Electronic Journal of Applied Statistical Analysis
EJASA (2012), Electron. J. App. Stat. Anal., Vol. 5, Issue 1, 1 – 14
e-ISSN 2070-5948, DOI 10.1285/i20705948v5n1p1

A COMPARISON OF PLOTS FOR MONOTHETIC CLUSTERING, WITH APPLICATIONS TO MICROBIAL COMMUNITIES AND EDUCATIONAL TEST DEVELOPMENT

Mark C. Greenwood*

Department of Mathematical Sciences, Montana State University - Bozeman, USA

Received 25 August 2010; Accepted 08 July 2011
Available online 26 April 2012

Abstract: *Monothetic clustering for multivariate binary data provides a method of identifying variables whose levels identify the different patterns of responses. The default graphical summary of the standard version of this method provides information on the variables used to split the observations into groups and the proportions of the responses in the groups but not the levels of the responses in each cluster. A modified graph is proposed to provide fuller interpretations of the clustering results. The methods are applied to two examples and contrasted with other approaches to clustering multivariate binary responses. One application involves the clustering different bacterial clones presence/absence throughout distinct strata (or sections) in ice cores from Pony Lake, Antarctica, where the interest is in the full patterns of responses in the different clusters. The second is from a test developed to measure elementary education mathematics coaches knowledge of different aspects of mathematics coaching. In this second application, the interest lies more with questions that define distinct groups of individuals than with the entire pattern of responses.*

Keywords: *Monothetic clustering, banner plot, multivariate binary data.*

1. Introduction

Cluster analysis is concerned with finding patterns in data. The interpretation of the cluster solution involves consideration of the number of individuals in each group, the patterns of the responses in each group, possibly the strength of the cluster membership and possibly which individuals are in each group. Depending on the application, each of these considerations may be

* E-mail: greenwood@math.montana.edu

weighed as more or less important. Graphical methods can often be used to enhance the researcher's ability to understand the cluster solution, engaging as many of those aspects of the solution as possible to aid in the interpretation of the results.

Clustering methods are often classified as divisive (hierarchical) or agglomerative and polythetic or monothetic [13]. Divisive algorithms split the observations from one large group into smaller groups; agglomerative methods start with each individual in their own cluster and combine them into larger and larger groups. Polythetic methods divide or combine observations based on information from the entire multivariate response for the observations; monothetic methods split the observations based on the responses on a single variable. Monothetic methods can either use a local criterion for choosing the split [12,13] or target more general measures across all the variables [2]. Each split is, however, confined to the responses on a single variable to define the groups. The vast majority of clustering methods are polythetic [3,13]. Monothetic clustering of multivariate binary data was first discussed in [12] in the context of clustering in a plant ecology context. In [6], monothetic clustering as in [7,12] is made available under the name MONA. More recently, [11] have made these methods available and [8] provided them in the `mona` function in the `R` [9] package `cluster`. This seems to currently be the only software where these methods are available so the focus of this paper is exclusively on `R`-based analyses.

For clustering multivariate binary (presence/absence) data, polythetic approaches rely on defining a suitable dissimilarity measure between the responses based on the type of response and a clustering algorithm. Euclidean distance is one possible choice that measures the distance between the patterns of 0s and 1s of the responses. A simple matching coefficient measures the percentage of matches and is more often recommended for binary responses [6], but is just one of many measures of dissimilarity that could be applied to multivariate binary data. Given a dissimilarity measure, clustering proceeds with forming groups either agglomeratively or divisively depending on the clustering algorithm selected. Methods such as single, average, complete linkage, and Ward's method are the most commonly discussed agglomerative clustering algorithms [4,6]. For agglomerative methods, the clustering results are presented in a dendrogram and that hierarchical display is used to interpret the cluster solution.

Monothetic clustering defines both the clustering algorithm and dissimilarity measure simultaneously, using associations between variables to select individual binary variables to split the observations into groups hierarchically. The focus here is on methods for displaying the results from monothetic clustering of binary data, their applications, and comparisons to some other options. The graphical enhancement could be employed with other monothetic clustering methods, but the clustering algorithm from [6,7,12] is used here. After explaining the monothetic clustering algorithm, a new graphical display of the results is proposed, illustrated using a commonly available data set and compared to other graphical displays and clustering techniques. Then, two applications of the methods are provided demonstrating potential uses of the clustering algorithm and the features of the enhanced graphical display. Some features of the clustering algorithm are made more obvious by the exploration of choices in the graphical display.

2. Monothetic Clustering Algorithm

The monothetic clustering algorithm used here of [6,7,12] for k binary responses is based on sequentially selecting the variables that are maximally associated with the remaining variables. It was also originally called association analysis, with different measures of association that could be used in the method. The preferred method is based on the pair-wise measure of association for variables f and g of $A_{fg}=|a_{fg}d_{fg}-b_{fg}c_{fg}|$, where the values a_{fg}, b_{fg}, c_{fg}, and d_{fg} are obtained from the contingency table in Table 1.

Table 1. Contingency table used in the measure of association between variables *f* and *g*.

		Variable *g*	
		1	**0**
Variable *f*	**1**	a_{fg}	b_{fg}
	0	c_{fg}	c_{fg}

To identify the variable that is most related to the others, a total measure of association for each variable is calculated as $A_f=\sum_{g\neq f}A_{fg}$. The variable that is most highly related to the others, the one with the largest A_f, is then used to perform a binary split of the data set. Then each partition of the data set is explored in exactly the same fashion, always selecting the variable that is most related to the other variables for the remaining observations. It is possible with the binary responses to have two variables that are equally related to the remaining variables and in this case one of the two is randomly selected to use as the split. This creates a tree of splits of the data set, with at most 2^k-1 splits possible until the groups are completely enumerated. It is possible to require fewer splits if the responses are highly related. This also may occur if some patterns are not very likely or impossible, as seen below.

