PalaeoMath 101

Groups III: Cluster Analysis

Over the last two essays we've discussed strategies for undertaking the analysis of multivariate datasets that are known to be characterized by a group-level substructure. This covers a lot of what we might need to do in terms of the evaluation of *a priori* group-based hypotheses. But what do we do if we suspect groups may present in our data, but don't have a very good idea who belongs to which group? If the groups are very obvious we can, of course, run the data through a procedure that assumes the presence of a single group (e.g., PCA, PCoord, Correspondence Analysis) and check the ordination plots. In such cases obvious clusters of data points that account for a high proportion of the sample variance should show up as distinct clouds of points in the space created the first few eigenvectors of the similarity matrix. But this will not always be the case, especially if the group-level structure is diffuse and/or swamped by other sources of variation. In such instances the standard approach would be to employ a formal 'cluster analysis'.

Cluster analysis is one of the oldest approaches to multivariate data analysis, tracing its origins back at least to the 1930s and 40s. It really came into its own, though, in the 1950s and 1960s when taxonomists began using numerical algorithms coded for processing on (the then new) computers to make the process of creating a classification more objective. This led to creation of the school of numerical taxonomy. Cluster analysis was the data analysis method of choice for most numerical taxonomists (see Sokal and Sneath 1963; Sneath and Sokal 1973). It is also one of the most widely used of all multivariate data analysis procedures with a solid literature of applications in fields ranging throughout the natural and social sciences and even on to areas such market research, advertising, and bioinformatics. On the face of it then, cluster analysis has an impressive history. Nevertheless, I must admit to finding the entire subject very *ad hoc*, lacking in organized development, and frustrating. So, with that personal caveat, and with a firm commitment to try not to let my own biases show through (too much), let's begin.

The best way to begin, of course, is with an example. Let's return to our old friends the trilobites and select a small subset of the previous data to illustrate some basic principles (Table 1).

Table 1. Trilobite data

Genus	Body Length (mm)	Glabella Length (mm)	Glabella Width (mm)
Acaste	23.14	3.50	3.77
Cheirurus	31.74	9.33	12.11
Phacops	27.23	5.30	8.19
Rhenops	55.94	19.00	13.10
Trimerus	89.43	23.18	21.52
Minimum	23.14	3.50	3.77
Maximum	89.43	23.18	21.52
Range	66.29	19.68	17.75
Mean	45.50	12.06	11.47
Variance	765.45	74.56	43.44

Since we wish to combine these taxa into groups based on the data we have collected (in this case distances between corresponding features of the body), our first task it to decide on a quantitative index we can use to summarize similarities and differences among these genera. These distances are represented by real numbers so we'll need to use an index designed to take advantage of fractional units. Since we are interested in relations between objects (= the *Q*-mode problem), the most obvious choice would to calculate a 'straight-line' or Euclidean distance between genera in the space formed by the three measured variables. Either of formulations of the Euclidean distance are the typical choices, the standard Euclidean distance …

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2} \qquad (12.1)$$

… or the squared Euclidean distance.

$$d_{ij} = \sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \qquad (12.2)$$

In both equations *i* represents the i^{th} specimen, *j* represents the j^{th} specimen, and *p* represents the total number of variables measured on each specimen. The only difference between these two indices is that the former produces a result whose units are the same as those of the original variables whereas the latter produces a result in squared units. Obviously this assumes all variables have been measured in the same units, which is the case for our data.

But there is a further decision we must make. Note the range of the body length variable is many times the magnitude of the lengths of the glabellar variables. This difference means the a greater proportion of the distance between genera will be due to differences in the body length than between glabellar length or width. Accordingly, the body length variable will 'count' more in expressing differences between genera than the glabellar variables.

As we've seen before, this forces us to decide whether differences between variables are part of the signal we're trying to assess or a nuisance factor. While most texts would recommend standardizing the variables to remove between-variable magnitude differences, my recommendation is to think more closely about this. If all variables are measured in the same units (in this case mm) differences between variables cannot always be regarded as artificial. In such cases the differences—and so the variables—should be maintained in their original form unless there is a good reason to do so. If the variable set includes mixed types some of which are intrinsically different in terms of their magnitude than others or if the difference between variables is not part of the hypothesis you wish to test (e.g., you're interested in a size-free analysis of similarity), it's best to standardize the variables as this operation forces them to account for the same proportion of overall sample variance. For our example we'll leave the variables in their raw form and use the Euclidean distance index (eqn. 12.1). Table 2 shows these distances for the Table 1 data.

Table 2. Euclidean distance matrix.

	Acaste	*Cheirurus*	*Phacops*	*Rhenops*	*Trimerus*
Acaste	0.00	13.32	6.29	37.46	71.39
Cheirurus	13.32	0.00	7.20	26.08	60.07
Phacops	6.29	7.20	0.00	32.18	66.07
Rhenops	37.46	26.08	32.18	0.00	34.78
Trimerus	71.39	60.07	66.07	34.78	0.00

By now the concepts behind, and overall form of, a *Q*-mode distance matrix should seem familiar. If not, go back and read the previous columns on *r*-mode and *Q*-mode analyses. There are other distance measures we could have used. The raw Euclidean distance ignores the between-variables covariance structure. If the covariance structure can be estimated to a reasonable degree of certainty, the Mahalanobis distance might be a better choice (see the Groups I column). Similarly, L. W. Penrose (1953) proposed a distance measure that could be used if multiple specimens from each genus were available. If we assume for a moment that Table 1 represents a matrix of means rather than individual measurements, the Penrose and Mahalanobis distance matrices for the example data are shown in tables 3 and 4.

