A COMPARISON OF HIERARCHICAL METHODS FOR CLUSTERING FUNCTIONAL DATA

Laura Ferreira and David B. Hitchcock
Department of Statistics
University of South Carolina
Columbia, South Carolina 29208
hitchcock@stat.sc.edu

Key Words: cluster analysis; cluster sizes; misspecification; periodic functions.

ABSTRACT

Functional data analysis (FDA) — the analysis of data that can be considered a set of observed continuous functions — is an increasingly common class of statistical analysis. One of the most widely used FDA methods is the cluster analysis of functional data; however, little work has been done to compare the performance of clustering methods on functional data. In this paper a simulation study compares the performance of four major hierarchical methods for clustering functional data. The simulated data varied in three ways: the nature of the signal functions (periodic, non-periodic, or mixed), the amount of noise added to the signal functions, and the pattern of the true cluster sizes. The Rand index was used to compare the performance of each clustering method. As a secondary goal, clustering methods were also compared when the number of clusters has been misspecified. To illustrate the results, a real set of functional data was clustered where the true clustering structure is believed to be known. Comparing the clustering methods for the real data set confirmed the findings of the simulation. This study yields concrete suggestions to future researchers to determine the best method for clustering their functional data.

1. INTRODUCTION

While an extensive amount of work has been done to compare various clustering algorithms, relatively little has been done in exploring how such methods perform in clustering

functional data, that is, data that arise as curves. In recent years, the analysis of such functional data has become increasingly common. A modern example occurs in proteomics studies when a mass spectrometer records macromolecular expression values almost continuously across a domain of mass/charge ratio values.

In particular, many scientists are using hierarchical cluster analysis to make sense of the huge mass of data available in certain fields, genetics currently being among the most well known of these (e.g., Eisen et al., 1998; Morley et al., 2004). Although a number of statistical researchers have developed cluster analysis methods suited specifically for functional data (e.g., Serban and Wasserman, 2005; Wakefield, Zhou and Self, 2003), scientists often prefer to cluster data using traditional methods, especially hierarchical cluster analysis. This may be because the classical approaches are well known in their fields and convenient software exists to carry out the traditional analyses. This paper reports the results of simulation studies to compare four methods of hierarchical cluster analysis when the data are functional.

1.1 Cluster Analysis

Cluster analysis involves sorting data objects (or items) into natural groupings based on similarity. Grouping data is important because it can reveal information about the data such as outliers, dimensionality, or previously unnoticed interesting relationships. In cluster analysis there is **usually** no prior specification about the number or nature of the groups to which the objects will be assigned. The grouping is often done based solely on similarity measures, **and the ideal number of groups is determined (often in an ad hoc manner) within the overall process of clustering, but possibly quite separately from the clustering algorithm itself**. These characteristics can make cluster analysis difficult, and a wide variety of algorithms have been proposed to produce the "best" clustering of objects based on a set of observed data.

Many commonly used clustering algorithms rely on the pairwise dissimilarities between the objects to guide the clustering. There are many different ways of defining (dis)similarity among objects, and the choice of dissimilarity measure depends largely on the type of data one is working with (discrete, continuous, binary, etc.). Usually when items are clustered,

the dissimilarity between any two items is indicated by some sort of distance. Some common measurements of distance between two multivariate data vectors include ordinary Euclidean distance or Mahalanobis distance. With functional data, a more appropriate dissimilarity measure for two functions $y_i(t)$ and $y_j(t)$ measured on some domain, say, $[0, T]$, is the squared L_2 distance between the two curves:

$$d(i, j) = \int_0^T [y_i(t) - y_j(t)]^2 \, dt.$$

An important class of clustering methods is hierarchical cluster analyses. There are two main types of hierarchical clustering methods, agglomerative and divisive. This study will focus on agglomerative methods. An agglomerative hierarchical method begins with each object as its own cluster. It then successively merges the most similar clusters together until the entire set of data becomes one group.

In order to determine which groups should be merged in agglomerative hierarchical clustering, various linkage methods can be used. Single linkage (Sneath, 1957) merges groups based on the minimum distance between two objects in two groups; therefore the distance between clusters R and Q is defined by

$$d_S(R, Q) = \min_{i \in R, j \in Q} d(i, j),$$

where $d(i, j)$ is the distance between the ith and jth objects.

Complete linkage (McQuitty, 1960; Sokal and Sneath, 1963) merges groups based on the maximum distance between two objects in two groups. In other words, the distance between clusters R and Q is defined as

$$d_C(R, Q) = \max_{i \in R, j \in Q} d(i, j).$$

Average linkage (Sokal and Michener, 1958) merges groups based on the average distance of all the objects in one group to all the objects in the other. The distance in average linkage is defined as

$$d_A(R, Q) = \frac{1}{|R||Q|} \sum_{i \in R, j \in Q} d(i, j),$$

where $|R|$ = the number of objects in cluster R.

Another important hierarchical clustering method is **that** of Ward (1963). While Ward's method is similar to the linkage methods in that it begins with N clusters, each containing one object, it differs in that it does not use cluster distances to group objects. Instead, the total within-cluster sum of squares (SSE) is computed to determine the next two groups merged at each step of the algorithm. The error sum of squares (SSE) is defined (for multivariate data) as:

$$SSE = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})' (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})$$

where $\mathbf{y}_{ij}$ is the jth object in the ith cluster and n_i is the number of objects in the ith cluster.

