

PalaeoMath 101

Multidimensional Scaling and Ordination

This column is devoted to completing our discussion of basic multivariate data analysis by tying up a bit of a loose end. In previous columns I've approached the description of various bivariate regression and multivariate analysis methods from a variable-centred point-of-view, at least for the most part. That is to say, I've focused on describing how the geometry of the calculations allows us to sense, summarize, and understand relations among variables across the example datasets. The obvious exceptions have been my discussions of principal coordinate (PCoord) and Q-mode factor analysis (Newsletter 61) as the explicit goals of both these techniques is to construct of a picture of similarity relations between the objects from which measurements and/or observations were obtained.

Nevertheless, I'm sure it has not escaped you attention that the resolutely *r*-mode methods of principal components analysis and factor analysis (newsletters 58 and 59 respectively) also produce pictures of similarity relations among objects in the form of scatterplots of those objects within the space formed by the new component or factor variables. Correspondence analysis yields similar plots of between-object similarity based on frequency data, with the added advantage of enabling geometric relations between variables to be represented in the same space (see Newsletter 62).¹ Likewise, canonical variates (Newsletter 65) are convenient for displaying relations between objects in spaces that emphasize differences between *a priori*-defined groups. The spaces formed by partial least squares analysis axes (Newsletter 63) are, again, used to construct images of between-object similarities, though in this case the focus of the procedure is the degree to which different datasets exhibited the same inter-object similarity structure.

The *r*-mode/Q-mode distinction is real, but actually pertains to the means by which each technique's goals are achieved. All these approaches provide critical information about the structure of covariation among variables *and* can be used to obtain a visual sense of the structure of similarity relations among objects. Consequently, their use allows us to first optimize, then interpret the nature of the geometric spaces we within which portray object similarity patterns. This gives us the power to test a wide range of hypotheses because we can see—and so understand—the nature of relations between variables *and* objects. But this power comes at a subtle and little-understood price.

The feature that gives PCA, PCoord, factor analysis, correspondence analysis, CVA and PLS their power to bridge the gap between variables and objects is their common dependence on eigenanalysis. Eigenanalysis is a method for estimating the major directions of variation in sets of numbers. While the assessment of variation in terms of both its relative directions and magnitudes is a logical and common-sense approach to understanding the behaviour of variable sets, is this the only—or even the most appropriate—approach we could use to understand *similarity* relations between objects?

Note that cluster analysis (Newsletter 66) is very different in this respect from the eigenanalysis group of techniques. Cluster analysis provides a means whereby object-based similarity relations can be represented, but does not do so by assessing the structure of relations among variables. Rather, it treats each object as set of descriptions (= the state of the variables the object manifests) and simply calculates a measure between-object similarity, usually in terms of a distance. Once the structure of inter-object distance relations has been estimated cluster analysis summarizes that structure, but does so in a manner that provides no direct insight into the nature of relations between the variables. In other words, cluster analysis provides an answer to the question of how similarity relations are organized across objects in a dataset, but has a very limited capacity to help us understand why that answer was obtained in terms of pattern relations among the original variables. Should this be a cause for concern?

¹ It is possible to scale PCA and factor axes to portray relations between object and variables in the spaces so formed (these are called 'biplots'), but this represents a step beyond the calculation of principal components or factor axes *sensu stricto*. Such scaling is a fundamental part of correspondence analysis.

The only reasonable answer to this question is 'It depends.'. If the hypothesis we are testing can be resolved solely by determining the structure of between-object similarity relations there may be no need to understand the nature of the space within which those similarity relations are portrayed. How the space related to the original variables, or whether it has been optimized in terms of the representation of major patterns of variation across the dataset as a whole, may be irrelevant. But one thing that doesn't change is the need to have some way of sensing how 'good' the resulting picture of between-object similarity relations is.

Of course, the eigenanalysis family of techniques provides this quality-control information through the eigenvalues. These are the lengths of the major and minor axes of a hyperdimensional spheroid fit to the variables when represented in a variance-covariance space (Fig. 1). For sets of highly correlated variables, the first few eigenvectors will typically subsume a much greater proportion of the overall variance than any single variable axis. If these vectors are then used to portray inter-object similarity relations, a 'good' result will typically be that in which the overwhelming majority of the observed variation is represented by 2-3 eigenvector-based axes. Decision criteria vary depending on the problem under investigation, but most analysts would feel that being able to summarize 90 to 95 percent of the observed variation in a dataset in 2-3 eigenvectors is accurate for the purpose of authoritative interpretation. Note, this feature of eigenanalysis is delivered by the technique's ability to summarize the structure of covariance relations among variables. It has nothing to do with assessments of inter-object similarity *per se*.

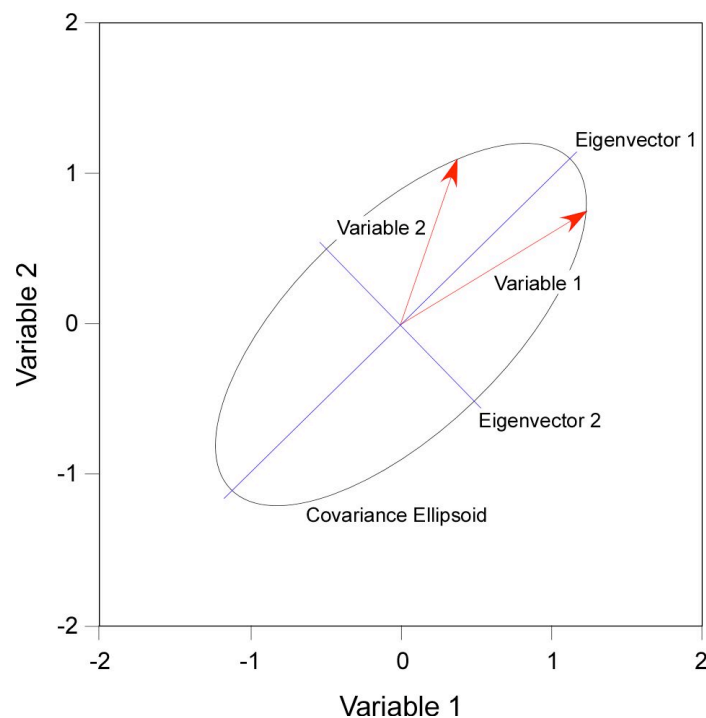


Figure 1. Geometric relations between variables and eigenvectors. Eigenvectors are the major axes of a covariance spheroid (ellipsoid in two dimensions) centred at the origin of a variance-covariance space and recursively fit to the swarm of vectors representing the variance-covariance structure of a set of variables. The eigenvector coefficients are the cosines of the angles between the eigenvectors and the variable vectors. The eigenvalues are the lengths of the eigenvectors. The sum of these lengths is proportional to the total amount of variance across all variables in the system.

