. 2003 for some examples). Our results show that the Bayesian method produces intuitively sensible classifications of catchments, and also provides more information on the classification than most alternative methods. The probabilistic treatment of results at all scales also provides substantial information on classification uncertainty, which is a major advantage for interpretation.

Bayesian mixture modelling with AutoClass

For a detailed description of the algorithms used in AutoClass see Hanson et al. (1991), Cheeseman and Stutz (1996), and the AutoClass Project home page (<<http://ic.arc.nasa.gov/ic/projects/bayes-group/autoclass>>). We provide here a précis of its operation.

The method used for clustering is finite mixture modelling (McLachlan and Peel 2000). The observed distribution of data is modelled as a mixture of some finite number of component distributions, the objective being to determine the number of distributions, their parameters, and memberships. The computational approach taken for mixture modelling allows the software to handle very large data sets (up to millions of cases), with computational time increasing only linearly with size of the data set (Cheeseman and Stutz 1996). The most defining characteristic of the Bayesian approach to finite mixture modelling is the fully probabilistic nature of the classification and the presentation of results. At all levels of the classification (data specification, class specification and the classification chosen), uncertainty is explicitly reported.

Prior probability distributions are required for all model parameters because of the Bayesian underpinnings of the method. AutoClass requires the use of flat (uninformative) prior distributions, and this approach has one important implication for assessing the resulting classification. Through a property that Cheeseman and Stutz (1996) refer to as the “Occam factor”, the most parsimonious classification is guaranteed to have the highest posterior probability; the probability of the model being correct given the data; $p(\theta|y)$. There is an automatic trade-off between model complexity (the number of classes) and model fit (posterior probability of the model being correct). The use of flat priors for multiple models across the same data set ensures that the marginal likelihood – the probability of the data being found given the model $p(y|\theta)$ – is proportional to the posterior probability (Cheeseman and Stutz 1996, Gelman et al. 2004). Thus, in AutoClass, the competing classifications are assessed based on an estimate of marginal likelihood (Cheeseman and Stutz 1996). This avoids the need to calculate posterior density functions. This “criterion” for model selection achieves a trade-off similar to that achieved by post-hoc information-criteria approaches (Burnham and Anderson 1998), but without the need to calculate any additional criterion following model fitting. It also avoids the problem that different information criteria will lead to selection of different models (McLachlan and Peel 2000).

The algorithm leads to multiple plausible classifications, which are ranked on their estimated marginal

likelihoods. Given that the models are very simplified abstractions of the data, the model-specific probabilities are always vanishingly small. They are also imprecise (Cheeseman and Stutz 1996). These limitations notwithstanding, the probabilities allow the user to assess the relative merits of the alternative classifications, but if the marginal likelihoods of two classifications are within approximately $e^5 - e^{10}$ of each other (in the order of 100 to 10 000 times more probable), the user should examine both and consider them to be approximately equally probable (Anon. 2002).

Within a classification, individual cases are probabilistically associated with each class. The vast majority of cases usually have a very high probability of membership of one class and can be regarded as “belonging” to that class. However, there are usually a few cases that have substantial ($p > 0.20$) probabilities of belonging to two or more classes.

We used AutoClass C (ver. 3.3.4), in which continuous data can be modelled as normally or log-normally distributed, and discrete data are treated as a multinomial distribution. Two or more continuous attributes can be specified as covarying, which leads to cases being classified by changes in the relation between the covarying attributes rather than by absolute differences in their values. This property can be illustrated by considering a group of objects described by length, width and depth. If these attributes are considered independent, the objects will tend to be classified based on size. However, if the attributes are specified as covarying, they will tend to cluster according to shape, since the correlation between attributes differs between shapes (Anon. 2002). AutoClass requires the user to specify measurement uncertainty for continuous variables. Data that are more precise have more influence on the final classification.

The reports generated contain much information that can be used to understand and interpret results. As discussed above, the marginal likelihood for the classification is supplied, which can be used to compare alternative classifications. The breakdown of probabilities of membership of each class also is supplied for each case, allowing the identification of specific cases that do not fit “neatly” into classes. For each class, there is information on the class strength – the probability that the attribute distributions at the class level can be used to predict the class members. There is also information on the importance of the individual attributes, both for the classification overall and for each class. The divergence measure is the Kullback-Leibler distance (or relative entropy; Cover and Thomas 2005). This is a useful measure of distance between data distributions because it takes into account both the centre of the distribution and the variability of the data around the centre. However, it is not a true metric because distances are not symmetric. Thus the distance

from distribution Q to distribution P does not necessarily equal the distance from P to Q. The D_{KL} of P to Q for discrete variables is defined as

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

with the equivalent for continuous variables, where p and q denote the densities of P and Q, defined as

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

The modelled distribution of each attribute for each class is also supplied, ranked by the attribute's relative importance in describing the class. Lastly, the overall divergence of each class from the overall distribution of cases, calculated as the summed D_{KL} over all attributes, is reported.