The algorithm can be applied to situations with missing observations, at least in a sense. Missing values are imputed in [6] prior to performing the cluster analysis. The same metric of association is used to identify the single variable that is most highly associated with the variable with the missing values. Based on that identified variable, the missing value is imputed based on the patterns of association. This could serve to enhance the apparent size of some clusters but is a reasonable method to fill in the missing values. In the results considered below, the demonstration data set from [8] and the education data set contained missing responses. The enhanced graphics proposed below do display the response categories for all observations and thus may display the imputed responses used by the clustering algorithm.

3. Enhanced graphical display

The data set provided to demonstrate monothetic clustering in the `cluster` package is based on clustering n=20 types of animals based on k=6 binary characteristics (warm blooded, can fly, vertebrate, endangered, live in groups, and have hair). Four of the responses had missing values either on whether they were endangered or lived in groups. To make a fair comparison to other methods, those missing values are first imputed using the monothetic clustering algorithm discussed previously. That "complete" data set is clustered using the percentage of mismatching $[1-(a_{fg}+d_{fg})/(a_{fg}+b_{fg}+c_{fg}+d_{fg})]$ and two different linkage methods, with the dendrograms of results displayed in Figure 1. These two solutions are built agglomeratively but are interpreted from the largest cluster down. These types of solutions provide no information about the pattern of

responses that led to particular individuals being grouped together, the researcher is left to revisit representatives from each group to interpret the cluster solution. If the response variables tend to be correlated, then it might be possible to simplify the interpretation of the main divisions in the animals based on the presence or absence of a certain key trait or traits, but polythetic clustering does not make this type of interpretation easy. With six binary variables, there are 64 different possible patterns of responses. In this data set, there are twenty observations and eleven different patterns observed. This suggests some relationships between presence/absence of the different traits and that the interpretation of the larger scale groups could be driven by and interpreted based on information from a subset of the measured traits.

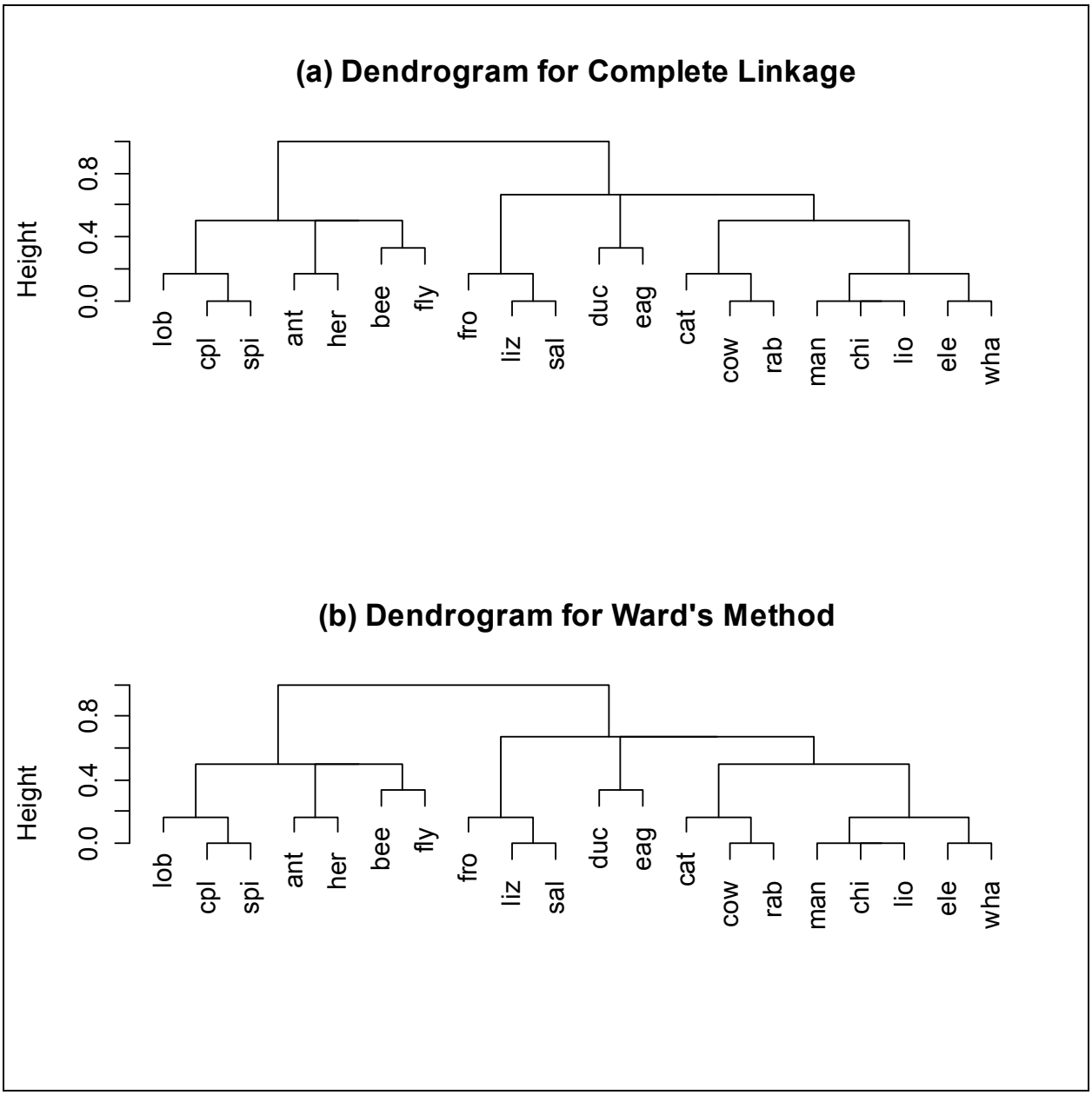

Figure 1. Polythetic clustering results for the animals data set using mismatching percent distance metric with (a) complete linkage and (b) Ward's method.