As I've alluded to above, cluster analysis has no parallel to the eigenanalysis in terms of numerical quantities that summarize the amount (e.g., eigenvalues) and kind (e.g., eigenvector coefficients) of information present along the axis against which the cluster pattern is formed. Under agglomerative clustering this is typically done by estimating the

amount of overall distortion inherent in the portrayal of those relations as a dendrogram. Operationally this is accomplished via back-calculation of an implied similarity matrix and comparison of that to the original similarity matrix using the cophenetic correlation coefficient (see Newsletter 66), which provides a very rough assessment of the overall distortion. As we have seen, force-fitting a hierarchical model to the data in order to distinguish clusters of objects can result in substantial distortion being introduced. Partition clustering approaches (e.g., *K*-means clustering) are often better in this respect as they do not attempt to fit a hierarchical model to the data (and so avoid generating distortions). However, these approaches assume the data exhibit subgroup-level structure and will always find that structure regardless of whether it is the dominant pattern in any particular dataset.

In other words, trying to portray the pattern of high-dimensional data in low-dimensional spaces will always involve distortion. But let's go back to first principles for a moment to regain our perspective. The objects in a dataset exhibit patterns of attributes that exist at (potentially) all scales. If distributions of attributes are strongly clustered there will be a clear distinction between large-scale (e.g., between groups) and small-scale (e.g., within-groups) patterns. Canonical variates analysis takes explicit advantage of this distinction by using a two-step eigenanalysis procedure to focus the analysis on portraying between-groups differences. However, even though small-scale patterns play a role in determining the orientation of eigenvectors, the resulting vector orientations will always be most closely aligned with the large-scale patterns present in the data. If all the data in an analysis are normally distributed, this difference between large and small-scale patterning may not be that important—hence many multivariate statisticians' love of qualifying their statements in terms of normality assumptions. But many (most) biological and paleontological datasets exhibit distributions that are far from normal. In those cases where the focus of the analysis is on the portrayal of inter-object similarity relations is cluster analysis or eigenanalysis the best we can do?.

Of course, the answer to this question is 'No!'. The name of the technique that can help us in this regard is multidimensional scaling, or MDS for short. Like cluster analysis, MDS is actually a family of techniques with many different variants on the common theme. Also like cluster analysis, MDS can be used to analyse an astounding variety of data. Best of all, MDS makes no prior assumptions about the nature of similarity relations present in the data (e.g., hierarchical or non-hierarchical), the scales of similarity patterning they contain, the distribution of the data within variables, or the presence/absence of well-defined sub-group structure within the overall dataset.

The term 'multidimensional scaling' means different things to different people (see below). In this column I'm going to refer to it as a specific type of numerical procedure that has seven basic steps.

1. A Q-mode distance matrix δ is calculated between all pairs of objects across a set of variables and/or observations where δ_{ij} is the distance between object i and object j .
2. These objects are then arranged in a configuration within a k -dimensional space. The typical starting configuration is random, but may also be the configuration of the first k principal coordinates.
3. A new distance matrix d is calculated between all pairs of objects across the k dimensions of configuration space where d_{ij} is the distance between object i and object j .
4. A regression of d on δ is performed. The purpose of this regression is to specify the functional relation between the configuration space and the original (high-dimensional) distance data. Based on this regression the quantity \hat{d} is then determined using the following equation.

$$\hat{d}_{ij} = a\delta_{ij} + b + \varepsilon_{ij} \quad (13.1)$$

where a and b are constants (e.g., the slope and y -intercept of a d on δ linear regression) and ε is the error associated with the regression. This quantity (\hat{d}) is called the ‘disparity’.

5. A test is made of the fit between the disparity distances and the configuration distances in the k -dimensional space. This goodness of fit is summed into single statistic that represents the amount of distortion represented by the test configuration of points in the k -dimensional space.
6. The disparity values (\hat{d}) are then substituted for the configuration distance values (d) and used to calculate a new configuration space.
7. Steps 4–6 are repeated until the test performed at step 5 indicates no further improvement in the goodness-of-fit of the configuration space on the original distance data matrix can be achieved, or until a predetermined iteration limit is reached, whichever comes first.

Note first how similar this fitting procedure is—at least in principle—to the procedure for obtaining an eigenanalysis (see Newsletter 58). If only two variables are involved the equation of the first eigenvector (= first principal component) can be calculated directly. This is identical to calculating the slope of a major axis regression (see Newsletter 56). For all higher dimensional datasets the first eigenvector must be estimated using an iterative procedure that compares the goodness of fit of a test vector to the major dimensions of variation within the sample. The orientation of subsequent eigenvectors are fit in the same way subject to the additional constraint that they must be orthogonal to the orientation of all preceding eigenvectors.

The MDS approach uses a similar strategy to estimate the configuration of points in the k -dimensional space. But note that the MDS algorithm (1) is restricted to assessing the fit across a pre-determined set of dimensions (e.g., does not try to use all possible dimensions to obtain a perfect solution), (2) assesses the fit between the disparity and configuration distances globally, across the entire scale of distances present in the dataset, and (3) places no constraints on the orientation of the solution vectors. Thus, unlike strictly eigenanalysis-based methods no preference is given to fits across large-scale patterns in the data; and, like CVA axes, there is no guarantee the final MDS axes will be aligned with major patterns of variation in the original data or even that they will be orthogonal to one another in the space of the original variables.

Of course, the goodness-of-fit test is very important to the outcome of the procedure. While there are alternatives, most MDS programs take Kruskal’s (1964a) Stress (1) statistic as a reference point to obtain this estimate.

$$S = \sqrt{\frac{\sum_{i=1}^n \sum_{j>i}^n (\delta_{ij} - d_{ij})^2}{\sum_{i=1}^n \sum_{j>i}^n d_{ij}^2}} \quad (13.2)$$

This is a scaled sum-of-squared-differences estimator.

Let’s take a look at an MDS analysis now. But instead of jumping head-first into the trilobite data, let’s analyze a dataset about which we can develop some expectations using intuition alone. For this purpose the analysis of geographic map distances presents a compelling target of opportunity. For our example, let’s use a area with which many UK palaeontologists are reasonably familiar, the Isle of Wight

Figure 2 shows a simple map of the Isle of Wight with the location of 13 named locations marked and Table 1 shows a road distance triangle for these locations.