Table 3. Mahalanobis distance matrix.

	Acaste	*Cheirurus*	*Phacops*	*Rhenops*	*Trimerus*
Acaste	0.00	2.83	1.51	2.56	2.74
Cheirurus	2.83	0.00	1.47	2.55	2.73
Phacops	1.51	1.47	0.00	2.56	2.27
Rhenops	2.56	2.55	2.56	0.00	2.78
Trimerus	2.74	2.73	2.27	2.78	0.00

Table 4. Penrose distance matrix.

	Acaste	*Cheirurus*	*Phacops*	*Rhenops*	*Trimerus*
Acaste	0.00	0.72	0.17	2.21	6.06
Cheirurus	0.72	0.00	0.20	0.68	2.99
Phacops	0.17	0.20	0.00	1.38	4.48
Rhenops	2.21	0.68	1.38	0.00	1.11
Trimerus	6.06	2.99	4.48	1.11	0.00

A careful inspection of tables 2-4 will show that the various estimates of between-genus similarity are quite different. Which distance index is best? That tends to be a matter of opinion. For this very simple dataset I'd argue the original Euclidean distance matrix makes the fewest assumptions about the data. But if we were analyzing a larger dataset, the choice might not be so clear. Moreover, these are the only three alternative indices. We don't have the space to go into all of the various distance measures that have been devised, but exotica such as the Bhattacharya distance, Bray and Curtis distance, Canberra distance, Gower distance, Chebychev distance, Chi-square distance, squared-chord distance, geodesic distance, Manhattan distance, etc. would all be potential alternatives. And that's just for the distance indices. If we were looking for a index that represented similarity as an angle between the vectors representing specimens in the variable space[1] we'd have another range of choices. By the same token, if we had a data matrix composed of binary, state codes (e.g., present/absent, large/small, simple/complex) we'd have another very extensive set of association indices that could be used to express similarity based on the proportion of shared presences and, in some cases, shared absences (e.g., Jaccard Index, Dice Index, Otsuka Index). There are also probability-based indices, indices for use with proportions, the list is virtually endless. To be honest, any of these similarity measures could also be used as the basis of a principal coordinates or *Q*-mode factor analysis (see the columns covering those methods). But in practice you simply don't see as much variation in the manner inter-object or inter-variable similarity is expressed in the theoretical development, or practical application, of these approaches compared to cluster analysis.

Once you've settled on an index you feel is appropriate to gauging similarity among your specimens, your next decision involves a choice of overall clustering strategy. Almost every textbook treatment offers a different taxonomy of clustering approaches. I'm going to restrict my discussion to the two most frequently used approaches: hierarchical agglomerative clustering and partition-based clustering.

Hierarchical agglomerative clustering is a classic top-down approach. At the beginning of the process each specimen is regarded as its own unique group. Then, as the level of similarity is progressively lowered, the separate groups are allowed to merge and the agglomeration history tracked. The procedure ends when all specimens have been collected into a single group.

The simplest hierarchical agglomerative clustering method is called *single linkage* or *nearest neighbour* analysis. The single linkage agglomeration history for the matrix shown in Table 2.

[1] This is the approach we'd use if we had mixed variable types in our data matrix.

Table 5. Cluster formation sequence. Abbreviations as follows: *A – Acaste, C – Cheirurus, P – Phacops, R – Rhenops, T - Trimerus*

Distance	Grouping History
0.00	*A,C,P,R.T*
6.29	*(A,P),C,R,T*
7.20	*(A,P,C),R,T*
13.32	-
26.08	*(A,P,C,R), T*
32.18	-
34.78	*(A,P,C,R,T)*
60.07	-
66.07	-
71.39	-

Under single-linkage cluster analysis the values in similarity/dissimilarity matrix are placed into rank order either from greatest to least (similarity) or least to greatest (dissimilarity, left column) and the objects or specimens joined into groups based on the greatest similarities between them (right column).

Stepping through this analysis, after 0.00 level, at which level all genera exist as discrete groups, the next greatest similarity value is 6.29. This is the distance between *Acaste* and *Phacops* (Table 2). So, at this level these two genera are joined to form a single group symbolized by the parentheses in the right-hand column of Table 6. Moving down the similarity value list we find 7.20, which is the level of distance-based similarity between *Cheirurus* and *Phacops*. Since *Phacops* is part of the *Acaste-Phacops* group, *Cheirurus* joins that group at the 7.20 level. The fourth smallest distance value is 13.32, which links *Acaste* and *Cheirurus*. However, since these two genera were already linked into the same group in the previous step the group structure is maintained through this level. Next there is a big decrease in the set of similarity values to 26.08 where *Rhenops* joins the *Acaste-Phacops-Cheirurus* group by virtue of its similarity with *Cheirurus*. This structure is also maintained through the 32.18 similarity mark that links *Rhenops* with *Phacops*. Lastly, *Trimerus* joins the main cluster at a distance level of 34.78 through its similarity with *Rhenops*. At this point all genera have joined the same group, so the analysis is complete. Graphically the structure of the Euclidean distance matrix under single linkage cluster analysis can be summarized by tree-like diagram called a *dendrogram* (Fig. 1).

