

## Multidimensional Scaling of some Lexicostatistical Data

Annette J. Dobson

*University of Newcastle, N.S.W., Australia*

Paul Black

*Australian Institute of Aboriginal Studies, Canberra City, A.C.T., Australia*

(Received: 25 October 1977. Revision received 8 December 1977)

### Abstract

Multidimensional scaling is a technique for representing relationships amongst objects geometrically by points arranged so that the distances between points correspond to empirical measures of the dissimilarities between the objects. The algorithm for non-metric multidimensional scaling is described briefly and the procedure is used to analyse some data about the lexical similarities amongst a group of Australian Aboriginal languages. The configuration obtained bears a striking resemblance to the geographical distribution of the tribes who spoke the languages.

### 1. Introduction

Multidimensional scaling is a useful method of data analysis when a spatial configuration is sought to summarise a data matrix consisting of dissimilarities (or similarities) between objects. If there are  $n$  objects (which may be people, species, nations, psychological stimuli, etc.), the data consist of  $\frac{1}{2}n(n-1)$  real numbers which are empirical measures of the similarity or dissimilarity between pairs of these objects. For definiteness, suppose that the data are dissimilarities and let  $\delta_{ij}$  be the value for the pair of objects  $i, j$  with  $\delta_{ij} = \delta_{ji}$  and  $\delta_{ii} = 0, i, j = 1, \dots, n$ .

If for every ordered triple  $(i, j, k)$  the data satisfy the triangle inequality  $\delta_{ij} \leq \delta_{ik} + \delta_{kj}$ , then the objects can be represented by points in  $(n-1)$ -dimensional Euclidean space,  $E_{n-1}$ , so that the  $\delta_{ij}$ 's are the distances between the points. If the data do not satisfy these constraints, the objects can still be represented by points in  $E_{n-1}$  but the distances  $d_{ij}(d_{ij} = d_{ji}; d_{ii} = 0; i, j = 1, \dots, n)$  between the points will only approximate the data values. The objects can also be represented by points in spaces of successively smaller dimensionality but the distances  $d_{ij}$  are likely to become increasingly poorer approximations to the data values  $\delta_{ij}$ . The aim of multidimensional scaling is to find a suitable configuration of points with the smallest dimensionality so that the fit between the distances and the actual data remains reasonably good.

For example, suppose that the objects are twenty cities and the  $\delta_{ij}$ 's are obtained by accurately measuring the (Euclidean) distances between the cities on a map. Given only these data and no information about what they represent, since they satisfy the triangle inequalities one could obtain a configuration of points in  $E_{19}$  exactly corresponding to

the data. However, using multidimensional scaling one should be able to find a configuration of points in a plane for which the distances also correspond exactly to the data. This two-dimensional configuration would in fact be a reconstruction of the relative positions of the cities, although it might be necessary to rotate or reflect it to obtain the proper geographical orientation.

Real data are of course unlikely to be exactly representable in  $E_2$ . Suppose that the data for the twenty cities are not the Euclidean distances but are the distances between them by road. To the extent that roads are designed to connect cities by the shortest distances, a two-dimensional configuration might still provide a reasonably close approximation to both the data and the geographical positions of the cities. Even if the network of roads is markedly affected by such factors as terrain and political boundaries, there could still be a configuration in two or perhaps three dimensions which fitted the data well and the shape of this configuration would relate to these other factors as well as to the geographical positioning of the cities.

## 2. Non-metric multidimensional scaling

The objective is to obtain a configuration of points  $x_i$ ,  $i = 1, \dots, n$  in  $E_t$  which matches the data well, is visually assimilable and can be meaningfully interpreted. If the data values  $\delta_{ij}$  are known fairly precisely and they all satisfy the triangle inequality, then a configuration can be sought which will minimise

$$\sum_{i < j} (\delta_{ij}^2 - d_{ij}^2);$$

this technique, called principal co-ordinate analysis by Gower (1966), is a popular way of obtaining spatial representations of ecological data. However, data obtained by psychologists are typically less precise and do not satisfy the necessary constraints. Therefore psychometricians have developed methods which depend only on the rank order of the  $\delta_{ij}$ 's and not their absolute values.