Figure 2. Isle of Wight with selected locations marked. Modified from Google Map (2008).

Table 1. Road distances (in miles) between selected Isle of Wight locations (see Fig. 2).

| | Bembridge | Blackgang | Brightstone | Cowes | Newport | Northwood | Ryde | Sandown | Seaview | Shalfleet | Shanklin | Totland | Ventnor |
|-------------|-----------|-----------|-------------|-------|---------|-----------|------|---------|---------|-----------|----------|---------|---------|
| Bembridge | 0.0 | | | | | | | | | | | | |
| Blackgang | 18.0 | 0.0 | | | | | | | | | | | |
| Brightstone | 17.8 | 5.9 | 0.0 | | | | | | | | | | |
| Cowes | 1.7 | 13.7 | 12.2 | 0.0 | | | | | | | | | |
| Newport | 10.7 | 9.0 | 7.1 | 5.0 | 0.0 | | | | | | | | |
| Northwood | 14.0 | 12.4 | 10.2 | 1.6 | 3.2 | 0.0 | | | | | | | |
| Ryde | 6.9 | 16.2 | 13.6 | 11.4 | 6.4 | 9.6 | 0.0 | | | | | | |
| Sandown | 5.1 | 10.4 | 15.4 | 14.0 | 8.9 | 12.3 | 6.4 | 0.0 | | | | | |
| Seaview | 3.9 | 17.3 | 15.8 | 15.2 | 9.0 | 12.3 | 3.0 | 6.3 | 0.0 | | | | |
| Shalfleet | 17.0 | 11.1 | 4.9 | 7.5 | 6.1 | 8.1 | 12.3 | 15.0 | 16.2 | 0.0 | | | |
| Shanklin | 7.3 | 9.5 | 12.0 | 14.2 | 9.1 | 12.4 | 8.6 | 2.2 | 8.5 | 15.1 | 0.0 | | |
| Totland | 23.2 | 12.6 | 8.3 | 14.1 | 12.4 | 14.3 | 18.7 | 21.2 | 22.5 | 6.3 | 21.4 | 0.0 | |
| Ventnor | 11.0 | 6.0 | 12.6 | 15.7 | 10.7 | 14.1 | 12.1 | 5.9 | 12.1 | 16.9 | 3.6 | 19.3 | 0.0 |

Given that the road distance matrix locates each destination relative to every other, we should be able to use MDS to reconstruct the map shown in Figure 2 from the data present in Table 1. Some error will be generated owing to the fact that road distances are not the same as straight-line map distances. This discrepancy will help make an important point about MDS a bit later on. Nonetheless, given the scattered nature of these locations across the island the rank order of road distances should provide relatively accurate estimates of their map distances from one another.

Since MDS is a computation-intensive, iterative procedure I'm not able to take you through all the individual calculations performed. A sample of the first few steps following the simple MDS procedure outlined in Jackson (1991) is provided in the *PalaeoMath 101* spreadsheet (see url below). Indeed, because the initial configuration of points can be random—in which case it will differ with each analysis depending on the value used to seed the random number generator—the calculations will differ in their details with each run of the program, even for the same dataset! What we can do (below) is discuss what options were selected for the analysis, why, and how those affect the results we obtain.

The first option we need to specify when implementing MDS analysis is the dimensionality of the solution. That's actually an easy decision for the Isle of Wight data. Even though we are working with a 13 x 13 matrix of distances, we have good reason to suspect the correct

dimension of the solution is 2. After all, these are distances taken from a flat map and the result we're after is a reconstruction of that map.

In point of fact, virtually all MDS analyses try to fit the data into a two-dimensional or three-dimensional space for the simple reason that these spaces are able to be visualized using standard graphing methods. The MDS procedure can find solutions in high dimensional spaces. But given that the point of MDS is to visualize similarity relations existing in datasets, there usually seems little point in creating solutions that can't be visualized in their entirety. Naturally this principle also applies to other multivariate methods (e.g., PCA, PCoord, Correspondence Analysis) when these are used to achieve a data-scaling purpose. Regardless, the choice of the solution's dimensionality is more obvious with MDS since this value must be specified by the user at the outset of an analysis. Also, because it is an iterative procedure, the first two dimensions of a two-dimensional MDS solution will not necessarily bear any relation to the first two dimensions of a three-dimensional MDS solution.

The second option needing specification is the regression model the algorithm will use to estimate the disparity distance values. Since distances are ratio-type variables, with a true scale and true zero point, we could draw on analogy with PCoord analysis and use a linear regression as a basis for fitting the configuration of data points to the distance matrix. Selection of a linear regression model means we will try to match between-object (= inter-location) distances as closely as possible to the original distance matrix shown in Table 1 (see Fig. 3). Under this model, our result would be example of 'metric MDS' for we are using all the information present in the configuration and basis distance matrices to scale the data. I'll have more to say about this decision shortly.

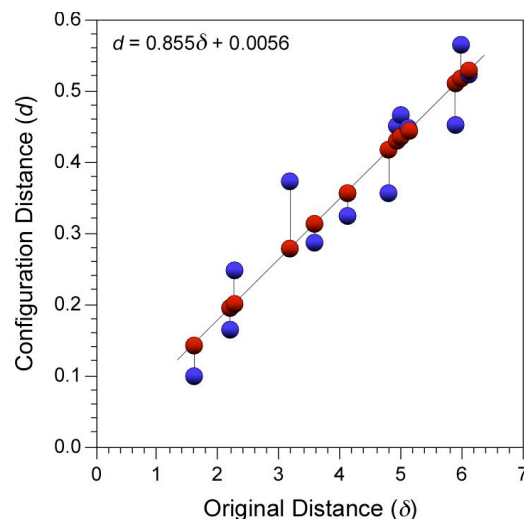


Figure 3. Example linear regression of MDS configuration distances ($k=2$) on the original distances for the first 13 between-location comparisons in the Isle of Wight data (see Table 1). Blue symbols represent raw data values for this fitting cycle. Red symbols represent disparity distances (\hat{d}) determined as a result of the regression analysis. Regression residuals are indicative of the error associated with this fitting iteration.