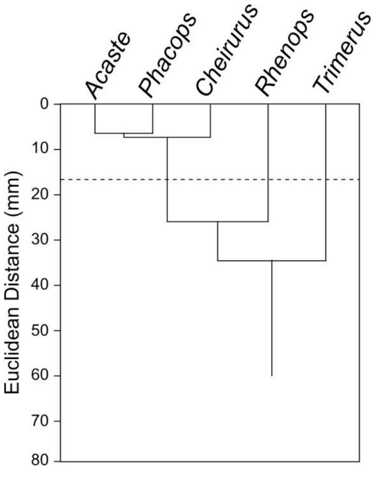

Figure 1. Single linkage dendrogram for the five-genus example trilobite data (see tables 1 and 2).

While dendrograms summarizes the structure of similarity matrices, by themselves they don’t tell you where the cluster or group boundaries are. What they do tell you is that, at a distance-similarity value 0.00, there are five groups and at 34.78 there is only one. A dashed line has been drawn in the middle of the greatest step between distance-similarity values that resulted in consolidation of the structure. This gap forms a natural subdivision in the distance data. With respect to these data this level corresponds to a 'morphological gap' of some type. Thus, it would make sense to set our grouping criterion at this level, in which case three clusters would be recognized: *Acaste-Phacops-Cheirurus, Rhenpops, and Trimerus*. This level-based approach to interpreting dendrograms is used throughout agglomerative, hierarchical clustering procedures with the level itself being referred to as the *phenon line* by numerical taxonomists and the *cut line* by statisticians. The problem, of course, is that the number of groups identified depends o where the phenon line is drawn, but there are no widely applied rules to guide this choice. The gap approach used above is one way to approach the issue o phenon line location. There are others (see below).

Generally speaking this result accords well with our data (see Table 1). *Acaste*, *Phacops*, and *Cheirurus* are all small individuals—in our dataset, at least—with the former two being noticeably smaller than the latter. *Rhenops* is over twice as long as these three taxa, though it has a proportionately smaller and decidedly elliptical glabella. *Trimerus* is larger still in overall body length, and a still smaller (proportionately) and circular glabella. Unfortunately, none of these geometric interpretations are evident from the dendrogram itself or from the information output by a cluster analysis (e.g., Table 5). Unlike eigenanalysis-based methods, the results of a traditional cluster analysis usually don't facilitate interpretation of the original data by any means other than *post hoc* comparison.

Of even more concern, however, is the issue of distortion of the data represented by the cluster analysis result—the dendrogram. Because of the rules used to create the clustering history and dendrogram important information about the structure of similarities among taxa is lost. Figure 1 implies that the distance between *Cheirurus* and *Acaste* is the same as the distance between *Cheirurus* and *Phacops*. This is not the case (se Table 2). Indeed, *Cheirurus* is almost twice as close (= similar) to *Phacops* as to *Acaste*. By the same token, *Rhenops* appears to be just as similar to *Cheirurus* as to *Acaste* and *Phacops*. This is also incorrect.

A measure of the amount of distortion present in the cluster analysis result can be derived by comparing the actual similarities to those implied by the dendrogram. For the trilobite data this comparison is illustrated in Table 6.

Table 6. Cophenetic distance matrix (see text for discussion).

	Acaste	*Cheirurus*	*Phacops*	*Rhenops*	*Trimerus*
Acaste		7.20	6.29	26.08	34.78
Cheirurus	13.32		7.20	26.08	34.78
Phacops	6.29	7.20		26.08	34.78
Rhenops	37.46	26.08	32.18		34.78
Trimerus	71.39	60.07	66.07	34.78	

Here the matrix's lower diagonal contains the Euclidean distance values (Table 2) observed in the raw data of Table 1. The upper diagonal contains those distances implied on the basis of the single-linkage dendrogram (Fig. 1). Numerical taxonomists refer to these implied values as the *cophenetic* values. This relation can also be expressed as a scatter diagram (Fig. 2) and summarized by calculating the correlation between the observed and implied distance values.

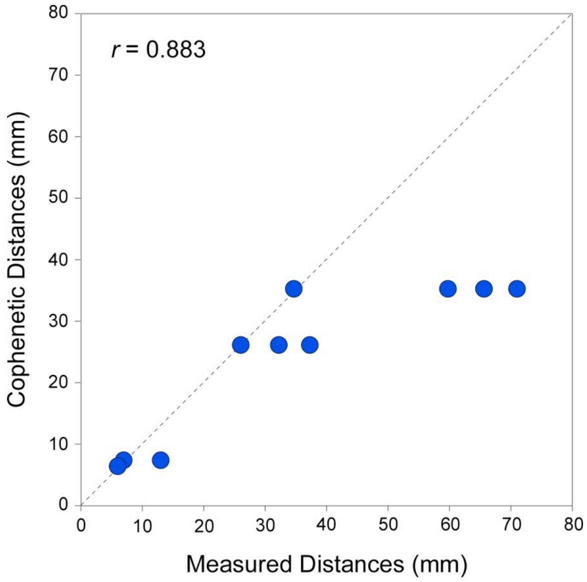

Figure 2. Distortion induced by a single-linkage cluster analysis of the five-genus example trilobite data. Dashed line represents model of perfect correlation.

In the context of a taxonomic cluster analysis this correlation is usually—and misleadingly—termed the 'cophenetic correlation coefficient' despite the fact that it is calculated using the standard Pearson product-moment formula. For our example analysis the level of distortion is substantial and most pronounced at the higher end of the distance scale—differentially affecting those data that are most important for inferring the deep structure of similarity relations.