There are certainly other classes of clustering methods besides hierarchical methods. Some of these include partitioning methods such as the well-known k-means (MacQueen, 1967) and k-medoids (Kaufman and Rousseeuw, 1987) algorithms, as well as model-based clustering methods (Banfield and Raftery, 1993). However, based on a survey of actual cluster analyses in the scientific literature, Kettenring (2006) indicated that hierarchical clustering was by far the most widely used form of clustering in practice. In this article, we will compare the performances of only hierarchical methods.

1.2 Clustering Functional Data

Functional data consist of observations that are intrinsically continuous functions, with the response measured over some domain such as time or space. Typically we have a single functional observation (in practice, observed at discrete measurement points) for each individual. Ramsay and Silverman (2005) provide a comprehensive introduction to functional data analysis. Typically, each discretely observed functional datum would be converted to a continuous functional observation via a smoothing method. In this article, we will apply a B-spline smoother to each functional object. **As a basic strategy, we choose to apply a relatively modest amount of smoothing to the observed functional data, as the goal was simply to remove noise and not to alter significantly the structure of the curves before clustering. While functional data clustering results can depend**

on the amount and nature of smoothing, as discussed by, e.g., Hitchcock, Booth and Casella (2007), we adopt a strategy of relatively conservative smoothing as our main purpose is to compare the clustering algorithms.

Once functional data are collected and smoothed, a method such as cluster analysis can identify patterns of variation both within and between clusters. By exploring which observations are in each cluster, we may be able to explain the source of the variation we see in the data by looking for common characteristics. By identifying these characteristics as potential explanatory factors, we can then collect new data with such factors in mind or use either cluster analysis or supervised classification on future data to look for similar patterns.

In this study, the squared L_2 distance between two curves was used as the basic dissimilarity measure. Since the data in practice consisted of discrete values representing measurements along continuous curves, the squared L_2 distance had to be approximated using a trapezoidal-rule approximation, where for a function $h(t)$ (which is $[y_i(t) - y_j(t)]^2$ in our case), the definite integral along the domain $[0, T]$ is approximately:

$$I_n = \frac{T-0}{2n}[h(0) + 2h(t_1) + \cdots + 2h(t_n) + h(T)],$$

where n is the number of measurement points used in the approximation. **Note that this I_n tends to the squared L_2 distance between $y_i(t)$ and $y_j(t)$ as the number of measurement points $n \to \infty$ within the fixed domain $[0, T]$ (in the sense of what Stein (1995) calls fixed-domain asymptotics).**

1.3 Previous Comparative Work

We now briefly review some previous work that has compared the performances of clustering algorithms on data types other than functional data.

Hands and Everitt (1987) examined five hierarchical clustering techniques on multivariate binary data, comparing their abilities to recover the original clustering structure. This paper will make similar comparisons for functional data. They controlled various factors including the number of groups, number of variables, proportion of observations in each group, and group-membership probabilities. All of the clustering techniques in the study were hierarchical: single linkage, complete linkage, group average, centroid, and Ward's method.

The simple matching coefficient, s_{ij}, was used as a similarity measure. They used two indices of fit (roughly Euclidean distances between the true membership vector and the clustering vector produced by the algorithm) as criteria to compare the algorithms. According to Hands and Everitt, most of the clustering methods performed similarly, except single linkage, which performed poorly. Ward's method did better overall than other hierarchical methods, especially when the group proportions were approximately equal.

Kuiper and Fisher (1975) compared six hierarchical clustering procedures (single linkage, complete linkage, median, average linkage, centroid, and Ward's method) for multivariate normal data, assuming that the true number of clusters was known. They first compared the ability of the clustering methods to recover two different samples from bivariate distributions with means $(0, 0)$ and $(d, 0)$ where $d > 0$. The second comparison involved a fixed number of clusters generated from multivariate populations. The authors used the Rand index, which gives a proportion of correct groupings, to compare the clustering methods. This index was defined as follows by Rand (1971):

$$Rand = \frac{N_{00} + N_{11}}{N_{00} + N_{01} + N_{10} + N_{11}}$$

where N_{00} counts the pairs of objects both placed correctly in different groups by a clustering algorithm, N_{01} counts the pairs of objects that are not placed in the same group but should be, N_{10} counts the pairs of objects that should be in the same group but are not clustered as such, and N_{11} counts the pairs of objects placed (correctly) in the same group. The Rand index serves as a measure of concordance between the true clustering structure and the output produced by a particular clustering algorithm. The Rand index will also be used in this study to evaluate the various clustering algorithms with functional data.

Kuiper and Fisher (1975) found that single linkage may only work well for long chain-type clusters as opposed to bunched clusters. Ward's method seemed to work well for equal sample sizes, but does not work as well for unequal sample sizes from bivariate data. The centroid and average linkage methods were quite similar, while complete linkage was most similar to Ward's method. In general, classification was better as the original number of clusters was increased. For clusters of equal sizes, Ward's method and complete linkage

worked best. With very unequal cluster sizes, centroid and average linkage worked the best.