There are a few other decisions that must be made, but these will differ from between application programmes. Because I'm using the MDS procedures included in the *XL-Stat* programme package to perform these example calculations I have the ability to select between several stress statistics, though for simplicity and consistency I chose the Kruskal's Stress (1) index. It is also worth noting that *XL-Stat* employs a stress majorization algorithm based on de Leeuw's (1977) iterative majorization in order to adjust positions of objects in the configuration space during refinement of the MDS solution. This approach has been shown to yield results that converge on the optimal configuration more quickly Kruskal's (1964a) steepest decent approach.

Figure 4 shows results of the two-dimensional, metric MDS representation of the Isle of Wight distance matrix (Table 1).

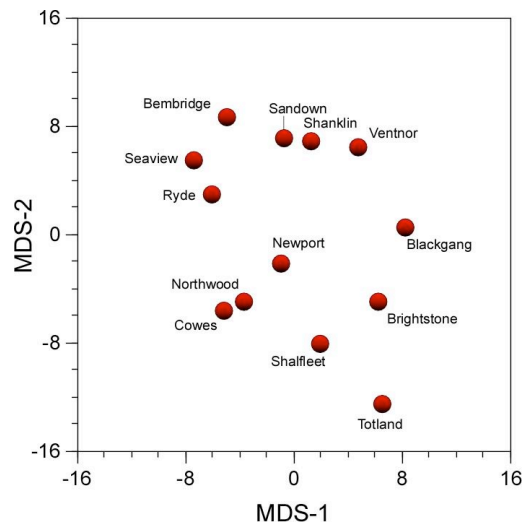


Figure 4. Metric MDS configuration space ($k=2$) for the Isle of Wight location distance data.

Inspection of the diagram reveals that all locations are in their correct relative positions. However, note that the MDS-1 axis is not aligned with the major axis of the point distribution, as it would be of this was a PCA or PCoord space. There are two reasons for this. First, the final point configuration is, to some extent dependent on the initial configuration (= configuration from which the iterative procedure begins). Because *XL-Stat* package uses a random configuration as the starting point, there is no guarantee the first MDS axis will lie close to the first PCA/PCoord axis.

Note also there are no equations for the MDS axes that can be used to relate the MDS space back to the space of the original variables. This is because the MDS space represents a series of adjustments of the relative point spacing in the context of the original or starting configuration space. Essentially, information is added piecemeal to the MDS solution by the sequence of regression-based adjustments (see Fig. 3) with the final configuration bearing no simple relation to the space of the original variables. This might be seen as a difficulty when it comes to interpretation and it is truly what separates MDS from PCA/PCoord. But this piecemeal procedure it is also what gives MDS its power. If a simple relation between the configuration space and the variable spaces were desired a decision would have to be made regarding what aspect of the original variable space to align the MDS axes with. By using iterative regression the focus remains on achieving the best global fit of the configuration space to the original data.

In order to present the results of a MDS analysis in a form that more closely resembles a PCA/PCoord result you can perform a PCA on the raw MDS coordinate values (see Fig. 4). In addition to looking more like a PCA result, this procedure also ensures that the MDS-PCA axes are uncorrelated with one another. If you use this secondary PCA procedure note that the eigenvalues refer to the amount of variance exhibited by the MDS configuration space scores, not the variance present in the original data. If the dimensionality of the PCA and MDS spaces is the same, the total percent variance express by these hybrid MDS-PCA axes will always be 100.

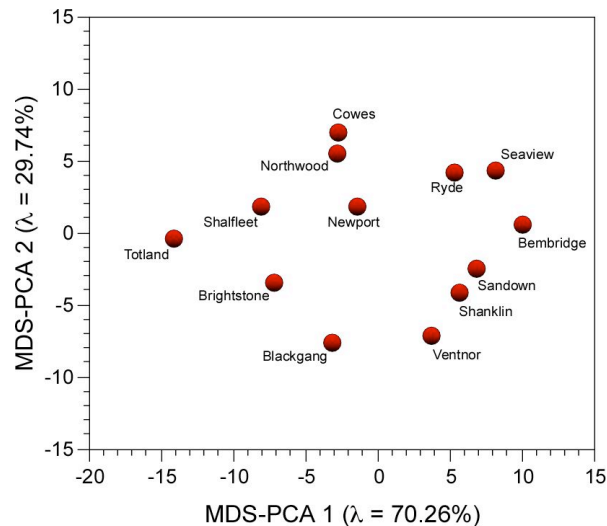


Figure 5. PCA-metric MDS configuration space ($k=2$) for the Isle of Wight location distance data.

Comparison of figures 2 and 5 shows some obvious distortions. Shanklin has been placed a bit too close to Sandown and Northwood is too close to the chord joining Cowes and Newport. To a large extent these distortions are only apparent, resulting from the fact that road distances were used to fix the relative locations instead of simple Euclidean distances. Of course, except for artificial map examples such as this we won't know what the MDS configuration for a set of data 'should' look like. Setting these issues aside, we can say that, to a very high degree, the MDS procedure was able to reconstruct a surprisingly accurate map of the our 13 location across the Isle of Wight from road distance data. But how accurate is this configuration, really?

Two accuracy measures are available in MDS. The first is the final value of the stress statistic used to assess the quality of the point configuration during configuration space adjustment. For the Isle of Wight data this value is 0.070. Since a perfect solution would have a stress coefficient of 0.0, and any stress value less than 0.1 is usually considered 'good', our metric MDS result would be considered 'good' if this had been an actual analysis. The point distribution certainly agrees well with a qualitative 'eyeball' test, comparing figures 2 and 5. But the stress statistic is a measure of overall distortion, akin to the cophenetic correlation coefficient of a cluster analysis. What the stress statistic doesn't do is to provide information about the degree to which our data match the expectations of the linear model we've used to refine the configuration or whether the model had trouble representing data from a particular range within the overall scaling problem.

These assessments can be made by looking at relations between the original distances (δ), the final configuration distances (d) and the disparities (\hat{d}). Graphs of these quantities are known as Shepard diagrams, after R. N. Shepard, one of the founders of modern MDS. Figure 6 shows the relationship used to assess the goodness of fit of the result in terms of the final configuration distances and the disparities.