Case-study: classifying catchments in the Murray Darling Basin (MDB)

We classified catchments in the MDB as a first step in modelling the effects of catchment-scale influences (e.g. land use) on the ecological condition (e.g. fish diversity) of rivers and streams. To date, most attempts to explicitly model the effects of catchment-scale influences on waterways have had limited success (Allan 2004). By grouping the MDB into catchments with broadly similar physiographic characteristics, we expect a greater similarity of ecological processes within classes, and hence more similar responses to catchment-scale influences. This, in turn, should lead to simpler models that are better predictors of ecological condition within that class of catchments.

To obtain catchment boundaries within the MDB, the latest version of the 9'' Digital Elevation Model (DEM) of Australia was used (Hutchinson et al. 2000). This model has an approximate cell size of 250×250 m. Using standard terrain analysis functions in ArcGIS 9.0 (Anon. 2004), catchments of different stream orders were defined using Strahler ordering (Strahler 1957). Here, we report only on order-4 catchments, of which there were 401 within the MDB. Based on the Strahler-order catchment definition, some parts of the MDB were not considered (e.g. an order-3 catchment draining directly into an order-5 catchment; see empty spaces in Fig. 1). The classification described in this paper was based on physiographic characteristics of the catchments, which are largely independent of direct human influence (Table 1). The attributes described for each catchment were generated from either the DEM, the Climatic atlas of Australia (Anon. 2003), or from the MDB Soil Information Strategy (Bui and Moran 2003). We considered both

means and variation for many attributes (Table 1), as variation in conditions can be as important as average conditions in ecology (Palmer et al. 1997).

The majority of attributes were modelled as normally distributed continuous variables, after \log_e -transformation of the original data. Data for catchment area and perimeter were treated as covarying so that their relation (i.e. area:perimeter ratio), rather than catchment size, contributed to the classification. Lithology did not fit any standard distribution, and often contained many 0% values. To incorporate these data into the classification, we discretized the % cover data into 10 equally-sized bins (0–10%, 11–20% etc.) and treated the bins as multinomial data.

Estimation of measurement errors for the spatial data was not possible. The only data for which any statement of measurement error was available was the peaks of the elevation model, which had a root mean square error of 20 m (Anon. 2001). Because an error estimate is only available for peaks, it is not possible to calculate errors for attributes derived from fields of adjacent cells within the DEM. Issues such as averaging of errors at the catchment scale and spatial autocorrelation would further complicate matters. For the climatic data, there was no indication of the accuracy of the monthly or yearly estimates. These errors are likely to be greater than those for data derived from the DEM because the climatic surfaces are based on interpolations of fewer data. With these great difficulties in estimating uncertainty, the measurement error was set as 5% of the data range for all continuous attributes. Our early attempts to directly estimate errors lead us to believe that this is an overestimate for most of the attributes considered and is therefore conservative. We also conducted a second classification with the error set at 10% for all attributes. This addressed the question of whether the Bayesian classification can deliver useful results if measurement error of spatial data is greater than we believe it to be.

The different clustering results were compared using the Adjusted rand index (ARI; Hubert and Arabie 1985). The statistic is based on the relation of every pair of cases in the study, and whether these relations differ between two solutions. This avoids the need to specify one solution as "correct", and then assess how well the second solution reproduces the first. The index takes a value of 1 for perfect agreement between two clustering solutions, and a value of 0 if agreement is equal to that expected solely due to chance. Values of <0 are also possible. The ARI assumes a hard classification, and to calculate the statistic we assigned cases to their most probable classes.

We also investigated the effects of data order on the clustering result. We re-ordered the data randomly 4 times, and re-ran the original classification (i.e. 5% error for continuous variables). The most probable