Fortunately (or not as we shall see), single linkage isn't the only clustering game in town. The logical complement to single linkage is *complete linkage* or *furthest neighbour* linkage in which links are set at the level of the furthest or most dissimilar comparisons. Table 7 shows the linkage history for a furthest neighbour analysis of the example trilobite data. Note that the first two genera (*Acaste* and *Phacops*) join together at the same distance as before (6.29). This is because there is only one similarity value involved. After this, though, the order and level of group joining in set by the largest (instead of the smallest) similarity value (e.g., *Cheirurus* joins the *Acaste*-*Phacops* group at a distance of 13.32 instead of 7.20 (see Table 2).

Table 7. Cluster formation sequence using the complete-linkage approach. Abbreviations as in Table 5.

Distance	Grouping History
0.00	*A,C,P,R.T*
6.29	*(A,P),C,R,T*
13.32	*(A,P,C),R,T*
34.78	*(A,P,C),(R,T)*
71.39	*(A,P,C,R,T)*

The corresponding dendrogram and cophenetic correlation analysis are provided in Figure 3.

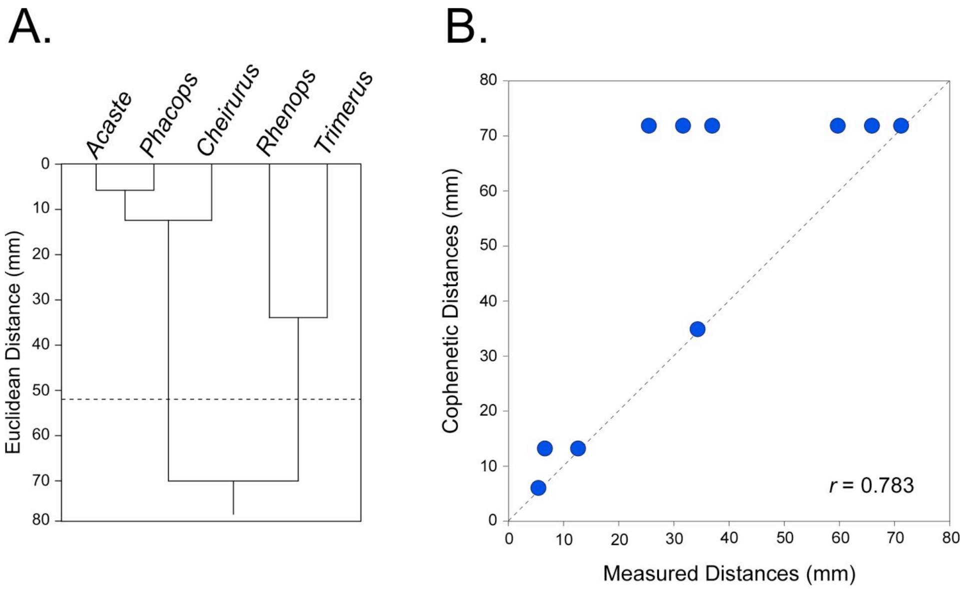

Figure 3. Results of a complete-linkage cluster analysis of the example trilobite data. A. Complete-linkage dendrogram. B. Associated cophenetic correlation scatterplot. Compare with figures 1 and 2.

Note the change in the structure of apparent distance relations among these taxa. Under complete linkage *Rhenops* and *Trimerus* are seen as forming a locus of similarity of their own, though as before, whether this substructure becomes part of the interpretation depends on where the phenon line for group recognition is located. If we use the same criterion as for the single-linkage analysis the *Rhenops-Trimerus* cluster would be recognized. Unfortunately, the level of distortion for this analysis is even higher than for the single linkage example with the greatest distortions, once again, occurring at the deeper levels of the hierarchy.

In order to overcome the obvious limitations of single-linkage and complete-linkage cluster analysis approaches a variety of alternative agglomerative procedures have been developed. One of the most popular among numerical taxonomists (and paleontologists) has been unweighted pair-group mean averaging (UPGMA). In most instances UPGMA maximizes the cophenetic correlation coefficient of a cluster analysis and so produces results with minimum levels of distortion (Farris 1969; Sokal and Rohlf 1970). The UPGMA approach does this by attempting to use a greater proportion of the information present in the similarity matrix. Let's work through the procedure as applied to the five-genus trilobite example.

Table 8. shows the linkage history for a UPGMA analysis.

Table 8. Cluster formation sequence using the UPGMA approach. Abbreviations as in Table 5.

Distance	Grouping History
0.00	*A,C,P,R.T*
6.29	*(A,P),C,R,T*
10.26	*(A,P,C),R,T*
31.91	*(A,P,C,R),T*
58.08	*(A,P,C,R,T)*

As before, five groups exist at the 0.00 distance level. Also as before, the shortest distance (greatest similarity) exists between *Acaste* and *Phacops* at the 6.29 level. Thus, during the first round of analysis these two genera join to form a group. *Cheirurus* exhibits the next smallest differences, both with *Acaste* and *Phacops* (both now members of the same group). However, the level at which *Cheirurus* joins this group is now set to the average of its distances-based similarity with both members of the group (= [7.20+13.32]/2, or 10.26). In other words, the UMPGA procedure attempts to 'split the difference' between these discrepant levels of similarity and so estimate the level of similarity between *Cheirurus* and the *Acaste-Phacops* group in as unbiased a manner as possible. *Rhenops* is next up, exhibiting a distance of 32.18 with *Phacops*. But in order to estimate its level of similarity with the *Acaste-Phacops-Cheirurus* group we average its similarity with all three of these taxa (=[37.46+26.08+32.18]/3, or 31.91). Finally, *Trimerus* is added to the group at its average distance to the other four genera (=[71.39+60.07+66.07+34.78]/4, or 58.08). The dendrogram determined form the UPGMA clustering history and associated cophenetic correlation plot are shown in Figure 4.