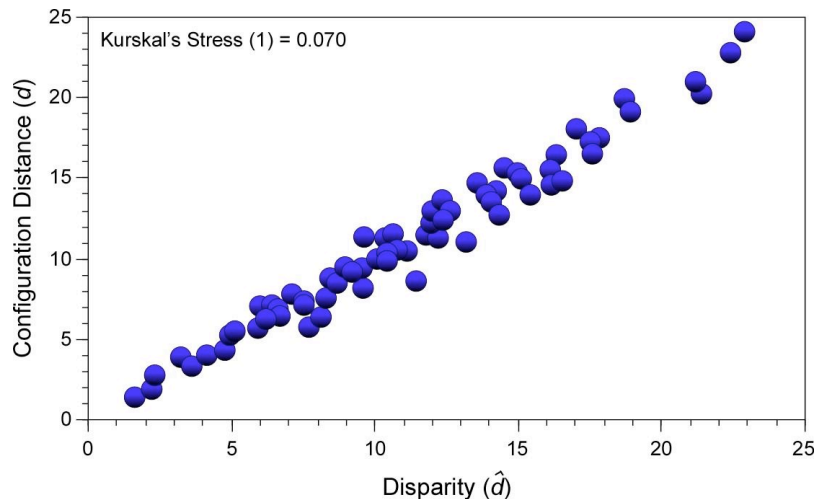


Figure 6. Shepard diagram for the Isle of Wight metric MDS result ($k=2$).

Note how well the MDS disparity predicts the final configuration distance. In a perfect analysis the predicted distance would agree perfectly with the configuration distance and all the points would fall along a straight line. Our metric MDS result isn't perfect, but it certainly provides evidence that a high-quality fit has been achieved across the entire range of scales present in the dataset.

Recall, the metric MDS model uses linear regression to guide its search for an optimal configuration (Fig. 3). This is an obvious choice, but it is not the only option available. Indeed, linear regression is one of the most restrictive choices we could make. We would use a metric MDS approach if we had high confidence in the quality and type of our original data. Certainly we do have confidence in the data type. The distances we've calculated are true metric distances. They are not, for example, 'distances' inferred from a set of ordinal variables whose quantitative relations to one another are unknown. We can use MDS to analyze ordinal data, but in that instance the metric MDS model would not be our best choice.

What about the quality of our data? Remember these are road distances, not Euclidean distances. On the Isle of Wight, road tracks are much denser on the eastern side of the island (see Google maps 2008). Also, as there are no bridges across the inlet extending from Cowes to Newport, this represents a natural barrier forcing routes between locations on the eastern and western sides of the island to detour through Newport, occasionally adding significant distance to the trip. As a result the road distance data represent an estimate of the distribution of locations across the island, but it's a biased estimate, strongly influenced by natural barriers to travel and idiosyncrasies in the island's transportation network. In more palaeontologically realistic examples, we often deal with proxy variables, variables that measure some quantity we are interested in, but only in an indirect manner. In such situations we might suspect there is substantial error in the data, so much so that we would be ill-advised to slavishly apply the metric-MDS (regression-based) model that treats the original distance data as an error-free standard. What to do?

Fortunately, there are a wide range of alternative models we can use with MDS to explore the configuration space. Full discussion of the ins and outs of all fitting models is well beyond the scope of this essay. However, one of the most commonly used, flexible and innovative 'non-metric model is isotonic (also called monotonic) regression.

Isotonic regression of two variables (say y on x) reorders both such that the x -values are uniformly increasing and then adjusts the values of y such that those values increase or remain constant relative to the x -values. This is accomplished by finding all pairs of y -values in which the second is smaller than the first and replacing both with their average. Figure 7 uses the same data analyzed in Figure 3 as an example of isotonic regression.

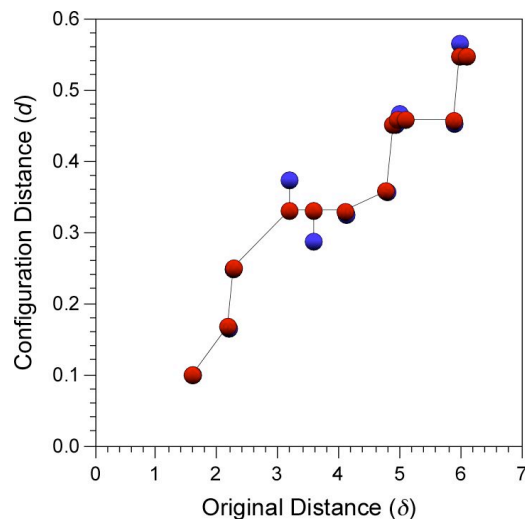


Figure 7. Example isotonic regression of MDS configuration distances ($k=2$) on the original distances for the first 13 between-location comparisons in the Isle of Wight data (see Table 1). Blue symbols represent raw data values for this fitting cycle. Red symbols represent disparity distances (\hat{d}) determined as a result of the regression. Regression residuals are indicative of the error associated with this fitting iteration. Note how much less error is generated by the isotonic, as opposed to the linear regression (see Fig. 3). This results from the more flexible, non-linear character of the isotonic model. Non-metric MDS using isotonic regression usually results in a better fit of the configuration space to the original data, irrespective of data type.

As can be seen from the figure, isotonic regression results in the specification of smaller deviations between configuration distances (d) and the disparities (\hat{d}), thereby minimizing the overall stress of the result. When isotonic regression is employed the technique is referred to as ‘non-metric multidimensional scaling’. Generally speaking, the greater flexibility of the isotonic regression approach will lead to less adjustment of the configuration space, and so lower stress values. When applied to the Isle of Wight location data the magnitude of this improvement can be appreciated readily (see figs 8 and 9).

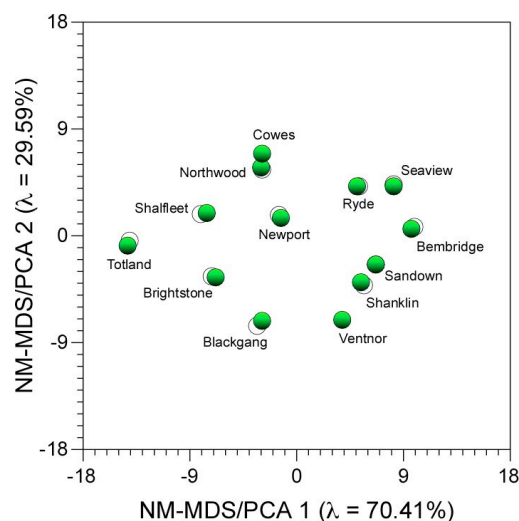


Figure 8. Non-metric MDS configuration space ($k=2$) for the Isle of Wight location distance data. Closed symbols represent non-metric MDS configuration, open symbols represent metric MDS configuration (see Fig. 4). See text for discussion.