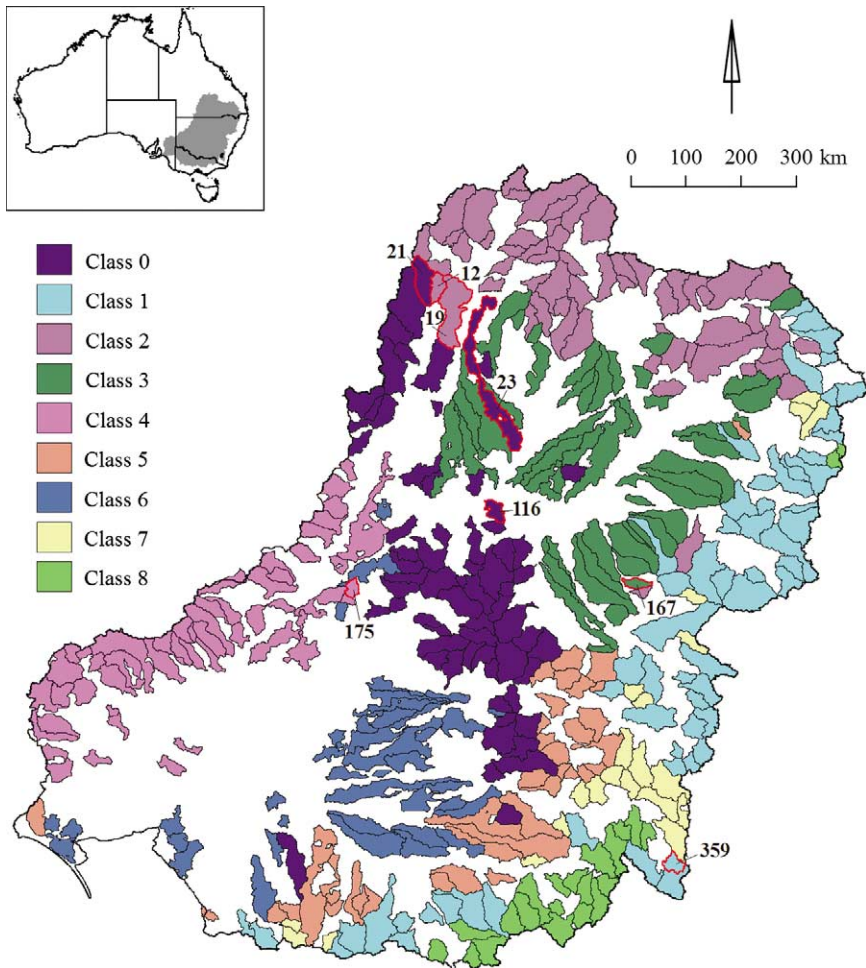


Fig. 1. Classes of order-4 catchments within the Murray Darling Basin. Catchment colour indicates most probable class number for M1, according to the key. Blank areas within the MDB are those that cannot be included in the definition of an order-4 catchment as described above. Numbered catchments are those for which the primary class membership probability was <0.80 (detailed in Table 4).

solutions for all 5 data orderings were compared using the ARI between pairs of results.

Last, we compared the results obtained from the Bayesian fuzzy method to those from a non-Bayesian hard clustering algorithm. We chose the Partitioning around medoids method (PAM; Kaufman and Rousseeuw 1990) implemented in the R package “Cluster” (Maechler 2007) for several reasons. First, as with the Bayesian method used in this paper, it is a partitional algorithm. Second, through the use of the “Daisy” distance matrix function (Maechler 2007), the method can operate on mixed data types (here continuous log-normal variables and multinomial variables). Last, like AutoClass, the software is freely available and free to use. The PAM clustering was performed on the original data set (i.e. prior to random re-ordering). The only difference in the data submitted to the two methods was

that for PAM area:perimeter ratios (of log-transformed areas and perimeters) were calculated prior to clustering. This was necessary because PAM does not have the capacity to treat variables as correlated. Being a deterministic hard clustering method, PAM does not require any statement of measurement uncertainty for continuous variables, but does require the user to specify the number of clusters a priori. This was set equal to the number in the most likely Bayesian solution to provide a direct comparison of the two results.

Case-study: results and discussion

Bayesian solutions

The most likely classification (M1) from AutoClass identified nine classes of catchment type within the

Table 1. Catchment attributes used in the classification.

Attribute
Area and perimeter*
Average slope
Stream segment length (mean)
Stream segment length (SD)
Stream density (i.e. km km ⁻²)
Stream confluence density
Annual rainfall (mean)
Annual rainfall (range)†
Annual actual evapotranspiration (mean)
Annual actual evapotranspiration (range) †
Annual potential evapotranspiration (mean)
Annual potential evapotranspiration (range) †
% coarse grained sediments
% fine grained sediments
% acid volcanic substrate
% basic volcanic substrate
% granite substrate
% limestone substrate
% water bodies

*Area and perimeter were entered separately, but are treated as correlated variables in the classification.

†Data for rainfall and evapotranspiration were supplied as monthly averages. Range is highest monthly value minus the lowest.

MDB at order-4 scale. The classes are shown in Fig. 1, with the colour indicating the most probable class for each catchment. The spatial association of the class members suggests that actual geographic gradients were identified. The classes reflect altitudinal and latitudinal gradients. Neither of these variables was included as an

attribute in the model (Table 1), although they are likely to be correlated with other included variables.

To describe how the classes differ from one another, we used the Kullback-Leibler distance for the various attributes in each class. A $D_{KL} > 1$ was used to define the point at which an attribute distribution was sufficiently different from the global distribution to warrant attention. This was an arbitrary, but useful, choice because it focussed attention on a relatively small number of attributes during the interpretation of results. Class-by-attribute combinations that met this criterion are shown in Table 2. We interpreted these values to describe the classes as shown in Table 3.