Blashfield (1976) attempted to compare four types of hierarchical clustering methods (single linkage, complete linkage, average linkage, and Ward's method) for accuracy in recovery of original population clusters. Blashfield simulated many data sets, which were all mixtures from different populations. Each subpopulation had a randomly chosen mean and variance. Once the populations for a mixture were determined, the number of populations in the mixture and the number of variables sampled from each population were randomly selected. He used Cohen's κ statistic (Cohen, 1960) to measure the accuracy of the clustering methods. Blashfield compared the clustering methods for 50 distinct simulated mixed data sets. Ward's method performed significantly better than the other clustering procedures; the second best was complete linkage; average linkage gave relatively poor results.

Johnson and Wichern (2002) informally contrast various hierarchical clustering methods, without specifying a particular data type (the implication is that they are considering multivariate continuous data). For instance, they note single linkage tends to perform poorly with clusters that are truly elliptical. This problem occurs especially when clusters are elliptical when plotted, but the ends of the clusters are geometrically close. This causes certain points to be clustered together early in a hierarchical algorithm when in fact they should be placed with the rest of their own ellipse.

Tarpey (2007) compared several clustering methods for functional data, but in a quite different manner than this paper. He focused on k-means clustering and examined the effect on the clustering outcomes based on how the observed data were smoothed (using the raw data, and data smoothed via a B-spline basis, Fourier basis, and power basis, respectively). He concluded that the results of clustering functional data depend on how well the smooth curves fit the raw data, but that the choice of best smoothing method depends on the true mean curve of each cluster.

In addition, Schwaiger and Rix (2005) used a simulation study to compare two-mode hierarchical clustering methods, which attempt to simultaneously cluster the objects and the variables of a data set, but such cluster analyses are not considered in this article. Edelbrock

(1979) examined the performance of clustering algorithms when the hierarchical dendrogram is cut at several places. Milligan (1980) noticed that complete linkage and Ward's method reacted badly when outliers were introduced into the simulated data. Hubert (1974) and Baker (1974) separately compared single-linkage and complete-linkage clustering for several data types and structures. Cunningham and Ogilvie (1972) compared seven hierarchical methods based on the association between the input dissimilarity values and corresponding distance values obtained from the final clustering hierarchy.

An important conclusion arising from most of these studies was that the superiority of one method over others was not uniform, but rather depended greatly on the form of the data. We will establish analogous conclusions in the case of clustering functional data.

2. SIMULATION STUDY

2.1 Setup

This study was designed to test the performance of four hierarchical clustering algorithms on functional data: single linkage, complete linkage, average linkage, and Ward's method. The data simulated varied in three ways: the groups of signal functions being used, the standard deviation of the error added to the signal functions, and the number of objects in each true cluster.

There were three main groups of signal functions used in this study. The functions in each group were chosen to lie reasonably close to one another so that with the addition of error to the functions, it would not be trivial for the clustering algorithms to partition the resulting curves. The first group of signal functions involved some form of periodic data. These functions are plotted in Figure 1 (top) and defined as follows:

$$
\begin{aligned}
\mu_1(t) &= (1/28)t + \exp(-t) + (1/5)\sin(t/3) + 0.5, t \in [0, 100] \\
\mu_2(t) &= (1/20)t + \exp(-t) + (1/5)\sin(t/2), t \in [0, 100] \\
\mu_3(t) &= (1/15)t + \exp(-t) + (1/5)\cos(t/2) - 1, t \in [0, 100] \\
\mu_4(t) &= (1/18)t + \exp(-t) + (1/5)\cos(t/2), t \in [0, 100]
\end{aligned}
$$

[Figure 1 here]

The second group of signal functions had no periodic tendencies and were strictly decreasing. These functions are plotted in Figure 1 (middle) and are defined as follows:

$$\begin{aligned}
\mu_1(t) &= 50 - t^2/500 - 7\ln(t), t \in (0, 100] \\
\mu_2(t) &= 50 - t^2/500 - 5\ln(t), t \in (0, 100] \\
\mu_3(t) &= 50 - t^2/750 - 7\ln(t), t \in (0, 100] \\
\mu_4(t) &= 50 - t^2/250 - 4\ln(t), t \in (0, 100]
\end{aligned}$$

The third group of signal functions, all of which had a decreasing trend, had a mix of periodic and strictly decreasing functions. These functions are plotted in Figure 1 (bottom) and are defined as follows:

$$\begin{aligned}
\mu_1(t) &= -t/2 + 2\sin(t/5), t \in (0, 100] \\
\mu_2(t) &= -t/2 + 2\cos(t/3), t \in (0, 100] \\
\mu_3(t) &= -t^2/250 - 4\ln(t), t \in (0, 100] \\
\mu_4(t) &= -t^2/250 - 2\ln(t), t \in (0, 100]
\end{aligned}$$

For each simulated data set, 40 discretized curves were generated based on the four signal functions from the specified group. In general, the data were simulated over 201 (or 200) points from $t = 0$ (or $t = 0.5$ for clusters containing the $\ln(t)$ function) to $t = 100$ in increments of 0.5.