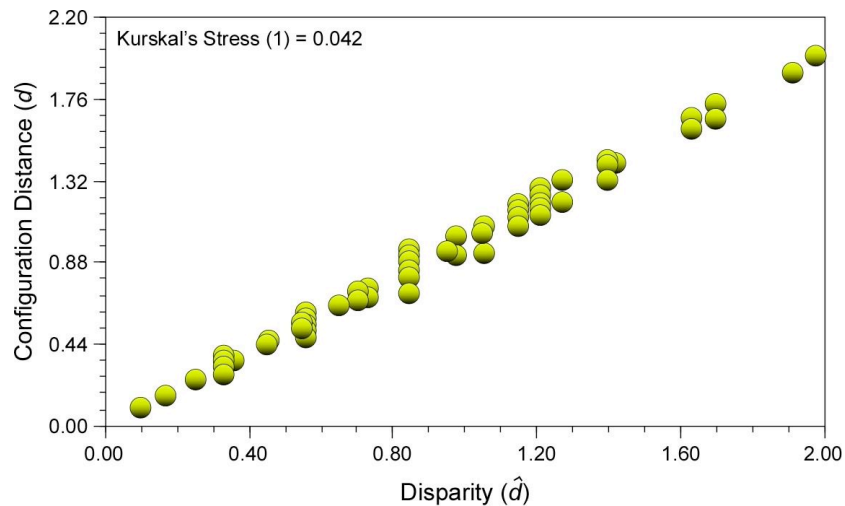


Figure 9. Shepard diagram for the Isle of Wight non-metric MDS result ($k=2$).

As can be seen clearly in Figure 8, use of isotonic regression to estimate the final MDS configuration space led to different relative positions in over half of the points in the Isle of Wight data. While these differences may not seem large, Figure 9 shows they reduced the amount of distortion in the overall result by almost 50 percent. Qualitatively speaking, this moves the result from the 'good' to the 'excellent' range. But even more importantly, use of the non-metric approach resulted in a better 'fit' between the data we have collected and the analytic approach we have chosen. This is the real goal we're after, to have the analysis capitalize on the strengths, and compensate for the weaknesses inherent in the data. The fact that this match yielded something close to the best possible result is satisfying, but getting a 'close-to-perfect' result isn't the point of data analysis. Understanding the system of observations is.

The example analysis result should not be taken to indicate that non-metric MDS is the correct choice for all MDS situations. Far from it. For example, if I had used straight-line map distances between Isle of Wight locations, the metric MDS approach would arguably have been more appropriate for those data. Nevertheless, non-metric MDS often represents the best choice for the greatest range of data we typically come across in palaeontology, provided our interest is confined to understanding the character of inter-object similarity.

Now we are in a position to apply MDS to our trilobite data. Table 2 represents the Euclidean distance matrix calculated from our three trilobite variables.

Table 2. Euclidean distance matrix for trilobite data

| Genera | <i>Acaste</i> | <i>Balizoma</i> | <i>Calymene</i> | <i>Ceraurus</i> | <i>Cheirurus</i> | <i>Cybantyx</i> | <i>Cybeloides</i> | <i>Dalmanites</i> | <i>Deiphon</i> | <i>Ormathops</i> |
|--------------------|---------------|-----------------|-----------------|-----------------|------------------|-----------------|-------------------|-------------------|----------------|------------------|
| <i>Acaste</i> | 0.00 | | | | | | | | | |
| <i>Balizoma</i> | 8.84 | 0.00 | | | | | | | | |
| <i>Calymene</i> | 30.30 | 38.58 | 0.00 | | | | | | | |
| <i>Ceraurus</i> | 2.60 | 6.92 | 31.70 | 0.00 | | | | | | |
| <i>Cheirurus</i> | 13.32 | 19.92 | 20.06 | 13.67 | 0.00 | | | | | |
| <i>Cybantyx</i> | 16.99 | 24.42 | 14.90 | 17.78 | 5.82 | 0.00 | | | | |
| <i>Cybeloides</i> | 4.64 | 11.40 | 27.23 | 4.75 | 8.97 | 13.11 | 0.00 | | | |
| <i>Dalmanites</i> | 11.22 | 19.25 | 19.48 | 12.38 | 6.21 | 6.29 | 8.11 | 0.00 | | |
| <i>Deiphon</i> | 6.40 | 9.44 | 30.19 | 4.81 | 10.68 | 15.68 | 4.01 | 11.60 | 0.00 | |
| <i>Ormathops</i> | 9.40 | 1.18 | 38.79 | 7.28 | 19.95 | 24.47 | 11.59 | 19.43 | 9.39 | 0.00 |
| <i>Phacopidina</i> | 4.95 | 8.20 | 30.76 | 3.00 | 11.83 | 16.31 | 3.75 | 11.61 | 2.25 | 8.19 |

| | | | | | | | | | | |
|----------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| <i>Phacops</i> | 6.29 | 13.62 | 25.22 | 7.03 | 7.20 | 11.49 | 2.75 | 6.85 | 5.72 | 13.90 |
| <i>Placoparia</i> | 16.87 | 24.88 | 13.77 | 18.04 | 7.26 | 2.74 | 13.50 | 5.92 | 16.53 | 25.04 |
| <i>Prilocylopyge</i> | 22.46 | 29.42 | 12.48 | 23.02 | 10.14 | 5.69 | 18.34 | 11.90 | 20.39 | 29.35 |
| <i>Ptychoparia</i> | 40.30 | 48.78 | 10.75 | 41.87 | 30.76 | 25.41 | 37.55 | 29.60 | 40.71 | 49.02 |
| <i>Rhenops</i> | 37.46 | 45.16 | 9.44 | 38.47 | 26.08 | 20.82 | 33.88 | 26.26 | 36.44 | 45.18 |
| <i>Sphaerexochus</i> | 0.91 | 9.01 | 29.88 | 2.41 | 12.55 | 16.40 | 3.83 | 10.77 | 5.58 | 9.51 |
| <i>Toxochasmops</i> | 24.66 | 32.90 | 6.26 | 26.06 | 14.44 | 9.93 | 21.57 | 14.23 | 24.46 | 33.16 |
| <i>Trimerus</i> | 71.39 | 79.47 | 41.13 | 72.66 | 60.07 | 55.13 | 68.07 | 60.39 | 70.66 | 79.58 |
| <i>Zacanthoides</i> | 27.89 | 35.75 | 4.75 | 28.99 | 16.70 | 11.42 | 24.38 | 16.80 | 27.05 | 35.84 |