The monothetic clustering algorithm proceeds in a very different manner, attempting to find the important variables in defining the groups of observations. After imputing missing values, the clustering algorithm splits on warm-blooded or not and then splits within these two groups using the remaining variables, as seen in Figure 2a. The default summary of the results is using a type of banner or bar plot that displays horizontal bars for the observations as well as the variables used to make the splits within the graph.

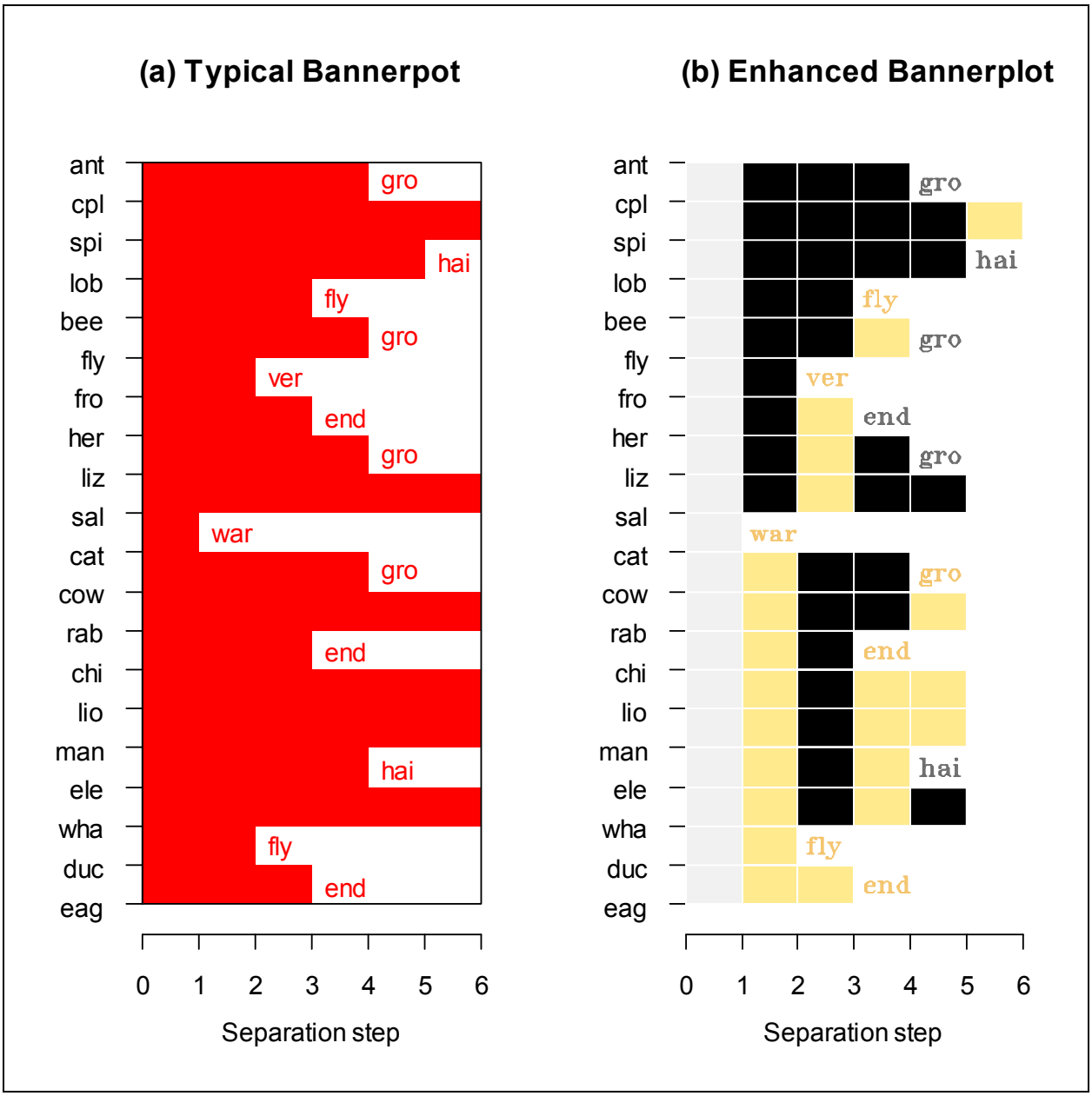

Figure 2. Monothetic clustering of "animals" data based on seven characteristics. Conventional graph in (a) and enhanced version in (b). In (b), black relates to absence of the trait, yellow to presence, and light grey bars to the data set before any splits have occurred. The y-axis denotes the different animal species included in the data set.

It provides a useful summary of the proportional sizes of the different groups and which variables were used to generate the different groups, but that is the limit of this graphical device. This graph appears to be a straightforward translation of a graph originally developed for the text-based graphics in [6], and is best when used with large data sets. It is possible to enhance this graph to incorporate the responses of the individuals in each group, to more clearly delineate the location of each response in the graph, and to eliminate implications of splits that are not occurring from the graph. It can be difficult to track a specific observation through the splits in the default graph, which is typically not done in large data sets and suggests the intended use of the graph. This contrasts with the figure in [6] where each line in a text-based graphic corresponds to an observation, which is problematic with even with moderate sized data sets. The result of plotting the cluster solution in `R` is displayed in Figure 2a.

To improve on the graph without fundamentally impacting its useful qualities, we propose four changes to enhance in the information available in the graph. First, adding borders around the bars provides the reader the ability to track the individual response across the graph and to assess the number of splits occurring for an observation, and to easily count the number of observations within a group if needed. It is important to note that individuals are actually displayed between the bars, with the bars displayed between two observations that have the same level at that level in the clustering. This idea informs the second and third modifications. This feature is turned off for large data sets because it is not possible, or likely of interest, to track individual observations in larger sample sizes.