The classes describe groups of catchments with different physiographic profiles, with rainfall and slope being key distinguishing attributes (Table 2). These characteristics greatly affect the stream environment, and are therefore likely to moderate the effects of land-use change on stream ecological condition.

There are different degrees of divergence from the global distribution among classes. Class 8 exhibits the greatest class-level divergence with respect to the global class, while classes 0 and 5 have the lowest values (Table 2). However, divergence of a class from the global distribution does not necessarily imply class strength. A strong class will tend to have tight distributions of attribute values, better allowing the class to predict its own members (see definition of class strength above). Conversely, a class that is greatly divergent from the global distribution may have wide distributions, implying lower strength. For these results,

Table 2. Distinguishing attributes for the nine classes found in M1. Filled cells in the table are those for which $D_{KL} > 1$. For continuous attributes, the figure indicates the number of class-level standard deviations separating the class-level mean from the global mean, with “+” indicating that the class-level distribution is on average greater than the global distribution, while “-” implies the opposite. “0” indicates less than one class-level standard deviation separates the two means. For the discrete attributes, the class – “C” and global – “G” distributions are compared at a single point. “96% > 80%” indicates that 96% of catchments in the class had > 80% cover of that lithology. Symbols: μ = mean, σ = standard deviation. Attributes for which no D_{KL} was > 1 for any class are not included. The relative strength and class divergence (as measured by summed D_{KL}) for each class are also shown.

Class # →	0	1	2	3	4	5	6	7	8
Mean slope		+3					-3	+4	+7
Seg. length (μ)				+2					
Seg. length (σ)								-2	
Stream density							0		
Conf. dens.				-1					
Rainfall (μ)	+3				-3			+2	+5
Rainfall (range)			+2		-1				+5
AAET (μ)	+2				-3				+3
AAET (range)	+2				-1		-2		+3
APET (μ)									-1
APET (range)									-1
% coarse grained sediments			C: 96% > 80%						
			G: 30% > 80%						
% granite substrate									C: 86% > 10%
									G: 20% > 10%
Rel. class strength	0.0507	0.0005	1.0000	0.0037	0.0002	0.0022	0.0004	0.0017	0.0034
Class divergence	6.23	8.91	8.50	9.51	9.60	6.10	10.6	10.9	17.3

Table 3. Interpretation of the characteristics of the classes of landscapes generated by the most probable classification.

Class	Description
0	No attributes substantially different from the overall distribution of each attribute for the MDB.
1	Wet and hilly. High slope, rainfall and actual evapotranspiration.
2	Geology dominated by coarse-grained sediments, variable rainfall.
3	Long stream segments, few confluences. A relatively uniform landscape, but not necessarily flat.
4	Dry. Low rainfall and actual evapotranspiration.
5	No attributes substantially different from the overall distribution of each attribute for the MDB.
6	Flat. Low slope.
7	Similar to class 1, but with less evapotranspiration.
8	Wet, hilly and cold. High slope and rainfall with cold temperatures yielding relatively low potential evapotranspiration, but with high actual evapotranspiration relative to other classes because of high rainfall.

class 2 has the greatest strength, while class 4 has the least (Table 2). Classes 0 and 5 cannot be readily distinguished from the global distribution of attribute values by the criteria used here, but they are different to the other classes generated (Table 2). Moreover, the catchments of these classes were aggregated geographically (Fig. 1) and so represent sets of catchments that are more similar to each other than they are to the global set.

The colouring of classes in Fig. 1 does not represent the probabilities of class membership for the catchments, and we have previously argued that the probabilistic classification results are a major strength of the Bayesian mixture modelling approach. The vast majority of cases had a high probability of belonging to only one class (92% of cases had $p > 0.95$ for their most likely class). However, eight catchments had a highest class membership probability of < 0.80 (numbered in Fig. 1), and each of these catchments had a substantial probability of belonging to another class (Table 4). “Low probability” catchments lie on the edges of groups of catchments that belong to a single

class (Fig. 1). For all cases, except catchment 167, the second-most probable class for each catchment corresponds to the class of an adjacent group of catchments. We think that these kinds of catchments potentially are of great ecological interest because they may be transition areas in which many physiographic characteristics change over a short distance. Hence their identification is very important.

The posterior probability of the second most likely classification (M2) was about 1/20th that of the most likely classification – M1 – [$p(y|M1) = 6 \times 10^{-9170}$ vs $p(y|M2) = 3 \times 10^{-9171}$]. M2 identified eight classes within the MDB compared to the nine of M1. The ARI comparing the two classifications was 0.750. To put this figure in context, Steinley (2004) found that ARI values of 0.86, 0.77, 0.67, and 0.60 corresponded to the 95th, 90th, 85th and 80th percentiles of a distribution of 168 000 ARI values that compared randomly generated clustering results to “true” results. Thus, a value of 0.75 indicates close agreement of the two solutions (close to the 90th percentile). The map of most likely class memberships for M2 (not shown) was very similar to that for M1, with geographically distinct groups of catchments following altitude and latitude gradients.