We added random error to the signal functions using a discretized approximation of the stationary Ornstein-Uhlenbeck process (a zero-mean Gaussian process in which the covariance between errors at measurement points t_l and t_m is $\sigma^2(2\beta)^{-1}\exp\{-\beta|t_l - t_m|\}$), with varying σ^2 and $\beta = 1$. The Ornstein-Uhlenbeck error process results in an autoregressive covariance structure for the equally-spaced discretized data in the simulation. Increasing the value of σ essentially added more random noise to the signal functions. The starting value

of the pattern of σ values explored for each group was chosen to give the various clustering methods a "difficult" test. This value tended to be smaller for the periodic functions and larger for the non-periodic functions. The value of σ was increased by the same increment (0.5) across simulation settings. **We consider here only the case in which the noise magnitude (as measured by σ) is the same for curves in different clusters.** The values of σ explored for each group are listed in Table 1.

Table 1: Values of σ Used for Each Group of functions

Group	Values of σ			
1	1.5	2.0	2.5	3.0
2	4.0	4.5	5.0	5.5
3	3.5	4.0	4.5	5.0

The final setting that varied in each simulation was the set of the simulated cluster sizes. Every simulated data set contained a total of 40 observed functions; however, the allocation of these 40 observations to the true signal functions was varied. There were five different cluster size options which represented potential data scenarios. The first scenario was equal cluster sizes, having 10 objects generated from each of the four signal functions. The next scenario represented three equal-sized clusters and one smaller cluster (12 objects from three of the signal functions and only four were generated from the remaining signal function). The third scenario included two relatively larger clusters and two smaller outlying clusters (18 objects from each of two of the signal functions and 2 observations from each of the remaining signal functions). The fourth scenario had one large cluster and three very small outlying clusters (35 objects from one signal function, 2 objects from two other signal functions, and 1 from the last signal function). The final scenario had four different cluster sizes: one large, two medium, and one small (20, 10, 8, and 2 observations for the different signal functions). In order to eliminate bias, particularly in situations with unequal cluster sizes, the size for the cluster corresponding to each signal function was randomly chosen for each simulation of new data.

For each simulated data set, the functional observations were smoothed using B-splines with 30 knots. **Visual inspection on a number of examples of this nature revealed the choice of 30 knots to provide a reasonable amount of smoothing across all values of σ considered here**, allowing a consistent smoothing method to be used over the entire study. The approximate squared L_2 pairwise distances between the smoothed simulated curves were used as dissimilarities input into each clustering algorithm. The clustering algorithms were implemented using the `agnes` function in the `cluster` package of R (R Development Core Team, 2009). For the main portion of this study, the **resulting dendrogram** was cut off at the correct number of clusters ($k = 4$). **(We note that cutting the dendrogram to obtain a clustering partition, while very commonly done, is not without risk. A simple cut may ignore the overall hierarchical clustering process, especially when some clusters have very few objects. We do not necessarily recommend a simple dendrogram cut for a single cluster analysis, but in a simulation study in which hundreds of cluster analyses are done at once, this seems the only practical means of judging the results.)** Section 2.3 discusses a secondary simulation study in which the number of clusters was misspecified ($k = 3$ and $k = 5$).

Once the clustering algorithm was stopped at the specified number of clusters, the Rand (1971) index was calculated for each clustering method, to indicate the performance of the clustering method. For each combination of signal-function group, cluster-size structure, and σ value, 1000 data sets were generated. For each simulated data set, the Rand index for each clustering method was calculated and stored. Based on these 1000 repetitions, the mean Rand index was found for each clustering method. The Monte Carlo standard error for this mean Rand index was also calculated. The mean Rand value was the criterion used to compare the clustering methods.

We note that the Rand index is merely one of a number of possible criteria for measuring the accuracy of a clustering partition. Others include the κ statistic of Cohen (1960) and a version of the Rand index (Hubert and Arabie, 1985)

adjusted for agreement due to chance; Warrens (2008) discussed an equivalence between these latter two measures. Several other approaches are mentioned in Section 1.3. For the purpose of comparing algorithms, any of the standard measures should be sufficient, and we choose the Rand (1971) index for its simplicity and for consistency with the study of Kuiper and Fisher (1975).

2.2 Results

Tables 2 through 4 (presented in the Appendix) give the exact mean Rand values for the four clustering methods at each of the settings we explored. For each mean Rand value, an associated Monte Carlo standard error (MCSE) was calculated; to save space, we simply report the maximum MCSE for each section of the tables. **The MCSE measures the variability of the Rand index across simulation iterations; we may loosely judge the difference between two mean Rand values to be genuine (rather than reflective of the vagaries of the simulated data sets) if they differ by more than, say, twice the associated MCSE. It is apparent that the MCSEs in these simulations are generally quite small in this practical sense.**

The comparison of methods using the Group 1 functions (periodic in nature) is displayed in Table 2. For almost every pattern of cluster sizes, Ward's method had the highest mean Rand index. Complete linkage often rated second best. The only time Ward's method was not superior was in the case of one very large group and three small groups (cluster sizes 35, 2, 2, 1), when average linkage did better for lower values of σ while single linkage (usually a straggler) performed best when the noise was increased. Figure 2 (top) (for the equal-cluster-size case) and Figure 2 (bottom) (for the case of one large cluster and three small clusters) show the mean Rand index for different values of σ for each clustering method.