The second modification is to color the bars relative to the response of the individual on the variable that is being used to split the observations. The baseline (transformed into a coding of 0 in the `mona` function) response is colored black and the positive response (coding of 1) is colored yellow; these colors are chosen to provide contrasting colors and hues in either black and white or color reproductions. Since the bars are being colored based on the responses on a particular variable, the initial column, which corresponds to a single, initial cluster that has not been split, is left as a light grey.

The third modification is a companion to the coloring of bars. The text color for the text for the variable names used in a particular split is also modified to denote the response of the individual at the bottom of the bar where that text is placed. This is consistent with the fact that the observation on the top of the graph is only represented with a line, not a bar, and then bars or text are entered depending on whether the next response down the column is the same or different from the observation. Coloring the text also clarifies the results in a situation where there are only two neighboring individuals in a group being split.

The final modification is to drop columns on the graph where the observations are no longer being split. This is possibly a contentious choice as it changes the length of the bars in the original graph that were used to convey the observations remaining in each cluster across additional variables. The enhanced graph now contains colors based on the responses in each column and it is unclear what color the bars should be made if no splits are occurring and multiple variables remain. If only one variable is left, then it is possible to include that column and color the bars accordingly. This sort of situation occurs when the initial variable splits are able to completely purify the clusters, with additional variables un-addressed in those clusters because the responses are all identical.

The resulting in graph in Figure 2b contains the enhancements discussed above. The `R` function to make this plot is available from the author. With this additional information available, the

clustering results for the different types of animals become much clearer, as the levels of variables accompany the splits. For example, the top half of the observations are cold-blooded and bottom half are warm-blooded. For the warm-blooded animals, only two fly, with the eagle being endangered and the duck not. That last bit of information is based on the text for "end" in the fourth column, bottom row, corresponding to the negative status. For the cold-blooded animals, slightly less than half are vertebrates. The interpretations could continue but are reserved for more interesting applications below.

The other graphical display of monothetic clustering that is used in [2], [6], and [12] involves displaying the tree of splits, presenting the categories for each branch in the hierarchy on the arrows for the different splits. It is also typical to display the sample sizes in each terminal node in the tree. This approach is very similar to another type of binary splitting algorithm, classification or regression trees [1]. Those use a suite of explanatory variables to create binary partitions in the responses to optimally explain the responses. In this case, the measure for selecting splits is driven by a single response variable in contrast to monothetic clustering where all the variables are treated equally in deciding to split the observations into groups. Tree graphs provide a useful summary of the partitions and some information on the number of observations having each complete pattern. However, it is difficult to visualize the number of observations involved in each split in the hierarchy, especially as a proportion of the total or a proportion of the subset being considered for a split. The graph proposed here attempts to retain the information from the banner plot version while incorporating the information on the categories of the responses used to make the splits.

4. Applications

Two different real applications of monothetic clustering are used to demonstrate the features of the plot and to highlight different uses of the results from monothetic cluster analysis. The first data set provides a small data set (n=36) with a small number of response variables (k=3) for microbial presence/absence in different levels of ice cores. The interest is in fully describing the patterns of each individual and describing what those patterns are, both in general for each larger cluster and for describing unusual responses. The second data set contains responses on a set of k=7 test questions for n=67 individuals, with five missing responses out of the total possible of 7*67=469. This is a subset of responses in the overall study to protect privacy of the study participant responses; the questions are renamed and not reported to allow repeated use of the instrument. The point here is to illustrate the potential benefits in exploring the patterns of responses on a set of test questions; it is not necessary to know the full details of the questions to explore the use of the technique. In this situation, the main interest is in the patterns of responses to the questions. Specifically, which questions provide the primary separation between the response patterns of the subjects? How are correct or incorrect responses related across all the questions? Is it possible to provide most of the discrimination between the individuals based on a subset of the questions? These questions are discussed further below.

4.1 Clustering presence/absence of bacterial clones identified in different sections of an ice core

First, an application to clustering presence/absence of different bacterial clones across three different levels of an ice core collected from Pony Lake located at Cape Royds (77° 33' S, 166° 00' E), Ross Island, Antarctica [4]. Except for a few weeks during mid-summer the lake is ice covered or frozen solid to its base. Eight ice core samples were obtained along two transects in the center of the lake and then divided into three sections (top, middle, and bottom). They were cleaned, melted, and pooled to create representative samples of ice melt water at each of the three levels of the lake ice. Using 16S rRNA gene sequencing, thirty-six different bacterial clones were identified across all the ice melt water. The presence/absence of the detected clones in the different ice core sections were the focus of the monothetic cluster analysis for the n=36 clones on the k=3 variables. It was of interest to summarize memberships of the different clones to the 2^3-1=7 different patterns of presence/absence across the three levels, summarizing the number of clones with similar patterns and being able to explore which clones had particular patterns across the levels. Only seven patterns were possible since the pattern of absence would correspond to the bacteria that were not measured in the gene sequencing.

Three dominant patterns were identified using the monothetic cluster analysis displayed in Figure 3. One group of clones was present in the top core and absent in the other levels, another was present in all three levels of the cores, and the largest group was absent in the top two cores. Clones could not be considered in the data set if they were not present in at least one level, so it is obvious that the clones that are absent in the top two cores must be present in the remaining variable not considered, the bottom level. In fact, in [4] the resulting presence of the clones in the bottom level was added to plot. In other studies, it may take further investigation of the data set to understand the results that are not present in the hierarchy of the cluster analysis. In this data set, the patterns observed are nearly completely enumerated based on the results for the top and middle cores, suggesting a strong relationship between information at those levels and the remaining level. The remaining patterns were relatively rare, with four clones present only in the middle core, one present in the middle and bottom cores, two present in the top and bottom cores, and one that only present in the top and middle cores. This suggests some separation in the clones that have been encapsulated at the top and bottom levels of ice in the lake, with few common strains at the different levels.