The effect of increasing the nominal measurement error for continuous variables was greater than the difference between M1 and M2. The most likely classification using 10% measurement error – M3 – had a greater likelihood than the most likely classifications based on 5% error [$p(y|M3) = 2 \times 10^{-7848}$]. This results from reduced certainty of the data, because uncertain data can be more easily “shoe-horned” into classes. Increased uncertainty of continuous variables will also act to effectively increase the contribution of the categorical variables to the solution obtained. The M3 classification had fewer classes – five – than M1 or M2. The ARI comparing M3 to M1 was 0.503. This figure would place in approximately the top 30% of Steinley’s (2004) distribution, indicating a much poorer match, with the large difference in number of classes probably driving much of this result. The map of most likely class

Table 4. Intermediate cases for the M1 clustering. Table shows catchments that had a highest membership probability of $p < 0.80$, along with information on their second most likely class.

Case no.	Most likely class		Second most likely class	
	Class	Prob.	Class	Prob.
12	2	0.718	0	0.282
19	2	0.534	0	0.464
21	0	0.714	2	0.286
23	0	0.618	3	0.382
116	0	0.785	3	0.171
167	3	0.587	5	0.413
175	4	0.662	0	0.332
359	1	0.685	7	0.315

memberships (Fig. 2) shows that the five classes are geographically distinct. However, whereas M1 had several classes in the eastern part of the MDB, M3 has one. Greater data uncertainty effectively reduces the resolving power of the classification, which is to be expected given that more uncertainty implies a reduced chance that any two objects can be strongly differentiated from one another. Notwithstanding these differences, M3 still informs about the differentiation of catchments within the MDB. Thus, the classification may still be useful despite the lower precision assumed for the data, and the results indicate the robustness of the mixture modelling solutions.

The 5 different data orderings led to slightly different solutions, with 7–9 clusters being found. The average ARI between pairs of solutions was 0.814, indicating very minor differences. We did not investigate these results in more detail, but suspect most of the differences would lie with “uncertain” cases (i.e. those with a high probability of belonging to two or more classes).

Comparison with deterministic clustering

The non-Bayesian hard clustering result from the PAM algorithm produced a very different solution to those of the Bayesian method (Fig. 3). The classes of catchments produced by the PAM algorithm were generally not spatially cohesive, with the exception of classes C, D and G (Fig. 3), which corresponded partly to classes 2,

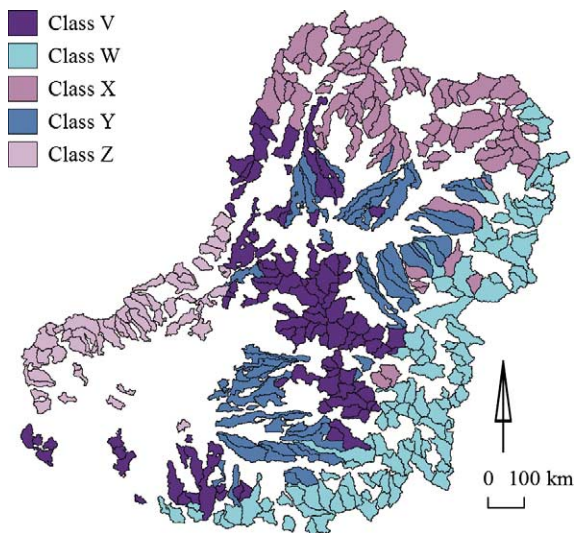


Fig. 2. Classes of order-4 catchments within the Murray Darling Basin when continuous data measurement error was specified as 10% of data range (M3). Catchment colour indicates most probable class, according to the key. Although the colours have been matched to those in Fig. 1, the classes are not equivalent.

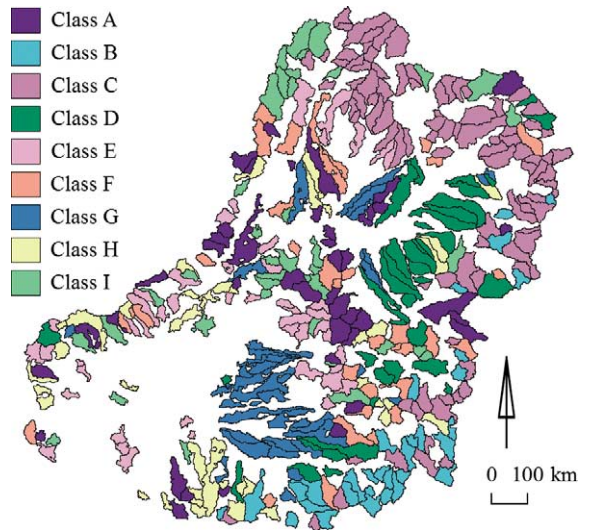


Fig. 3. Classes of order-4 catchments within the Murray Darling Basin as assigned by the PAM distance-based deterministic clustering method. Shading indicates most probable class number, according to the key. Although the colours have been matched to those in Fig. 1, the classes are not equivalent.