[Figure 2 here]

The comparison using the Group 2 functions (decreasing in nature) is given in Table 3. Again, Ward's method had the highest mean Rand index in most situations, with complete linkage often second best. Figure 3 (top) shows how the mean Rand index varies for different values of σ for Group 2 functions with equal cluster sizes. Again, Ward's method does

very poorly in the case of one large cluster and three smaller clusters, with average linkage doing best in this specific case (see Figure 3 (middle)). Average linkage also did better overall (much closer to Ward's method and complete linkage) with Group 2, compared to the periodic functions of Group 1. For instance, Figure 3 (bottom) shows that average linkage performs best with two large clusters and two small clusters at high noise levels.

[Figure 3 here]

The comparison using the Group 3 functions (a combination of the periodic and decreasing curves) is given in Table 4. The results were similar to those for the first two groups, with Ward's method and complete linkage usually performing best (as shown in Figure 4 (top)). In the situation with one large cluster, Ward's method had the lowest accuracy while average linkage had the highest, particularly with higher σ values (see Figure 4 (middle)). Average linkage again did well, especially when there were two large cluster sizes and two small cluster sizes (see Figure 4 (bottom)).

[Figure 4 here]

In general, Ward's method almost always performs the best, having the highest mean Rand index in most situations, with the exception of situations where the data contain one or two very large groups and a few other very small groups, when average linkage does best. While complete linkage performs well in general, it never has the highest Rand index. Similarly, single linkage does the worst overall, except in occasional situations. For these reasons, we do not recommend that either complete linkage or single linkage be used when clustering functional data. In short, if the analyst suspects there may be one or two outlying curves forming very small clusters, average linkage is recommended. Otherwise, if the investigator suspects several clusters, all of somewhat substantial size, we strongly recommend using Ward's method to obtain the most accurate classification.

Overall, all methods perform better when there are no periodic trends in the data; both Ward's method and average linkage perform best in Group 2 and worst in Group 1. This observation is somewhat intuitive, considering periodic functions contain more within-curve variation, and may be more difficult to cluster accurately.

2.3 Case of Misspecification

In practice, the true number of groups in a data set is often unknown and must be specified by the investigator. For this reason, misspecification of the true number of groups is a common error when using hierarchical clustering algorithms. As a secondary investigation, a comparison of clustering methods was done for situations when the number of groups was misspecified. The misspecifications examined were not severe; since there were four true clusters, we examined the ramifications of setting $k = 3$ or $k = 5$ on the accuracy of the resulting partitions. We present results for the case of equal cluster sizes and the case of cluster sizes of 20, 10, 8, and 2. Other situations produced very similar results.

The generation of data sets was identical to the main study, except that the clustering was stopped at the wrong number of groups (either 3 or 5). The Rand index was calculated the same way, except that a perfect clustering is impossible in this situation, so the Rand index could not reach 1. When comparing clustering algorithms, however, we are comparing the Rand indices to each other and not to a perfect clustering partition.

The results were similar to those in the original simulation but depended somewhat on whether there was under- or overspecification. Once again there were also some differences between the different groups of data. In Group 1, when the number of groups was underspecified at $k = 3$, Ward's method had the highest mean Rand index in most situations. This can be seen in Figure 5 (top), which shows how the mean Rand index changes over the different values of σ for each method at equal sample sizes. Once again, complete linkage often had the second highest mean Rand index, but was never the highest. When the number of groups was overspecified at $k = 5$, Ward's method had the highest mean Rand index in most situations. For the case when the cluster sizes were unequal, average linkage had the highest mean Rand index at the lowest σ value. This pattern can be seen in Figure 5 (bottom).

[Figure 5 here]

For the simulations for Group 2, while Ward's method did have the highest mean Rand index in most situations, complete and average linkage are often much closer to Ward's (see

Figure 6 (top)). When the number of groups is over-specified ($k = 5$), Ward's method again performs best with equal cluster sizes. However, average linkage is much closer to Ward's method than for Group 1. Also, in the case of unequal cluster sizes, average linkage has the highest mean Rand index for a low σ value, as illustrated in Figure 6 (middle).

[Figure 6 here]

The simulations for Group 3 had very similar results to Group 2, with Ward's method usually having the highest mean Rand index when the number of clusters is overspecified ($k = 5$). Ward's method also performs the best when the number of clusters is underspecified ($k = 3$), as seen in Figure 6 (bottom).

In summary, single and complete linkage are not the best clustering methods for functional data. Choosing between Ward's method and average linkage should be done on a case-by-case basis. Ward's method has the highest accuracy in most situations. If the goal of the analysis is to identify a few outlying clusters and one large cluster, however, average linkage is the best method. It also appears that average linkage is the best choice when there are different-sized groups and when there is a good chance of overspecification of groups, particularly when the data have fewer periodic tendencies. The performance of Ward's method remains more stable than that of average linkage when most of the data are periodic.

3. ANALYSIS OF A REAL DATA SET

To compare these clustering methods in practice, we examined a functional data set originally analyzed in Alter, Brown, and Botstein (2000). Originally presented by Spellman et al. (1998), it contained functional observations measuring (over time) the log-transformed expression ratios of 78 yeast genes. The expression ratios were measured at 18 timepoints at 7-minute intervals. This data set fits the nature of this paper well in that biologists believe there are five clusters (based on the cell cycle phase of each gene) of unequal sizes; the assumed clustering structure is given in Table 5. Thus we use the Rand index to compare the outputs of the clustering methods relative to this suspected structure.