Because of the small sample size and interest in each bacterial strain, being able to explore the pattern for each clone was one goal of the analysis. A second goal was to understand the popularity of each cluster pattern produced by the microbial community composition. The typical banner plot only partially succeeds in the second desire and completely fails in the first. The enhanced banner plot succeeds in both cases for this data set. It also clears up one potential source of confusion in the banner plot related to the group of clones that is absent in the top two levels and present in the bottom section. Because the responses from the bottom ice core level are used for all the splits in the last column, it is easy to assume that the bottom responses were used to split in that cluster as well. The only aspect of the bottom ice core responses used in forming the cluster was that they all had the same value. While it takes further exploration of the data set to understand the results for non-displayed variables, it also avoids mis-interpretation of the clustering algorithm to use the modified display.

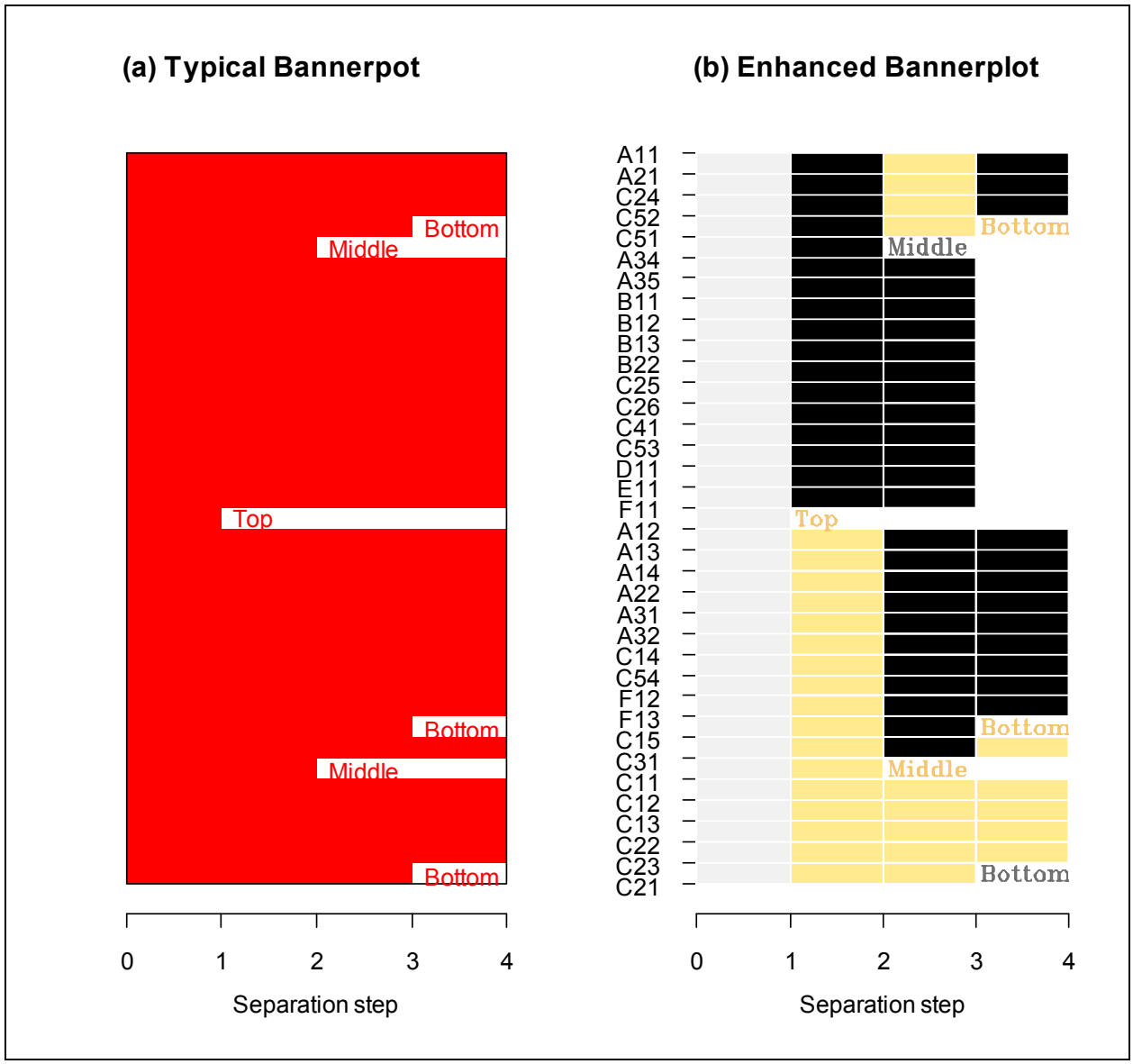

Figure 3. Monothetic clustering of bacterial clones identified in different ice core sections (top, middle, and bottom) from Pony Lake, Antarctica. Conventional graph in (a) and enhanced version in (b). In (b), black relates to absence of the clones at that level of ice, yellow to presence, and light grey bars to the data set before any splits have occurred. The y-axis denotes the different bacterial clones.

Dating back to the first paper on association analysis [12], tree-based diagrams have been considered for displaying the results of these analyse with convention of a capital letter representing presence for that trait (level of ice core in this application) and lower case for absence. An example of this type of display is provided in Figure 4, showing that information about the frequency of certain patterns and important variables is possible. However, the proportions of the total at any intermediate stages of the hierarchy are difficult to extract from the plot. In fact, rarely observed patterns are given the same treatment as extremely frequent patterns, except in the final line of the plot where a summary of counts at the terminal nodes is provided. This type of display was not attempted for situations with many response variables

because the exhaustive tree would become almost uninterpretable as it divided the data set into its potential 2^k different patterns. The sample size has no effect on this plot, but the number of different observed patterns does directly impact it.