3 and 6 of M1 (Fig. 1). The ARI comparing the 9-class PAM solution to M1 was 0.212, indicating that agreement between the two solutions was only slightly greater than that expected by chance. Overall, the result is far less intuitively sensible than those produced by the Bayesian mixture model. Why the results should be so different is unclear. The dissimilarity matrix generated by the Daisy function uses Gower’s (1971) general dissimilarity coefficient so that mixed data types can be used. In this index, there is no assumption that attributes follow any type of distribution. Perhaps this lack of assumption in the method leads to an effective loss of information in the distance matrix. Moreover, the output from the hard clustering method is far less informative than that of the Bayesian mixture model. There is no “description” of the classes generated as a by-product of the clustering process. This would have to be manually produced from the original data set. There is no assessment of the most important attributes for each class, or for the solution as a whole. This would make it more difficult to refine the clustering in terms of the attributes used. Being deterministic, PAM provides no uncertainties, nor is there any presentation of alternative plausible solutions.

There have been few other comparisons of Bayesian mixture modelling to other clustering techniques. Okada et al. (2005) classified DNA microarray data using complete-linkage hierarchical clustering, and compared their results to other methods (including k-means and AutoClass). AutoClass did well based on a

measure of cluster strength, but less well on cluster divergence. Ter Braak et al. (2003) compared Bayesian mixture modelling to maximum likelihood-based mixture modelling for aquatic macrofauna, and found that the maximum likelihood approach suffered more from problems of locally optimal solutions than did the Bayesian alternative. Upal and Neufeld (1996) used artificial data to compare AutoClass to a neural network classifier and a Minimum message length (MML) approach. They found that the MML approach had most success in retrieving pre-determined clusters, followed by AutoClass and then the neural network. Bryan (2006) compared an ecological land-use classification based on AutoClass with a neural network approach (self organising map) and k-means clustering with homogeneity analysis. He preferred the latter approaches, with the self organising map better for visualising the structure of multivariate data and k-means useful for making decisions about the scale of the classification. Here, we have shown that Bayesian clustering can be applied to catchment-scale physiographic data. Moreover, the outputs are more informative, and the results more intuitively appealing, than those produced from an alternative non-Bayesian hard clustering method.

Further use of the Bayesian clustering

For the MDB example, the number of classes is relatively small, and the relationships among them (i.e. which classes are most similar/different to one another) can be readily determined. For larger data sets, the Bayesian mixture modelling approach is likely to produce a larger number of classes (Cheeseman and Stutz 1996). When the number of classes is too large for the information to be assimilated, it may be useful to perform a meta-classification of the data (i.e. classify the classes). In such an analysis, the class descriptions from the original classification are used as data for a new set of "cases", which are then classified. This approach is an indirect method of establishing a hierarchy of relationships within the original data set, and can help with interpretation of the original classification (Cheeseman and Stutz 1996).

The classifications produced here should not be considered as definite. Cheeseman and Stutz (1996) stress that clustering of complex data sets should be an iterative process between those performing the clustering and what they term "domain experts". A close examination of these results for a specific purpose may reveal too much heterogeneity in some classes, suggesting that additional discriminating data should have been included in the classification. Conversely, multiple classes may appear to be functionally similar, suggesting that unnecessary discrimination occurred. Therefore, it

is likely that some iteration of the classification would be necessary before a "final" scheme is realised.

Having derived our classification and with the stated uncertainties, what is its use? There are two applications that we see as being immediately useful, but which are too detailed to be presented here except by brief reference. The first is to undertake an exactly analogous classification of the MDB using land-use variables rather than physiographic ones to examine the congruence between land use and catchment physiography. For example, if catchments grouped together physiographically are scattered among several land-use classes, then one might conclude that at least some of the land-use activities may be unsuited to some of the catchments in which they are undertaken. Such information should be useful to regional land-management agencies and lead to important contributions to longer-term planning. The second application relates to the detection of human impacts on ecological condition in catchments. If one had catchments that are similar in physiography and land-use patterns (i.e. classified together), then one might focus on a much more limited set of drivers and proximal stressors of ecological deterioration than if considering an entire region. Moreover, lessons learnt in managing those deleterious effects in any one of those catchments should be transferable with great effectiveness to other catchments of that class.