Table 5: Suspected Clustering Structure for Yeast Gene Data

Cluster Name	Cluster Size	Observations
G1	13	1-13
S	39	14-52
S/G2	8	53-60
G2/M	7	61-67
M/G1	11	68-78

S = Synthesis M = Mitosis G = Gap

Since the shapes (rather than the vertical positions) of the functional data were the crucial aspect on which to separate the genes into clusters, the data were initially centered by subtracting the sample mean response at each timepoint from each observed measurement. After the data were centered, the curves were smoothed using a B-spline smoother with 3 knots. **Note that the choice of 3 knots in this example is quite different from the 30 knots used in the simulation study, reflecting that an appropriate number of knots varies for different functional data. For instance, the simulated data had 30 knots interspersed within about 200 measurement points, while these real data have 3 knots interspersed within 18 measurement points, providing a set of curves that appeared visually to be moderately smoothed without loss of important structure.** The 78 smoothed curves are shown in Figure 7. Once the data were smoothed, a distance matrix was calculated using the trapezoidal-rule approximation of the L_2 distance between curves. Each of the four hierarchical clustering algorithms was applied to the matrix. Based on the clustering structure suspected by biologists, the Rand index was calculated for each method. For each method, the resulting clustering partition, as well as the Rand index, is shown in Table 6. **Dendrograms corresponding to each method are given in Figures 8 through 11.**

[Figure 7 here]

Table 6: Clustering Results for Each Hierarchical Method (Yeast Gene Data)

Method	Clustering Structure	Rand Index
Assumed Structure	{1-13}, {14-52}, {53-60}, {61-67}, {68-78}	N/A
Ward's Method	{1-2, 4, 6-11, 13, 70, 74-75}, {5}, {3, 12, 14-52, 61-67, 69, 78}, {54-60}, {68, 71-73, 76-77}	0.7949
Single Linkage	{1-4, 6, 8-64, 66-75}, {5}, {7}, {64}, {76-77}	0.4006
Complete Linkage	{1-4, 6-11, 13, 19, 22, 27, 30, 32, 41-42, 44, 47, 61-62, 65-66, 69-70, 74-75, 78}, {5}, {12, 14-15, 18, 21, 24-26, 28-29, 31, 33-35, 38-40, 43, 45-46, 48, 50, 52}, {16-17, 20, 23, 36-37, 50, 52, 53-60, 63-64, 67}, {68, 71-73, 76-77}	0.6857
Average Linkage	{1-4, 6-16, 18-19, 21-53, 61-63, 65-67, 69-70, 74-75, 78}, {5}, {17, 20, 54-60}, {64}, {68, 71-73, 76-77}	0.5951

Clustering partitions defined by the observation numbers in braces in the table.

The clustering results reflect our expectations based on the simulations. Ward's method performed the best with complete linkage second best. Average linkage did not perform as well, unsurprising since these data have periodic tendencies. Finally, as expected, single linkage was quite inaccurate.

We also briefly examine partitions other than the 5-cluster solution, calculating the Rand indices for these methods when the algorithms are stopped at 4 clusters and 6 clusters respectively, which represent misspecifications if the 5-cluster structure is the truth.

When the clustering algorithms are stopped at 4 clusters, Ward's method has the best Rand index (0.7972) by far, with complete linkage second best at 0.6787. Single linkage performs the worst (Rand index of 0.3640), with average linkage (0.5788) ranking third, again unsurprising due to the periodicity of the data.

When the clustering algorithms are stopped at 6 clusters, Ward's method again performs the best with a Rand index of 0.8162. However, unlike with the 4-cluster solution, average linkage is close behind Ward's method with a Rand index of 0.8035. Complete linkage has a Rand index of 0.7090, while single linkage again performs the worst with a Rand index of 0.4179.

[**Figure 8 here**]

[**Figure 9 here**]

[**Figure 10 here**]

[**Figure 11 here**]

A few objects and clusters bear closer scrutiny. All clustering algorithms appear to identify the fifth object in its own group when the clustering methods are stopped at five clusters. That object appears to be an outlier in cluster 1, and further investigation may be required to see if it truly warrants its own group. (With a 4-cluster solution, Ward's method and complete linkage place curve 5 with other curves, but average and single linkage still place it in its own cluster, **as indicated by the dendrograms**.) Generally, it appears that the true cluster 3 is well defined, being identified relatively well by all of the clustering methods except single linkage, which put most of the objects into one large cluster with the rest of the clusters only containing one or two objects **(see also Figure 9)**. The algorithms tend to merge some of the objects in clusters 1 and 5 into one group, and the true cluster 4 tends to be subsumed into the large second cluster. In short, the clustering results both display the idiosyncrasies of the different hierarchical algorithms and cast some doubt over the reliability of the assumed structure as a gold standard.

4. DISCUSSION

The goal of this study was to compare the performance of four major hierarchical clustering methods when applied to functional data. The Rand index was calculated for each method at each repetition of a simulation. The mean Rand index of the simulation for each method was compared. This index gives a proportion of pairs of objects that have been correctly clustered in the same group or correctly clustered into different groups.