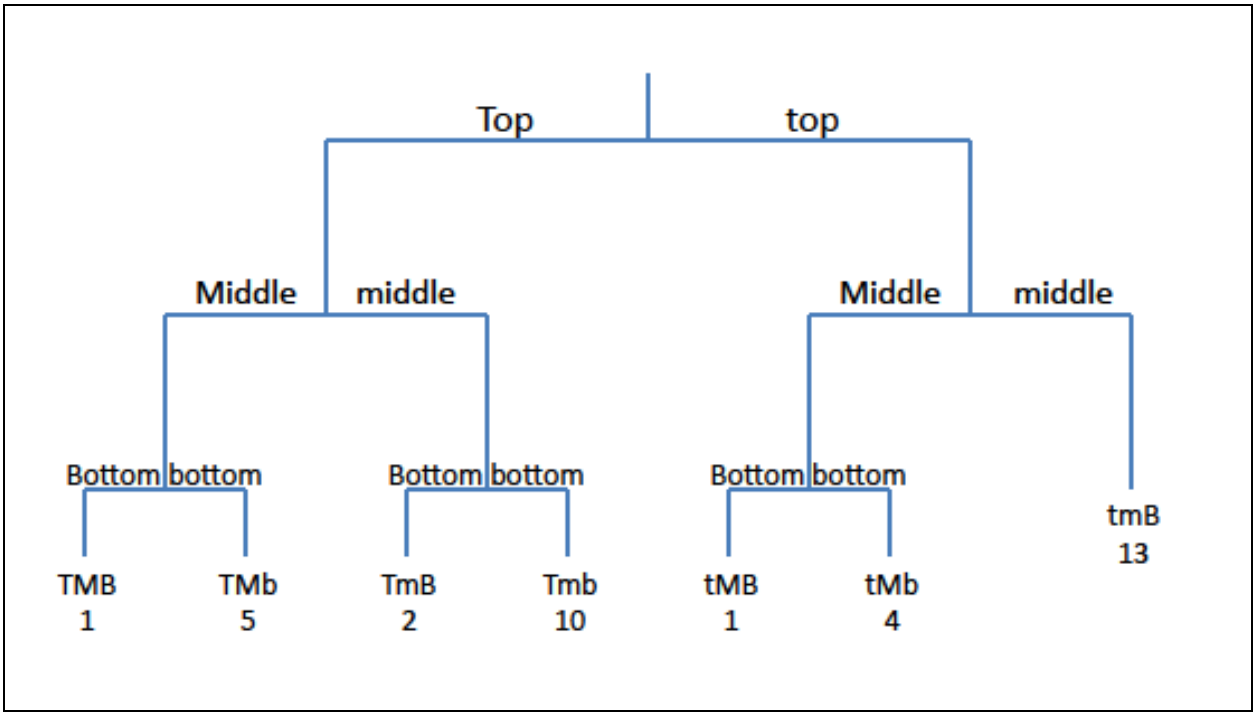

Figure 4. Tree-based display of monothetic clustering of bacterial clones. Capitalized labels indicate presence in ice layer, for example TmB suggests presence in the Top and Bottom layers but not in the middle. Counts are number of bacterial clones exhibiting each pattern of presence/absence through the ice core layers.

4.2 Clustering test responses in an educational testing situation

In a second application of monothetic clustering with slightly different goals, the results of a test being developed to measure elementary education mathematics coaches knowledge of coaching are considered. The development of this instrument is part of larger longitudinal study to evaluate the effects of mathematical and coaching knowledge on the teaching practices of teachers who are being coached in a K-8 environment. The preliminary results of the instrument are to be used to develop a training course to increase the knowledge of coaches on these items (as well as an additional suite of related questions). These were multiple choice questions and were graded as either correct or not. This subset of both responses and questions includes k=7 variables measured on n=67 subjects. Five missing values were encountered, two of those occurred for a single individual. The interests here are different from those in the previous application. There is no interest in every subject's responses, but in identifying common patterns of responses and the items that are most important in separating those patterns. To allow future use of the same instrument, the questions are not presented here. Figure 5 displays the results of the monothetic clustering of the data set. Because the data set was larger, the borders between observations and observation labels have been dropped.