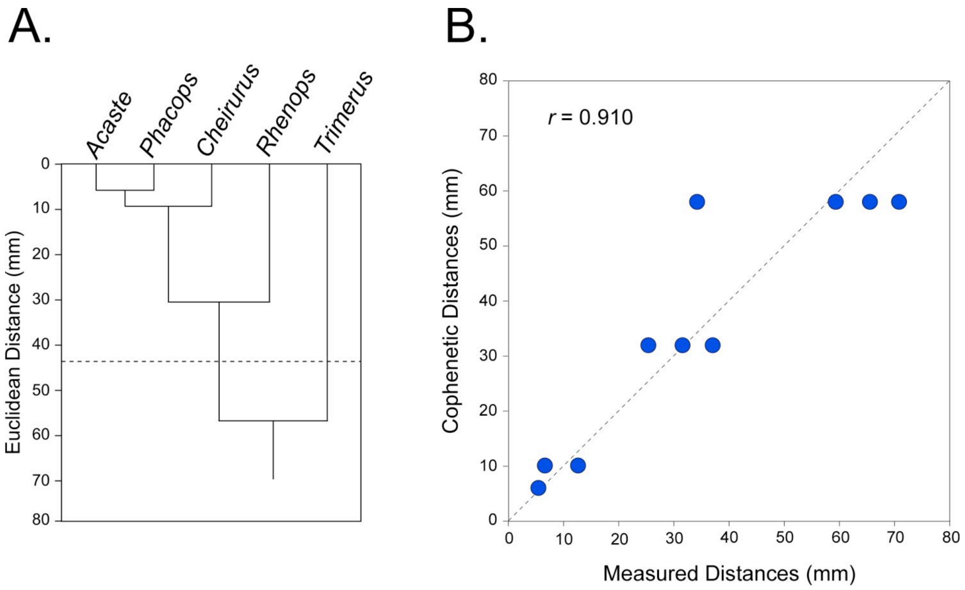

Figure 4. Results of a UPGMA cluster analysis of the example trilobite data. A. UPGMA dendrogram. B. Associated cophenetic correlation scatterplot. Compare with figures 1, 2 and 3.

The distortion resulting from the analysis is still on the large side (*r* = 0.910), but has been improved. Perhaps more importantly in terms of estimating the overall structure of the matrix and the deep structure of the cluster hierarchy, this distortion is now spread evenly across the entire range of distance values. Again, we can locate the phenon line at the level of the greatest morphological gap in which case two groups are identified. But note even though the pattern of relations between the single-linkage and UPGMA dendrograms (figs 1 and 4A respectively) are identical placement of the phenon line according to the morphological gap criterion yields different answers. If we changed the phenon-line location rule (e.g., first long branch) the group-recognition result would be identical. Which of these rules is 'best'? Both have their advantages and disadvantages. It is not clear which location rule should be used in this case.

A UPGMA analysis of the entire trilobite dataset (Table 9) is presented in Figure 5. This problem is a bit more realistic in size, though still not too large to trace detailed links between the dendrogram and original data.

Table 9. Trilobite data.

Genus	Body Length (mm)	Glabella Length (mm)	Glabella Width (mm)
Acaste	23.14	3.50	3.77
Balizoma	14.32	3.97	4.08
Calymene	51.69	10.91	10.72
Ceraurus	21.15	4.90	4.69
Cheirurus	31.74	9.33	12.11
Cybantyx	36.81	11.35	10.10

Cybeloides	25.13	6.39	6.81
Dalmanites	32.93	8.46	6.08
Deiphon	21.81	6.92	9.01
Ormathops	13.88	5.03	4.34
Phacopidina	21.43	7.03	6.79
Phacops	27.23	5.30	8.19
Placoparia	38.15	9.40	8.71
Pricyclopyge	40.11	14.98	12.98
Ptychoparia	62.17	12.25	8.71
Rhenops	55.94	19.00	13.10
Sphaerexochus	23.31	3.84	4.60
Toxochasmops	46.12	8.15	11.42
Trimerus	89.43	23.18	21.52
Zacanthoides	47.89	13.56	11.78
Minimum	13.88	3.50	3.77
Maximum	89.43	23.18	21.52
Range	75.55	19.68	17.75
Mean	36.22	9.37	8.98
Variance	346.89	27.33	18.27

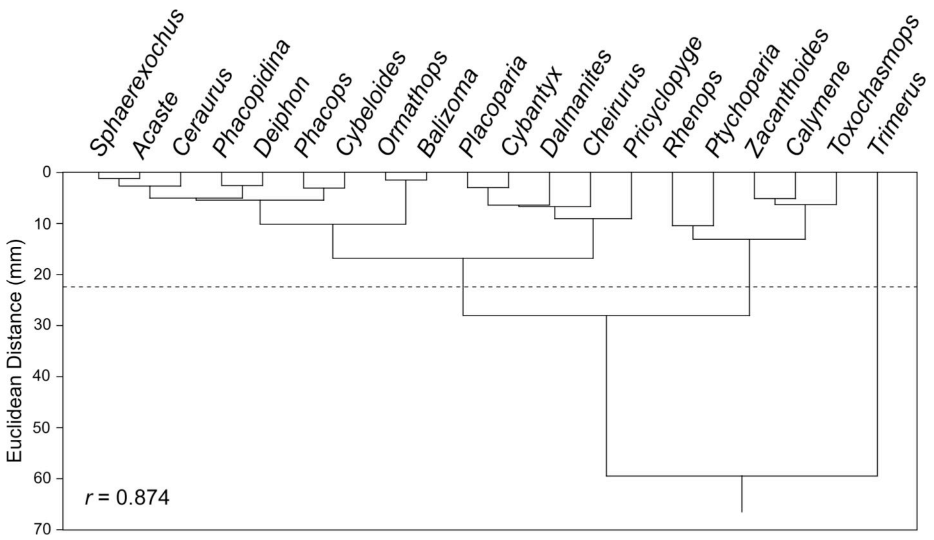

Figure 5. UPGMA dendrogram for the trilobite dataset.