In general, Ward's method had the highest mean Rand index in most situations, except when there were large differences among cluster sizes. This is a similar result to those in Hands and Everitt (1987), Kuiper and Fisher (1975), and Blashfield (1976) for other data types. In the situation with one large cluster and three very small clusters, average linkage performed the best. When the cluster sizes were generally different, average linkage performed well relative to Ward's method, particularly when there were fewer periodic tendencies in the data (a finding that relates specifically to clustering functional data). For certain values of σ, average linkage yielded a higher mean Rand index in such situations. While complete linkage usually performed well, it rarely performed better than all others. Single linkage did the worst in most (but not all) cases. This is another result similar to Hands and Everitt (1987).

Based on these findings, Ward's method is the best choice for clustering functional data, particularly when there are periodic tendencies in the data. Average linkage is recommended if the suspected clustering structure has one or two very large groups, particularly when the data are not periodic.

We found the results when the number of clusters was misspecified to be quite similar. Ward's method was usually the best, while average linkage performed best in some special situations, in particular when the number of clusters is overspecified.

In our cluster analyses of 78 yeast genes, Ward's method best recaptured the assumed clustering structure. The confirmation of the simulation results when analyzing the yeast data set indicates that the results of this study can be useful to researchers who cluster functional data.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the suggestions of an anonymous referee, which have improved this paper.

APPENDIX

Table 2: Mean Rand Values for Four Hierarchical Methods (Group 1 functions)

	$\sigma = 1.5$	$\sigma = 2.0$	$\sigma = 2.5$	$\sigma = 3.0$
Equal	W: 0.9768	W: 0.8154	W: 0.6981	W: 0.6474
Sizes	S: 0.6228	S: 0.3150	S: 0.3122	S: 0.3111
	C: 0.9593	C: 0.7785	C: 0.6615	C: 0.6146
	A: 0.9381	A: 0.6828	A: 0.4216	A: 0.3693
	Max. MCSE = 0.0058			
12	W: 0.9726	W: 0.8107	W: 0.6861	W: 0.6335
12	S: 0.6367	S: 0.3443	S: 0.3360	S: 0.3341
12	C: 0.9553	C: 0.7771	C: 0.6550	C: 0.6026
4	A: 0.9393	A: 0.6882	A: 0.4432	A: 0.3841
	Max. MCSE = 0.0062			
18	W: 0.9393	W: 0.7678	W: 0.6483	W: 0.5872
18	S: 0.7306	S: 0.4742	S: 0.4387	S: 0.4314
2	C: 0.9297	C: 0.7525	C: 0.6245	C: 0.5686
2	A: 0.9496	A: 0.7317	A: 0.5288	A: 0.4571
	Max. MCSE = 0.0067			
35	W: 0.7007	W: 0.5173	W: 0.4406	W: 0.4112
2	S: 0.9352	S: 0.8100	S: 0.7312	S: 0.6967
2	C: 0.7620	C: 0.5658	C: 0.4736	C: 0.4391
1	A: 0.9601	A: 0.8291	A: 0.7143	A: 0.6619
	Max. MCSE = 0.0046			
20	W: 0.9563	W: 0.7796	W: 0.6614	W: 0.6052
10	S: 0.6942	S: 0.4240	S: 0.4000	S: 0.3931
8	C: 0.9428	C: 0.7602	C: 0.6339	C: 0.5836
2	A: 0.9412	A: 0.7146	A: 0.5003	A: 0.4340
	Max. MCSE = 0.0062			

W = Ward's method, S = Single linkage, C = Complete linkage, A = Average linkage.

Table 3: Mean Rand Values for Four Hierarchical Methods (Group 2 functions)

	$\sigma = 4.0$	$\sigma = 4.5$	$\sigma = 5.0$	$\sigma = 5.5$
Equal	W: 0.9884	W: 0.9625	W: 0.9224	W: 0.8752
Sizes	S: 0.8654	S: 0.7814	S: 0.5232	S: 0.3445
	C: 0.9807	C: 0.9377	C: 0.8946	C: 0.8516
	A: 0.9678	A: 0.8998	A: 0.8689	A: 0.8384
	Max. MCSE = 0.0062			
12	W: 0.9870	W: 0.9591	W: 0.9188	W: 0.8731
12	S: 0.8774	S: 0.7887	S: 0.5460	S: 0.3792
12	C: 0.9768	C: 0.9360	C: 0.8922	C: 0.8499
4	A: 0.9674	A: 0.9083	A: 0.8727	A: 0.8441
	Max. MCSE = 0.0060			
18	W: 0.9658	W: 0.9170	W: 0.8791	W: 0.8337
18	S: 0.8909	S: 0.8054	S: 0.6367	S: 0.5197
2	C: 0.9603	C: 0.9081	C: 0.8715	C: 0.8266
2	A: 0.9603	A: 0.9124	A: 0.8894	A: 0.8574
	Max. MCSE = 0.0062			
35	W: 0.7799	W: 0.6958	W: 0.6405	W: 0.5788
2	S: 0.9642	S: 0.9473	S: 0.9127	S: 0.8679
2	C: 0.8266	C: 0.7460	C: 0.6933	C: 0.6332
1	A: 0.9757	A: 0.9517	A: 0.9297	A: 0.8939
	Max. MCSE = 0.0051			
20	W: 0.9784	W: 0.9360	W: 0.8890	W: 0.8423
10	S: 0.8873	S: 0.8123	S: 0.6145	S: 0.4576
8	C: 0.9663	C: 0.9167	C: 0.8789	C: 0.8331
2	A: 0.9659	A: 0.9105	A: 0.8846	A: 0.8512
	Max. MCSE = 0.0060			

W = Ward's method, S = Single linkage, C = Complete linkage, A = Average linkage.