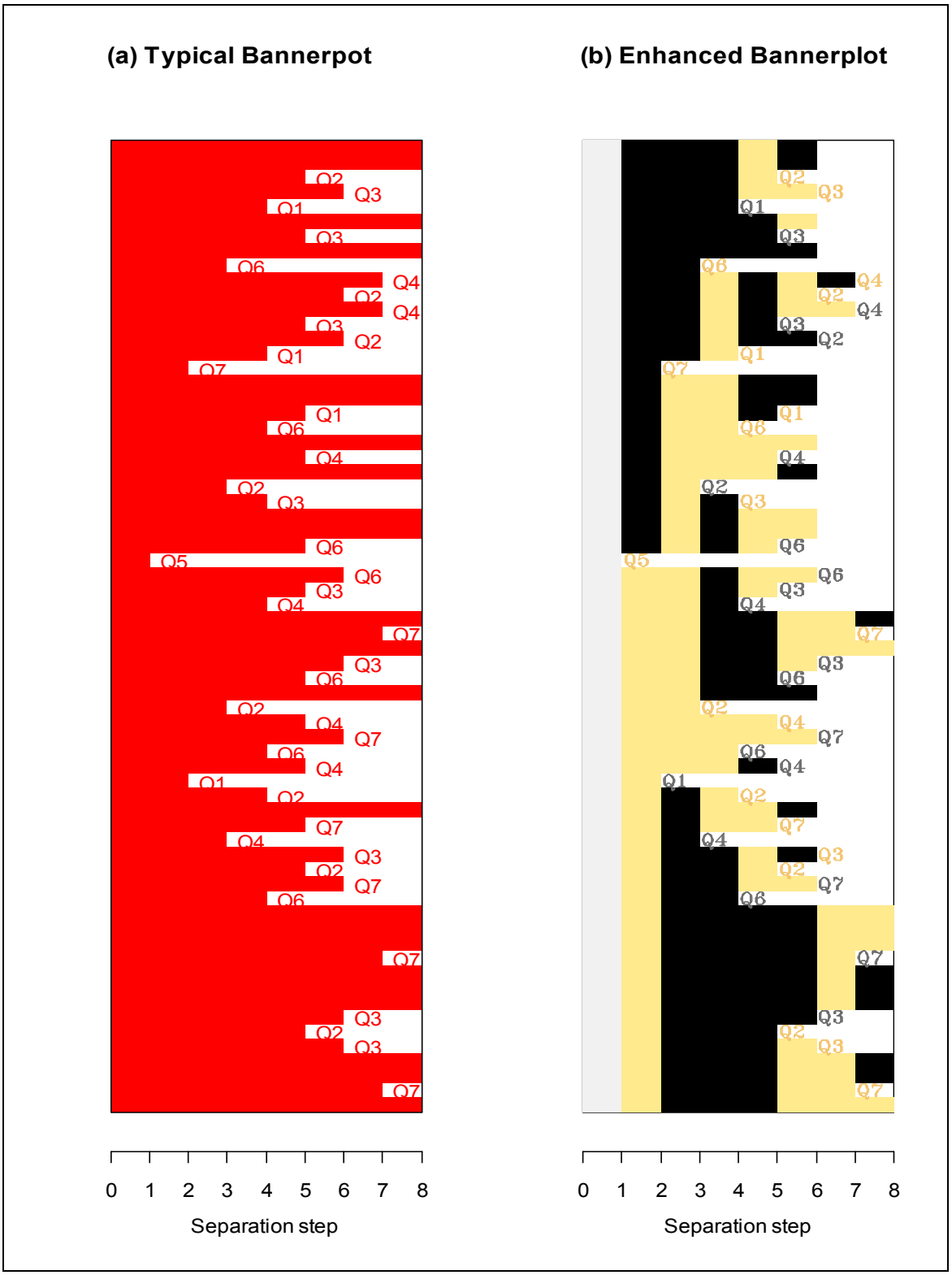

Figure 5. Monothetic clustering of educational testing data based on responses on seven questions (Q1, …, Q7). Conventional graph in (a) and enhanced version in (b). In (b), black relates to incorrect response, yellow to correct response, and light grey bars to the data set before any splits have occurred. The y-axis denotes the different subjects.

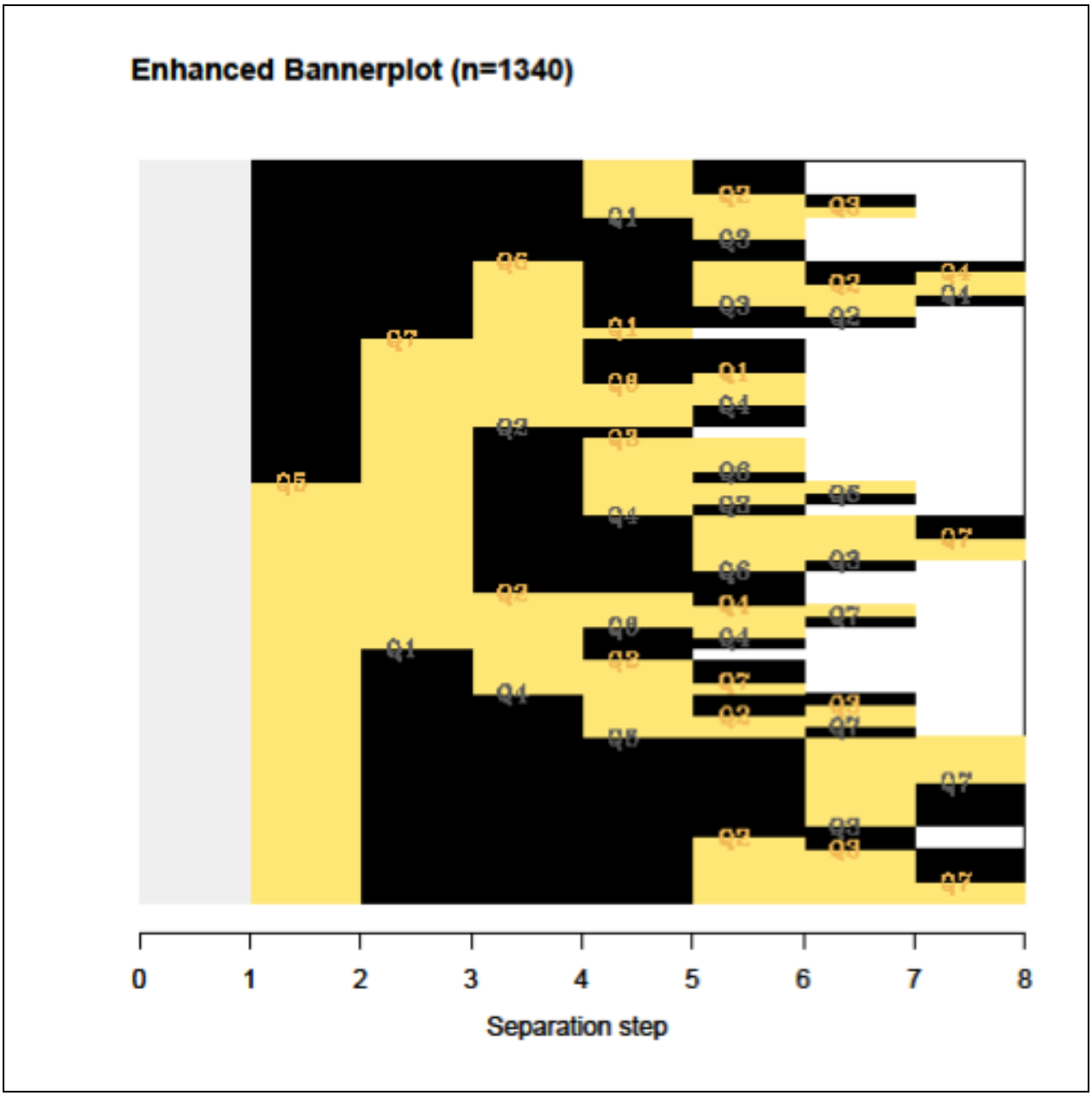

Figure 6. Monothetic clustering of educational testing data with sample size increased to n=1340. Performance of enhanced graphical display is un-effected by increases in sample size although the text labels do plot over the bars.