The UPGMA dendrogram for these data shows a profound difference between *Trimerus* and the rest of the genera. This is clearly a reflection of the larger size of the *Trimerus* specimen. If the data had been standardized this difference would not have been as apparent.

Again number of groups recognized is determined by where we set the phenon line. If we look for natural breaks in the dendrogram (= morphological gaps), and ignore the profound size-related gap between *Trimerus* and the other genera we could most objectively identify two additional groups (see dashed line on Fig. 5). This distinguishes *Calymeme*, *Zacanthoides*, *Toxochasmops*, *Rhenops*, and *Ptychoparia* from the remaining genera. But note how I've changed the location rule again. If I apply either the greatest morphological gap rule or first long gap rules (see above) only two groups are identified.

Inspection of Table 9 shows these genera are united in having body lengths in the range of 45-65 mm. The *Trimerus* specimen has a much greater body length and all the rest exhibit body lengths much less than 45 mm. So, the deep structure of the UPGMA dendrogram appears to primarily reflect body length whereas the higher-level structure reflects differences between relative glabella size and glabella shape. As you can see, for small numbers of taxa and/or small numbers of variables, dendrograms can be interpreted in ways that tell us something meaningful about the data used in their construction. This inspection-based approach becomes much less practical for dendrograms containing large numbers of objects and/or large numbers of variables. For those datasets such clean and compelling interpretations are rare.

Since there are so many different agglomerative hierarchical clustering methods—not to mention similarity-dissimilarity-association indices—it is natural to ask how stable any particular result is. The most straight-forward way of approaching this issue is to compare the results yielded by different cluster analysis approaches. Biologists have tended to prefer averaging approaches because these address the issue of similarity matrix distortion. Statisticians have largely focused on other issues, notably ’continuity’ which is a shorthand way of saying that ‘small changes in the data should result in small changes in the dendrogram’. Under this criterion single-linkage approaches usually perform better than averaging approaches. Figure 6 shows the result of the single-linkage analysis of the Table 9 data.

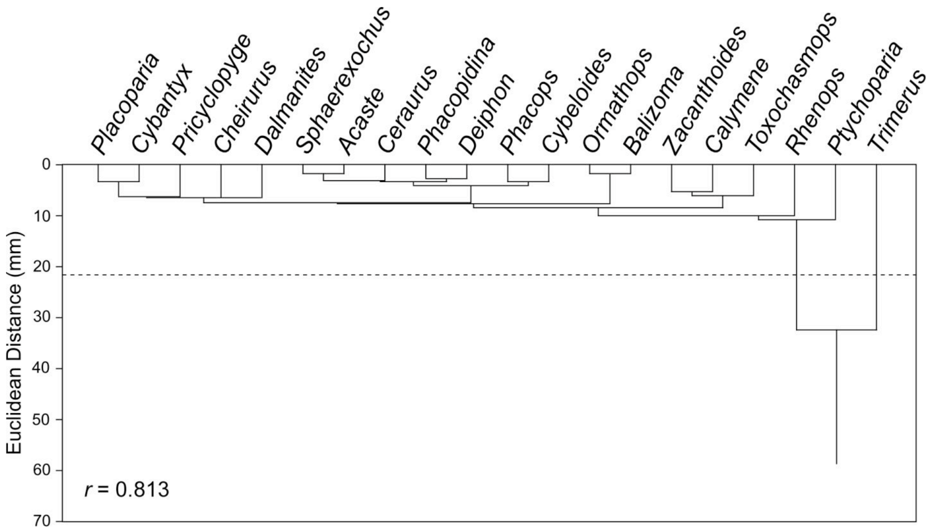

Figure 6. Single-linkage dendrogram for the trilobite dataset.

Obviously this is quite a different and more complex answer than was obtained by UPGMA analysis. Or is it? In many cases the top-level groups are the same, they’ve just been reordered. *Acaste* still links to *Sphaerexochus* and these two link to *Ceraurus*. But now this group is located in the centre of the dendrogram instead of on the left side. Since the order in which groups are presented has no significance, both the single-linkage and UPGMA patterns are equivalent for these taxa, though the UPGMA analysis links these taxa together at a lower level. This is the expected effect of averaging. The same can be said for the *Placoparia – Dalamanties*, *Ormathops – Balizoma*, and *Zacanthoides – Toxochasmops* groups.

Where these two dendrograms differ is in the manner in which the top-level groups are linked together. In the UPGMA analysis *Rhenops* and *Ptychoparia* are joined with the *Zacanthoides* group before this group joins the combined *Acaste-Placoparia* group. In the single-linkage dendrogram *Rhenops* and *Ptychoparia* chain together to link *Trimerus* to the other genera. *Trimerus*, of course, is the odd genus out in both analyses because of its large body length. What it all boils down to is a difference in the placement *Rhenops* and *Ptychoparia*. But this difference matters in terms of the interpretation. The UPGMA solution indicates that, if we accept *Trimerus* as a ‘group’, at least two other groups—possibly three—are present in the data. The single-linkage result suggests there is only one.