Use of Bayesian clustering with ecological data

Bayesian clustering, and AutoClass in particular, have been little used in ecological applications. Ter Braak et al. (2003) used mixture modelling to perform Bayesian clustering of aquatic macrofaunal and environmental data, in an approach very similar to that taken by the AutoClass software. However, their purpose-written FORTRAN software is not available for general use. It also required the user to specify the number of classes a priori, and could not process large data sets. AutoClass software has been used in two studies of an ecological nature. Crook et al. (2002) classified behaviour-related display patterns in cuttlefish based on display components, and Bryan (2006) classified physical environments in a mountain range in southern Australia using five principal components based on temperature, soil, rainfall etc. However, we feel that these authors did not take full advantage of the Bayesian clustering results. Crook et al. (2002) did not report uncertainty in clustering solutions, consider the effects of different error levels, or explore alternate classifications. Bryan (2006) evaluated alternative classifications, and used different error levels to determine the number

of clusters produced, but criticised the requirement to provide uncertainty estimates for data, claiming that it limited the utility of the method.

We are unaware of any other approach that routinely presents a set of plausible classifications rather than a single “best” solution. These multiple solutions can have different numbers of classes and/or different arrangements of cases within the same number of classes. For each classification, much more information is provided than is normally the case for clustering algorithms. Other approaches generally do not provide information on the relative strength of the classes generated, and such information is essential when interpreting results. For example, a non-intuitive class may be partly explained by it being relatively weak compared to other classes. That mixture modelling automatically produces a description of the class through the distributions of its most important attributes also is valuable.

The MDB example showed that Bayesian mixture modelling can deal with those cases that do not fit well into only one class. The hard clustering with PAM assigned these cases to the most likely class without reporting the poor fit or the probabilities of assignment to other classes. Other fuzzy clustering systems (Bezdek 1981) also deal better with uncertain cases than hard clustering methods, but as stated above the membership values are not probabilities, and therefore are not manipulated as easily.

A major advantage of the Bayesian approach implemented in AutoClass is that results are presented along with all aspects of uncertainty from the cluster analysis. Ecological analyses are normally conducted with much uncertainty, and results can only be as certain as the data allow. Taking uncertainty explicitly into account and providing estimates of uncertainty in the results is a more defensible approach to the clustering of ecological data than common hard clustering methods. The Bayesian approach requires the user to consider this uncertainty, and therefore leads to interpretations that reflect the contingent nature of the resultant classifications.

Bryan (2006) commented that the requirement to specify measurement errors limits the usefulness of the method because estimates of measurement error often are not available for landscape variables commonly used in ecology. However, the lack of information on uncertainty is a problem of the data, not of the statistical approach. Statistical methods that assume no error in measured data, or have no way to propagate errors through analysis to guide inference, provide a flawed view of nature. Approaches that require us to quantify such uncertainty will lead to more realistic results. A definite statement of measurement errors is highly desirable, and should be reported. This is a much-neglected area, and remains a major challenge for spatial data sets, both at the level of individual cells in the spatial databases, and also the way in which

multiple cells and spatial autocorrelation affect average errors at larger scales. However, despite these problems, sensible classifications can be produced using reasonable estimates of error rates for landscape data (both this study and Bryan 2006).

As with other partitioning clustering methods, Bayesian mixture modelling has weaknesses such as convergence to sub-optimal solutions, and dependence of the solution on data order. However, the presentation of multiple plausible classifications, and the reporting of information on uncertainty for each classification is a better and more transparent approach to this currently intractable aspect of clustering. Here, we have shown that a number of random re-orderings of the data lead to solutions that differ in fine detail, but are highly similar overall. Other shortcomings of the method as implemented in AutoClass include the mandatory use of uninformative prior distributions, and a relatively small number of distribution types allowable in the mixture model. These issues, however, did not prevent us from identifying spatially cohesive and intuitively appealing clusters for the case study data set.

Conclusions

Many ecological and environmental applications choose to allocate objects to groups based on their level of similarity. While the groups are artificial given that the world is continuous, such classifications are necessary for management (e.g. bioregions) and our capacity to comprehend highly complex systems. We are concerned that many such classifications become “set in stone” and the degree to which a classification is valid for a particular purpose is rarely stated transparently. While some classification methods allow one to state how different classes are (e.g. based on distance or dissimilarity thresholds), the probabilistic results of Bayesian mixture modelling are much superior, especially in providing several alternative scenarios and case-specific probabilities of association with classes.

Acknowledgements – This work was funded under CRC for Freshwater Ecology project B260, and e-Water CRC project IP198. We thank Danny Spring and Sam Lake for their input to the project. We also thank John and Janet Stein (CRES) and Kristin Milton (MDBC) for assistance with the data sets, and David Keith for his review of the manuscript.

References

- Allan, J. D. 2004. Landscapes and riverscapes: the influence of land use on stream ecosystems. – *Annu. Rev. Ecol. Syst.* 35: 257–284.