| | <i>Phacopidina</i> | <i>Phacops</i> | <i>Placoparia</i> | <i>Prilocylopyge</i> | <i>Ptychoparia</i> | <i>Rhenops</i> | <i>Sphaerexochus</i> | <i>Toxochasmops</i> | <i>Trimerus</i> | <i>Zacanthoides</i> |
|----------------------|--------------------|----------------|-------------------|----------------------|--------------------|----------------|----------------------|---------------------|-----------------|---------------------|
| <i>Phacopidina</i> | 0.00 | | | | | | | | | |
| <i>Phacops</i> | 6.22 | 0.00 | | | | | | | | |
| <i>Placoparia</i> | 17.00 | 11.67 | 0.00 | | | | | | | |
| <i>Prilocylopyge</i> | 21.22 | 16.81 | 7.29 | 0.00 | | | | | | |
| <i>Ptychoparia</i> | 41.12 | 35.62 | 24.19 | 22.63 | 0.00 | | | | | |
| <i>Rhenops</i> | 37.07 | 32.18 | 20.69 | 16.33 | 10.18 | 0.00 | | | | |
| <i>Sphaerexochus</i> | 4.30 | 5.52 | 16.37 | 21.83 | 39.97 | 36.97 | 0.00 | | | |
| <i>Toxochasmops</i> | 25.15 | 19.37 | 8.51 | 9.23 | 16.79 | 14.73 | 24.19 | 0.00 | | |
| <i>Trimerus</i> | 71.43 | 66.07 | 54.62 | 50.72 | 32.04 | 34.78 | 70.94 | 46.94 | 0.00 | |
| <i>Zacanthoides</i> | 27.71 | 22.53 | 11.03 | 8.00 | 14.66 | 9.81 | 27.39 | 5.70 | 43.74 | 0.00 |

These are the same data we used to illustrate PCoord analysis and Q-mode factor analysis (see Newsletter 61). There, our goal was to represent inter-object similarity relations using eigenanalysis-based approaches, which aligned the PCoord and factor spaces with the directions of greatest distance (= dissimilarity) across these data as a whole. This time, I'll use MDS to focus the analysis on the more general question of simply representing inter-object similarity relations in a low-dimensional space.

Because these distances are based on a small number of variables ($m = 3$), I'm going to draw a flat map of between-object similarity relations by setting dimensionality to 2. What about the accuracy of the variables? Here it gets interesting. In order to obtain the body length, glabellar length, and glabellar width distances I had to select landmark points on the trilobite images I used to obtain the original data. I then measured the Euclidean distances between those points. So far so good. But what about the landmark points themselves?

If I'm going to regard my distances as being correct I'll need to assume there's no parallax in the images. This is certainly incorrect for some if not all. While I have no way of quantifying the amount of parallax-based distortion, the most reasonable assumption would be that parallax varies across the image set. Also, in order to ensure strict comparability of the distances I'd have to know that I always selected the same points on all the specimens. While I'm confident I've selected approximately the same points, I can't guarantee I have. I also don't quite understand what 'same' means in this context. From a geometric point-of-view the points would need to correspond topologically whereas, from a biological point-of-view they would need to correspond functionally and/or represent homologous locations. All these definitional criteria are valid possibilities, and none need to be the same point. The fact of the matter is, in all the confusion of actually collecting the data, I did my best to select comparable points, but can't really be certain which points I selected, why, or whether another person collecting the same data would select exactly the same points and get exactly the same distance matrix. Given this, the situation is looking suspiciously close to the Isle of Wight road distance data; metric data with varying levels of inconsistency and approximation

arising from multiple factors. To be safe, it's probably better to use a non-metric MDS approach.

Figure 10 shows the results of a non-metric MDS analysis of the Table 2 data ($k = 2$). On the same plot I've shown the projected positions of the PCoord scores of these same data from the *Mind Your Rs and Qs* column (Newsletter 61) to illustrate the similarities and differences between the two approaches.

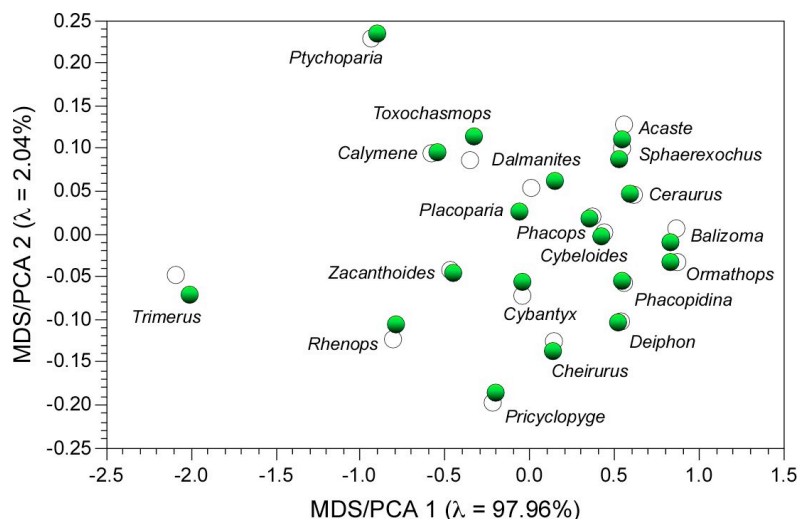


Figure 10. Non-metric MDS configuration space ($k=2$) for the trilobite distance data (Table 2). Closed symbols represent non-metric MDS configuration, open symbols represent projected positions PCoord scores for these same data configuration. See text for discussion.

Note how the relative positions of virtually all objects have been adjusted by the MDS analysis, some only slightly, but others substantially. For this particular dataset it's unlikely that the interpretation of the PCoord and MDS results would be very different. But recall this is a small and very well-behaved dataset. Differences between results for larger, more complex data could be much greater, certainly large enough to make a difference in their detailed interpretation, possibly sufficient to make a difference to general interpretation.

What about distortion? The Shepard plot for the trilobite non-metric MDS results is shown in Figure 11.