Which solution is correct? Strictly speaking, they both are in the sense that both are accurate and internally consistent representations of the structure of the distance matrix. They differ in the aspects of that structure they emphasize. Given the radical difference between the deep structure of the two dendrograms, the most well-supported generalized conclusion is that, for these trilobite data, which aspect of the distance matrix structure the analyst ends up emphasizing via their choice of approach makes quite a large difference to the answer obtained. Also recall, our trilobite data are very simple. For datasets containing more variables and or more objects, the differences between alternative clustering patterns would be expected to increase.

There is another approach to cluster analysis that I need to mention briefly: partitioning approaches (sometimes referred to as divisive or arbitrary origin methods). These are effectively the opposite of agglomerative methods. Instead of beginning with all objects as different groups and tracking the history of their agglomeration as the similarity threshold is decreased, divisive approaches begin with all objects as constituting a single group and track the history of their subdivision as the similarity threshold is increased. Whereas agglomerative approaches are ‘top-down’, partitioning approaches are ‘bottom-up’.

The most popular partition clustering approach is the *k*-means method. Here the user is required to specify the number of clusters known or expected to exist at the outset of the analysis. These are regarded as cluster seeds and usually initialized using random numbers scaled so that the seeds fall within the range of the observed data. During the first iteration the similarity between all objects and the seeds is calculated and the object most similar to each seed associated with it to form an initial group. The centroid between each seed group is then calculated and these centroids designated as new seeds. The process then repeats with the next most similar objects joining the seed groups and so on until all objects have joined a group. At higher levels in the analysis group joining is controlled by minimization of a group-level descriptor such as the trace of the group’s similarity matrix, that matrix’s determinant, or the Wilk’s λ statistic. Over the course of the iterations the seeds rapidly shift to the true centres of the emerging group clusters since the biasing effect of the artificial seeds diminishes with each iteration. A table of the three-group *k*-means solution for the entire trilobite dataset using the Wilk’s λ criterion as the clustering statistic is presented in Table 10.

Table 10: Three-group *k*-means solution (Wilks’ λ criterion).

Group 1	Group 2	Group 3
Acaste	*Calymene*	*Ptychoparia*
Balizoma	*Cheirurus*	*Rhenops*
Ceraurus	*Cybantyx*	*Trimerus*
Cybeloides	*Dalmanites*	
Deiphon	*Placoparia*	
Ormathops	*Pricyclopyge*	
Phacopidina	*Toxochasmops*	
Phacops	*Zacanthoides*	
Sphaerexochus		

Once again, note how different this result is from the UPGMA and single linkage dendrograms. The advantage of the *k*-means approach is that more specific grouping hypotheses can be evaluated and that the overall procedure can be performed much more quickly than agglomerative approaches, though given the speed of modern desktop computers this feature only matters for very large clustering problems. The primary disadvantage is that the *k*-means approach tends to produce suboptimal results owing to idiosyncrasies in the random placement of the original seeds. This can be overcome to some extent either by using actual objects as seeds, enhancing the specificity of the original hypothesis, but also requiring more be known about the problem at hand than is often the case. Another strategy is to test the result’s stability by performing the analysis multiple times using different starting seed values and comparing those results to the original for consistency, though this compromises the time-saving advantage.

So, what can we say about cluster analysis? Some degree of procedural variation exists for all the methods we've discussed to date. Usually these variants have focused on the manner in which the data are prepared (e.g., unstandardized, standardized, transformed) and type of similarity matrix used to quantify relations between variables and/or objects. However, once these decisions have been made, regression analysis, principal components/coordinates analysis, factor analysis, correspondence analysis, partial least squares analysis, and discriminant analysis all settle down to the application of standardized procedures (e.g., mostly forms of least squares analysis) whose statistical characteristics are well known. Cluster analysis differs because, in addition to the data type and similarity index variants, broad variation exists in the procedures used for undertaking the data analysis.

In addition, the statistical characteristics of these procedural variants are, by and large, not well known. This makes selection of the appropriate procedure for any particular dataset and data analysis situation much more difficult. Once a result is obtained, its interpretation is also complicated by the fact there is no widely agreed method whereby the phenon/cut line can be placed in agglomerative dendrograms, and by the instability of partition approach results. Moreover, the results of a cluster analysis do not lend themselves to efficient and nuanced interpretation in terms of the original variables in the manner in which eigenvector-based methods do.

Last, but by no means least, most cluster analysis methods fit a hierarchical model of inter-object similarity to the data even though there is nothing inherent in the structure of most similarity matrices that implies such a structure. Eigenvector-based methods also represent the structure of similarity matrices, but do not express that structure in terms of a hierarchy. Cluster analysis methods tend to adopt this approach because it prevents objects from being assigned to more than a single group. The upside of this is that, provided your data adequately capture the hierarchical structure you suspect is present, and provided the hierarchical structure is a clear and dominant feature of those data, cluster analysis will likely find it. The downside is that cluster analysis will find a visually compelling hierarchy in any dataset regardless of whether the hierarchical signal is actually there—even in random data (Fig. 7).