The initial split is based on Question 5 (Q5), with slightly more than half the respondents getting the question correct. For those that got Q5 correct, slightly less than half got Q1 correct. For those that missed Q1 in this group (at the bottom of the graph), most also missed Q4, Q5, and Q2. This suggests that there is a relatively large group of individuals who know the information on Q5 but may be struggling with the information in Q1, Q2, Q4, and Q6. The remaining questions provide further splits as these individuals get mostly correct responses on Q3 and about half get Q7 correct.

Explorations of the responses in the other main groups can provide similar interesting results. For the respondents that got Q5 and Q1 correct, they struggled with Q2 and Q4. One individual got all the questions correct, although Q3 was not used to separate him or her from the other subjects. Missing Q5 does not necessarily imply a poor performance on the test overall, as only two individuals missed all the questions. However, every subject that missed Q5 also missed at least one more question. These results, combined with information on the topics covered by each question, will prove invaluable in the development of materials to enhance the knowledge of coaches on these practices and which sets of topics combine to be easier or harder for groups of subjects.

In testing situations like this, it is common to consider item response theory [5] or Rasch models [10]. Those approaches allow one to assess the difficulty levels of different questions and to translate the original responses into underlying latent ability scores. The performance of subjects on different questions is implicitly of interest but explorations of connections between responses on the different questions are rarely performed. There is also little in the methods that can suggest if certain questions are redundant with others. A cluster analysis of the responses such as the one performed here can be a nice complement to a more formal test theory analysis. It could be used to refine test questions or, in our case, to inform sets of deficiencies of subjects to address.

This data set provides the framework for one additional exploration with a "large" data set. Although "large" is a matter of perspective, clustering twenty times (n=1340) the observations in the original data set is sufficiently large to see any impacts on the graph. Each observation is replicated twenty times to create the "data set" analyzed with monothetic clustering in Figure 6. The results are basically the same as those observed for the real version of the data set in Figure 5 except that the number of bars displayed increases. However, the proportions of the total and splits remain the same.

5. Discussion

Monothetic clustering for binary data is not a new method; it was initially developed prior to current graphical possibilities, especially those available in `R`. The proposed graph enhances the available information without significantly complicating the visual advantages of the standard approach. It was applied to three different situations, showing the different aspects of the cluster analysis that can be extracted from the graphical summary. The graph could be adapted for use in any monothetic clustering of binary data situation, such as when [2] is applied to data such as those discussed here.

Extending monothetic clustering to account for multi-category observations (nominal or ordinal) is possible using dummy or indicator function coding of the multi-level responses. The algorithm will only use j-1 of the categories of the j-level variable and may not use them in succession, making the interpretation slightly more complicated. But it does provide the possibility of an extension of the methods to this slightly more complicated clustering situation. Being able to visualize the levels that the dummy coded variables take on would be critical to understanding these more complicated results.

Acknowledgement

The author wishes to thank Drs. Markus Dieser and Christine Foreman for the initial inspiration to explore these methods, for the push to make the graphics more useful, and for providing the ice core data set. Drs. Elizabeth Burroughs and David Yopp made the application of the methods to educational testing possible and motivated the connections between these diverse areas of research. This research was funded under NSF Award No. 0918326. Any opinions expressed herein are those of the author and do not necessarily represent the views of the National Science Foundation. The author also wishes to thank two anonymous referees for the many useful suggestions which have led to very useful improvements to the manuscript.

References

[1]. Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth Advanced Books and Software.

[2]. Chavent, M. (1998). A monothetic clustering method. *Pattern Recognition Letters*, 19(11), 989-996.

[3]. Everitt, B. (1993). *Cluster analysis. Third edition*. London: Edward Arnold; copublished in the Americas by Halsted Press [John Wiley & Sons, Inc.].

[4]. Foreman, C., Dieser, M. Greenwood, M., Cory, R., Laybourn-Parry, J., Lisle, J., Jaros, C., Miller, P., Chin, Y., and McKnight, D. (2011). When a habitat freezes solid: microbes over-winter within the ice column of a coastal Antarctic lake. *FEMS Mircobiology Ecology*, 76(3), 401-412.

[5]. Hambleton, R., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.

[6]. Kaufman, L. and Rousseeuw, P. (1990). *Finding groups in data. An introduction to cluster analysis*. New York: John Wiley & Sons, Inc.

[7]. Lance, G. and Williams, W. (1965). Computer programs for monothetic classification ("Association analysis"). *The Computer Journal*, 8(3), 246-249.

[8]. Maechler, M., Rousseeuw, P., Struyf, A., and Hubert, M. (2005). *Cluster Analysis Basics and Extensions*, unpublished.

[9]. R Development Core Team (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.

[10]. Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

[11]. Struyf, A., Hubert, M., and Rousseeuw, P. (1997). Integrating robust clustering techniques in S-PLUS. *Computational Statistics & Data Analysis*, 26(1), 17-37.

[12]. Williams, W. and Lambert, J. (1959). Multivariate Methods in Plant Ecology: I. Association-Analysis in Plant Communities. *Journal of Ecology*, 47(1), 83-101.

[13]. Xu, R. and Wunsch, D., II (2005). Survey of clustering algorithms. *Neural Networks*, IEEE Transactions on, 16(3), 645-678.