Torgerson (for example, (1958)) worked on 'classical' multidimensional scaling in which a configuration is obtained so that the distances  $d_{ij}$  between the points approximate as nearly as possible the numbers  $f(\delta_{ij})$  where  $f$  is some specified monotone function of the dissimilarities  $\delta_{ij}$ . A more powerful technique, commonly known as non-metric multidimensional scaling, was suggested by Shepard (1962a, b) who sought a configuration which yielded any monotone relationship between the  $\delta_{ij}$ 's and the  $d_{ij}$ 's. By optimising a condition of monotonicity, he showed that the rank order of the  $\delta_{ij}$ 's determined the positions of the points in Euclidean space (except for reflection, translation and rotation of the axes or uniform stretching or shrinking of the space). Kruskal (1964a, b) provided a rigorous logical foundation for the technique. He defined the *stress* of a configuration of  $n$  points  $x_i = (x_{i1}, \dots, x_{it})$ ,  $i = 1, \dots, n$  in  $t$  dimensions by

$$S = \left[ \frac{\sum_{i < j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i < j} d_{ij}^2} \right]^{1/2}$$

where the  $\hat{d}_{ij}$ 's are numbers which are chosen so that they have the same rank order as the dissimilarities  $\delta_{ij}$  and they minimise  $S$ . Then a best fitting configuration is one obtained by moving the points  $x_i$  around until  $S$  is as small as possible.

For fixed dimensionality  $t$  and an initial configuration which is obtained either randomly or on the basis of some prior knowledge, the algorithm is as follows.

*Step 1.* Calculate the distances  $d_{ij}$  for the configuration. These may be any metric of the form

$$d_{ij} = \left( \sum_{k=1}^t |x_{ik} - x_{jk}|^r \right)^{1/r} \quad (1)$$

where  $r \geq 1$  is the same throughout the computation, in particular, if  $r = 2$  the distances are Euclidean.

*Step 2.* Find numbers  $\hat{d}_{ij}$  which minimise  $\sum (d_{ij} - \hat{d}_{ij})^2$  subject to the constraints that  $\hat{d}_{ij} \leq \hat{d}_{hk}$  whenever  $\delta_{ij} < \delta_{hk}$  (this is the process sometimes known as monotone regression).

*Step 3.* If, in the usual sense of steepest descent methods, a minimum of the stress  $S$  has been attained the computation is complete. If not a new configuration is obtained by moving the points in directions chosen to reduce  $S$  and the iterative procedure is continued by returning to Step 1.

Various computer programs are available to implement this algorithm, for example MDSCAL and KYST produced by Bell Laboratories, TORSCA, SSA and others.

It is common to begin the analysis by obtaining a configuration in  $E_t$  where  $t$  is quite large, say 5 or 6. Then the points are projected into  $E_{t-1}$  and the resulting configuration is used as a starting point for a scaling in  $(t-1)$  dimensions, and so on. If the minimum stress obtained in  $E_t$  is plotted against increasing values of  $t$ , the stress is initially high but it decreases and the curve flattens out as  $t$  increases. Therefore if the 'natural dimensionality' of the data is unknown, the optimal value of  $t$  is selected as one beyond which further decreases in stress are relatively small.

The distances between the points in a configuration produced by this algorithm will correspond to the rank order of the dissimilarities even if the configuration is translated, rotated, reflected or uniformly stretched or shrunk. Therefore to obtain a meaningful substantive interpretation it may be necessary to view the configuration from various directions and orientations.

Since the algorithm depends on numerical optimisation, it is susceptible to problems of local minima. For any initial configuration it is quite possible that the final configuration may not be one for which the stress is globally minimal. Shepard (1974) in an over-view of non-metric multidimensional scaling, suggests that this problem is particularly serious if  $t = 1$  or if the distances are given by Equation (1) with  $r = 1$ , the city-block metric. (This problem does not occur with metric scaling techniques which involve latent root and latent vector calculations instead of steepest descent methods.) In practice, when the nature of the data dictates the use of non-metric scaling, it is advisable to use an initial configuration based on some prior knowledge of the data or to use several different random initial configurations in order to reduce the chance of unknowingly obtaining only a local minimum.