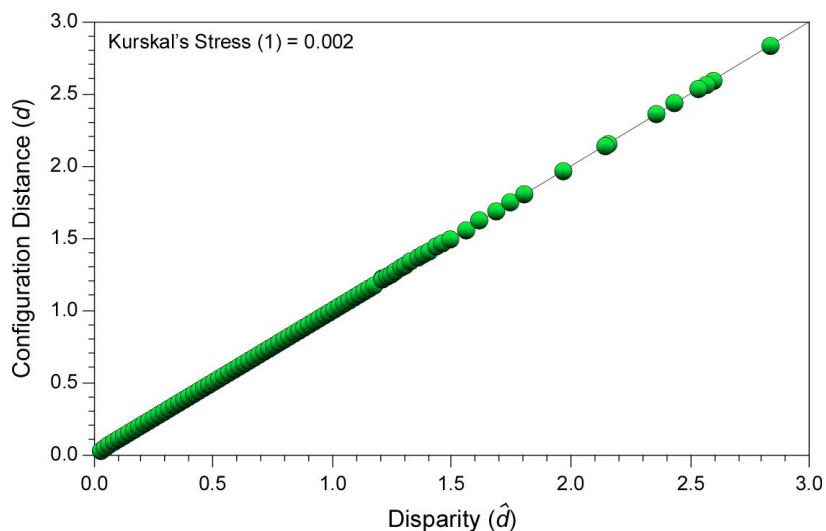


Figure 11. Shepard diagram for the trilobite non-metric MDS result ($k=2$).

The very low value of Kruskal's Stress (1) index and the high degree of conformance between the configuration distance and disparity across the entire range of the distance data provide direct confirmation of the high quality of the MDS result. For comparison the stress (1) value for the PCoord result can also be calculated from the reproduced distance matrix. That value (0.103), while acceptable (see above), is markedly suboptimal with respect to the non-metric MDS result.

Which distribution of points is correct? They both are! The relative point distributions are different because they are providing different information about the data: the PCoord result provides information about the distribution of points in a space aligned with the major dimensions of object dissimilarity. The MDS result provides information about the overall representation of inter-object distance at all scales. Both results have their place in data analysis strategies. But the really odd thing is, many data analysts interpret PCoord *and* PCA results—recalling that a PCoord analysis of a Euclidean distance matrix is the dual of a covariance-based PCA—as if they were absolutely accurate representations of overall inter-object distances. They are not the same and should not be described as such. The MDS approach delivers the globally optimised representation to a much greater degree than PCA/PCoord.

While I've presented MDS as a Q-mode method, it is possible to perform the analysis in the *r*-mode as well, though this is much less common. To do this the focus of the regression is the *r*-mode covariance or correlation matrix rather than the Q-mode distance matrix, but in all other respect the procedure, and the results in terms of minimizing distortion at all scales, are similar.

The origin of the MDS approach can be traced to the work Torgerson (1952, 1958), Shepard (1962, 1966) and Kruskal (1964a, b), Young (1970) and others with many of the important methodological improvements developed at Bell Laboratories (see Green et al. 1989). Despite the rarity of its application in systematic and palaeontological contexts (see Rohlf 1970 for an example), it is used routinely in many other fields, notably in social science, psychology, chemistry, and various fields related to economics, marketing, and advertising. Its use in these contexts is driven primarily by the need to analyze the datasets containing many 'state' variables (e.g., qualitative comparisons, customer survey results) in a PCA-like, non-hierarchical manner. The MDS approach is well-suited to such analyses, which of course are not uncommon in a wide range of physical and biological contexts, including palaeontology. But what I hope I've shown in this essay is that the advantages of MDS are much greater than simply being able to handle a wide range of data and produce PCA-like plots. The MDS approach focuses on a different, and somewhat more generalized question than PCA and other of eigenanalysis-based approaches focus on. It's a question that's commonly asked of data in our field.

In a larger sense though, all the methods I've discussed throughout this column can be thought of as 'multidimensional scaling' methods. Scaling, in its mathematical sense, refers to the act of representing some relation between objects on a numerical axis or scale. Any time we compare objects numerically using more than a single variable, we are engaging in an act of multidimensional scaling. A better term for the aspect of data analysis most authors refer to when they discuss scaling is 'ordination' (see Manley 1994). As a consequence of this somewhat inconvenient generality of the term scaling, the technical literature on MDS is quite complex with different authors drawing the boundary between MDS and 'not MDS' approaches at different places. Chatfield and Collins (1980) regard PCA as a type of MDS whereas Jackson (1991) does not. Both Jackson (1991) and Davis (2002), in turn, regard correspondence analysis as a form of MDS, whereas Pielou (1984) does not. Perhaps the best way to think of MDS is as the most generalized, and arguably the most accurate, of the set of scaling or ordination techniques. It certainly deserves to be much more widely used across the broad range of situations encountered in routine palaeontological data analysis.

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

REFERENCES

- Chatfield, C. and A. J. Collins.** 1980. *Introduction to multivariate analysis*. Chapman and Hall, London. 246 pp.
- Davis, J. C.** 2002. *Statistics and data analysis in geology (third edition)*. John Wiley and Sons, New York. 638 pp.
- de Leeuw, J.** 1977. *Applications of convex analysis to multidimensional scaling*. In J. R. Barra, F. Brodeau, G. Romier, and B. Van Custem, eds. *Recent developments in statistics*. Elsevier/North Holland, Amsterdam. 133–145 pp.
- Green, P. E., F. J. Camone, Jr., and S. M. Smith.** 1989. *Multidimensional scaling: concepts and applications*. Allyn and Bacon, Needham Heights, Massachusetts. 407 pp.
- Jackson, J. E.** 1991. *A user's guide to principal components*. John Wiley & Sons, New York. 592 pp.
- Kruskal, J. B.** 1964a. Multidimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis. *Psychometrika*, **29**, 1–27.
- Kruskal, J. B.** 1964b. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115–129.
- Manley, B. F. J.** 1994. *Multivariate statistical methods: a primer*. Chapman & Hall, Bury, St. Edmonds, Suffolk. 215 pp.
- Pielou, E. C.** 1984. *The interpretation of ecological data*. John Wiley & Sons, New York. 263 pp.
- Rohlf, F. J.** 1970. Adaptive hierarchical clustering schemes. *Systematic Zoology*, **19**, 58–82.
- Shepard, R. N.** 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, **27**, 125–140 (I); 219–246 (II).
- Shepard, R. N.** 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology*, **3**, 287–315.
- Torgerson, W. S.** 1952. Multidimensional scaling I: theory and method. *Psychometrika*, **17**, 401–419.
- Torgerson, W. S.** 1958. *Theory and methods of scaling*. John Wiley and Sons, Inc., New York. 460 pp.
- Young, F. W.** 1970. Nonmetric multidimensional scaling: recovery of metric information. *Psychometrika*, **46**, 357–388.

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