- Anon. 2001. GEODATA 9 Second DEM version 2: user guide. – Geosciences Australia (Commonwealth of Australia), <http://www.ga.gov.au/image_cache/GA4783.pdf> .
- Anon. 2002. AutoClass C documentation. – North American Space Administration, Ames Research Centre, <<http://ic.arc.nasa.gov/ic/projects/bayes-group/autoclass/>> .
- Anon. 2003. Climatic atlas of Australia. – Bureau of Meteorology.
- Anon. 2004. ArcGIS ArcInfo 9.0 (SP3). – Environmental Systems Research Inst.
- Belbin, L. 1987. The use of non-hierarchical allocation methods for clustering large sets of data. – *Aust. Comp. J.* 19: 32–41.
- Bezdek, J. C. 1981. Pattern recognition with fuzzy objective function algorithms. – Plenum Press.
- Bezdek, J. C. et al. 1984. FCM: the fuzzy c-means clustering-algorithm. – *Comput. Geosci.* 10: 191–203.
- Bolliger, J. and Mladenoff, D. J. 2005. Quantifying spatial classification uncertainties of the historical Wisconsin landscape (USA). – *Ecography* 28: 141–156.
- Bryan, B. A. 2006. Synergistic techniques for better understanding and classifying the environmental structure of landscapes. – *Environ. Manage.* 37: 126–140.
- Bui, E. N. and Moran, C. J. 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray-Darling basin of Australia. – *Geoderma* 111: 21–44.
- Burnham, K. P. and Anderson, D. 1998. Model selection and inference. – Springer.
- Cheeseman, P. and Stutz, J. 1996. Bayesian classification (AutoClass): theory and results. – In: Fayyad, U. et al. (eds), *Advances in knowledge discovery and data mining*, AAAI Press and MIT Press, pp. 153–180.
- Cover, T. M. and Thomas, J. A. 2005. *Elements of information theory*. – Wiley.
- Crook, A. C. et al. 2002. Identifying the structure in cuttlefish visual signals. – *Phil. Trans. R. Soc. B* 357: 1617–1624.
- Gelman, A. et al. 2004. *Bayesian data analysis*. – Chapman and Hall/CRC.
- Gower, J. C. 1971. A general coefficient of similarity and some of its properties. – *Biometrics* 27: 857–874.
- Hanson, R. et al. 1991. Bayesian classification theory. – Artificial Intelligence Research Branch, NASA Ames Research Center.
- Hubert, L. and Arabie, P. 1985. Comparing partitions. – *J. Classification* 2: 193–218.
- Hutchinson, M. F. et al. 2000. Upgrade of the 9 second Australian digital elevation model. – Centre for Resource and Environmental Studies, Australian National Univ., <<http://cres.anu.edu.au/dem/>> .
- Jain, A. K. et al. 1999. Data clustering: a review. – *ACM Comput. Surv.* 31: 264–323.
- Kaufman, L. and Rousseeuw, P. J. 1990. *Finding groups in data: an introduction to cluster analysis*. – Wiley.
- Kurita, T. 1991. An efficient agglomerative clustering-algorithm using a heap. – *Pattern Recognition* 24: 205–209.
- Lakoff, G. 1987. *Women, fire, and dangerous things*. – Univ. of Chicago Press.
- Maechler, M. 2007. The cluster package. – <<http://www.r-project.org>> .
- McLachlan, G. J. and Peel, D. 2000. *Finite mixture models*. – Wiley.
- Okada, Y. et al. 2005. Knowledge-assisted recognition of cluster boundaries in gene expression data. – *Artif. Intell. Med.* 35: 171–183.
- Palmer, M. A. et al. 1997. Ecological heterogeneity in streams: why variance matters. – *J. North Am. Benthol. Soc.* 16: 189–202.
- Rosen, K. H. 2003. *Discrete mathematics and its applications*. – McGraw Hill.
- Steinley, D. 2004. Properties of the Hubert-Arabie adjusted rand index. – *Psychol. Methods* 9: 386–396.
- Strahler, A. N. 1957. Quantitative analysis of watershed geomorphology. – *Trans. Am. Geophys. Union* 8: 913–920.
- Ter Braak, C. J. F. et al. 2003. Bayesian model-based cluster analysis for predicting macrofaunal communities. – *Ecol. Modell.* 160: 235–248.
- Upal, M. A. and Neufeld, E. M. 1996. Comparison of unsupervised classifiers. – In: Dowe, D. L. et al. (eds), *Proc. of the Conference, ISIS '96, Information, Statistics and Induction in Science*. World Scientific, pp. 342–353.
- Wang, P.-C. and Leou, J.-J. 1993. New fuzzy hierarchical clustering algorithms. – *J. Inform. Sci. Eng.* 9: 461–489.