Table 4: Mean Rand Values for Four Hierarchical Methods (Group 3 functions)

	$\sigma = 3.5$	$\sigma = 4.0$	$\sigma = 4.5$	$\sigma = 5.0$
Equal	W: 0.9751	W: 0.9351	W: 0.8907	W: 0.8460
Sizes	S: 0.8537	S: 0.7182	S: 0.5509	S: 0.3916
	C: 0.9574	C: 0.9076	C: 0.8658	C: 0.8166
	A: 0.9285	A: 0.8766	A: 0.8455	A: 0.7503
	Max. MCSE = 0.0046			
12	W: 0.9747	W: 0.9310	W: 0.8879	W: 0.8431
12	S: 0.8554	S: 0.7276	S: 0.5934	S: 0.4225
12	C: 0.9342	C: 0.9067	C: 0.8665	C: 0.8181
4	A: 0.9292	A: 0.8779	A: 0.8543	A: 0.7616
	Max. MCSE = 0.0049			
18	W: 0.9406	W: 0.8909	W: 0.8482	W: 0.7980
18	S: 0.8832	S: 0.7764	S: 0.6590	S: 0.5493
2	C: 0.9342	C: 0.8836	C: 0.8388	C: 0.7841
2	A: 0.9411	A: 0.8979	A: 0.8616	A: 0.7846
	Max. MCSE = 0.0065			
35	W: 0.7254	W: 0.6632	W: 0.6055	W: 0.5514
2	S: 0.9571	S: 0.9330	S: 0.8962	S: 0.8509
2	C: 0.7780	C: 0.7159	C: 0.6665	C: 0.6038
1	A: 0.9570	A: 0.9325	A: 0.9068	A: 0.8622
	Max. MCSE = 0.0047			
20	W: 0.9550	W: 0.9045	W: 0.8518	W: 0.8061
10	S: 0.8794	S: 0.7597	S: 0.6285	S: 0.4897
8	C: 0.9433	C: 0.8913	C: 0.8425	C: 0.7941
2	A: 0.9381	A: 0.8929	A: 0.8567	A: 0.7766
	Max. MCSE = 0.0052			

W = Ward's method, S = Single linkage, C = Complete linkage, A = Average linkage.

BIBLIOGRAPHY

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, **97**, 10101–10106.

Baker, F. B. (1974). Stability of two hierarchical grouping techniques – Case I: Sensitivity to data errors. *Journal of the American Statistical Association*, **69**, 440–445.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *The Psychological Bulletin*, **83**, 377–388.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Cunningham, K. M. and Ogilvie, J. C. (1972). Evaluation of hierarchical grouping techniques: A preliminary study. *The Computer Journal*, **15**, 209–213.

Edelbrock, C. (1979). Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody. *Multivariate Behavioral Research*, **14**, 367–384.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, **95**, 14863–14868.

Hands, S. and Everitt, B. (1987). A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. *Multivariate Behavioral Research*, **22**, 235–243.

Hitchcock, D. B., Booth, J. G. and Casella, G. (2007). The effect of pre-smoothing functional data on cluster analysis. *Journal of Statistical Computation and Simulation*, **77**, 1043–1055.

Hubert, L. (1974). Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *Journal of the American Statistical Association*, **69**, 698–704.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**, 193–218.

Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice Hall.

Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids. In: Dodge, Y. ed., *Statistical Data Analysis Based on the L_1-norm and Related Methods*, 405–416. New York: Elsevier/North-Holland.

Kettenring, J. R. (2006). The practice of cluster analysis. *Journal of Classification*, **23**, 3–30.

Kuiper, F. K. and Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics*, **31**, 777–783.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

McQuitty, L. L. (1960). Hierarchical linkage analysis for the isolation of types. *Educational and Psychological Measurement*, **20**, 55–67.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, **45**, 325–342.

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.

R Development Core Team. (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, (http://www.r-project.org/).

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. New York: Springer.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.

Schwaiger, M. and Rix, R. (2005). On the performance of algorithms for two-mode hierarchical cluster analysis – results from a Monte Carlo simulation study. In: *Data Analysis and Decision Support*, 141–148. Berlin: Springer.

Serban, N. and Wasserman, L. (2005). CATS: Clustering after transformation and smoothing. *Journal of the American Statistical Association*, **100**, 990–999.

Sneath, P. H. A. (1957). The application of computers to taxonomy. *Journal of General Microbiology*, **17**, 201–226.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, **38**, 1409–1438.

Sokal, R. R. and Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*. San Francisco: Freeman.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, **9**, 3273–3297.

Stein, M. (1995). Locally lattice sampling designs for isotropic random fields. *The Annals of Statistics*, **23**, 1991–2012.

Tarpey, T. (2007). Linear transformations and the k-means clustering algorithm. *The American Statistician*, **61**, 34–40.

Wakefield, J. C., Zhou, C. and Self, S. G. (2003). Modelling gene expression over time: Curve clustering with informative prior distributions. *Bayesian Statistics*, **7**, 711–721.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236–244.

Warrens, M. J. (2008). On the equivalence of Cohens Kappa and the Hubert-Arabie adjusted rand index. *Journal of Classification*, **25**, 177–183.