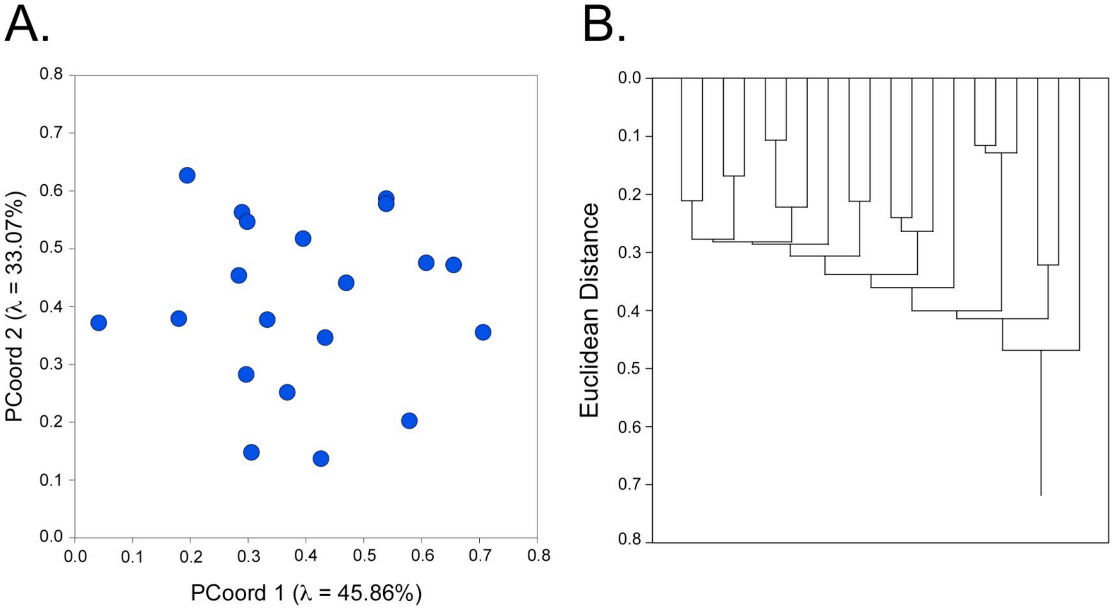

Figure 7. Quantitative analysis of a 20 x 3 table of random numbers. A. principal coordinates analysis result. B. Single-linkage cluster analysis result (distance matrix). Note that by forcing these random data to be represented by a hierarchical model the cluster analysis dendrogram displays much more apparent structure than the non-hierarchical PCoord ordination plot. While this random dendrogram differs from previous data-based patterns in the number of long terminal branches, this will not always be the case, especially for small datasets. Also, because of the statistical properties of eigenanalysis it is easier to test the PCoord result for the null hypothesis of random variation than it is to test the cluster analysis result.

I hope I haven't been *too* hard on cluster analysis. Like all data analysis methods, it has its place and can be very helpful when applied intelligently and with due caution. In this context the reader is well advised to remember that Cormack's (1971) observation about cluster analysis—that it is not a satisfactory alternative to clear thinking—actually extends to all numerical data analysis procedures.

There was a time when it was hard to page through an issue of the top half dozen systematics and/or palaeontology journals and not see a dendrogram. Not true now. Indeed, I'd say eigenvector-based methods are now more widely used by palaeontologists for routine data analysis that clustering methods. Regardless, cluster analysis lives on in the phylogenetics literature in the guise of numerical cladistics which was derived directly from the cluster analysis procedures developed by the phenetic school of numerical taxonomy (see Sokal and Sneath 1963; Sneath and Sokal 1973). Indeed, these books remain two of the most comprehensive treatments of cluster analysis, especially for biologists and paleontologists. Other, more recent references that focus on statistical issues, but are readable by non-mathematicians, include Kauffman and Rousseeuw (2005) and Fielding (2007).

In terms of computer programmes, cluster analysis is such a long-standing and popular technique that it is rare to find a commercial multivariate statistical package that doesn't include it in some form. These range from inexpensive plug-ins for MS-Excel (e.g., UNISTAT, XL-STAT, StatistiXL) to sophisticated stand-alone packages. A variety of older books on statistical analysis also come with cluster analysis software (e.g.,Davis,1981; Backer 1995; usually for DOS operating systems). As a last—or maybe as a first—resort, there are a large number of freeware and shareware cluster analysis applications available for download from the Internet.

Norman MacLeod
Palaeontology Department, The Natural History Museum
N.MacLeod@nhm.ac.uk

REFERENCES

Backer, E. 1995. *Computer-assisted reasoning in cluster analysis*. Prentice Hall, Upper Saddle River, New Jersey. 400 pp.

Cormack, R. M. 1971. A review of classification. *Journal of the Royal Statistical Society, Series A*, **134**, 321–367.

Davis, J. C. 1986. *Statistics and data analysis in geology (second edition)*. John Wiley, New York. 646 pp.

Farris, J. S. 1969. On the cophenetic correlation coefficient. *Systematic Zoology*, **16**, 174–175.

Fielding, A. H. 2007. *Cluster and classification techniques for the biosciences*. Cambridge University Press, Cambridge. 258 pp.

Kaufman, L. and **P. J. Rousseeuw**. 2005. *Finding groups in data: an introduction to cluster analysis*. Wiley-Interscience, Hoboken, New Jersey. 368 pp.

Sneath, P. H. A. and **R. R. Sokal.** 1973. *Numerical taxonomy: the principles and practice of numerical classification*. W. H. Freeman, San Francisco. 573 pp.

Sokal, R. R. and F. J. Rohlf. 1970. The intelligent ignoramus, an experiment in numerical taxonomy. *Taxon*, **19**, 305–319.

Sokal, R. R. and **P. A. Sneath**. 1963. *Principles of numerical taxonomy*. W. H. Freeman, San Francisco. 359 pp.

Don’t forget the *Palaeo-math 101* web page, now at a new home at:
http://www.palass.org/modules.php?name=palaeo_math&page=1