An example of the use of non-metric multidimensional scaling is in the analysis of lexicostatistical data. Using only comparisons between the vocabularies of languages it is often possible to obtain a spatial configuration which bears a striking resemblance to the actual geographical distribution of the languages.

### 3. Lexicostatistical data for some Australian aboriginal languages

Lexicostatistics involves measuring the similarity or dissimilarity between the vocabularies of pairs of related languages. The easiest and most commonly applied approach, discussed at length by Hymes (1960), begins with determining the words corresponding to one or two hundred relatively basic and universal meanings, such as 'tree', 'run' and 'hand'. For each pair of languages, the words having the same meaning are compared and judged by the linguist to be cognate (that is, descended from the same word in a shared ancestral language through a process of regular phonetic change) or not cognate. The proportion of cognates amongst the pairs of words is called the *lexicostatistical percentage* and it is taken as an index of the similarity between the languages.

Two distinct languages may have similar vocabularies either because they have inherited a high proportion of cognate words from an ancestral language (for example, as the Romance languages have from Latin) or because word borrowing has been extensive between contiguous or frequently interacting languages (the widespread literary influence of French is an obvious example). Suppose that the lexicostatistical percentages truly represent only inherited similarity, then if these data are analysed using hierarchical classification techniques (such as those reviewed by Cormack (1971)), the 'family tree' of the languages should be recovered. However if much of the lexical similarity is the effect of borrowing which the linguist has been unable to distinguish from true cognation, then the data will reflect this other, largely geographical influence. When languages are closely related both historically and geographically, it is difficult to judge, *a priori*, to what extent the relationships represented by the lexicostatistical data will be hierarchical or spatial. Black (1976) and others have shown how multidimensional scaling can provide a meaningful analysis of non-hierarchical linguistic relationships.

Table 1 shows Dixon's (1970) lexicostatistical percentages for ten Australian languages formerly spoken in the rainforest area southwest of Cairns, in northern Queensland. Dixon's determination of these percentages differs somewhat from the

Table 1. Lexicostatistical percentages

|          |       |         |      |         |         |          |         |         |          |
|----------|-------|---------|------|---------|---------|----------|---------|---------|----------|
| Dyabugay |       |         |      |         |         |          |         |         |          |
| 40       | Yidin |         |      |         |         |          |         |         |          |
| 14       | 27    | Ngadyan |      |         |         |          |         |         |          |
| 15       | 23    | 70      | Mamu |         |         |          |         |         |          |
| 15       | 22    | 62      | 87   | Dyirbal |         |          |         |         |          |
| 11       | 18    | 50      | 70   | 81      | Giramay |          |         |         |          |
| 5        | 12    | 30      | 47   | 53      | 60      | Wargamay |         |         |          |
| 9        | 11    | 13      | 21   | 23      | 24      | 30       | Nyawigi |         |          |
| 14       | 14    | 27      | 43   | 46      | 47      | 46       | 20      | Warungu |          |
| 9        | 16    | 15      | 17   | 18      | 15      | 9        | 8       | 13      | Mbabaram |

usual conventions. The list of meanings he used contained some items which are less 'basic' than those in the standard lists. Words for such meanings seem more prone to being borrowed by one language from another and borrowings are often difficult to distinguish from true cognates. In fact, Dixon based his calculation of the percentages on the numbers of related (cognate or borrowed) words rather than on cognates alone.

For these data, non-metric multidimensional scaling seems more appropriate than metric scaling methods. Although the scaling algorithm is described in Section 2 in terms of dissimilarities  $\delta_{ij}$ , the technique can equally well be applied to similarities  $\sigma_{ij}$  provided that the numbers  $\hat{d}_{ij}$  monotonically decrease with the  $\sigma_{ij}$ 's (instead of increasing with the  $\delta_{ij}$ 's). The data were analysed using MDSAL with this specification for descending monotone regression.

#### 4. Results

Figure 1, redrawn from Dixon (1970), is a map of the area showing the geographical distribution of the languages. Figure 2 is the configuration obtained by multidimensional scaling of the lexical data (suitably rotated to be comparable with Fig. 1). Since the purpose of this example is to compare the computed configuration

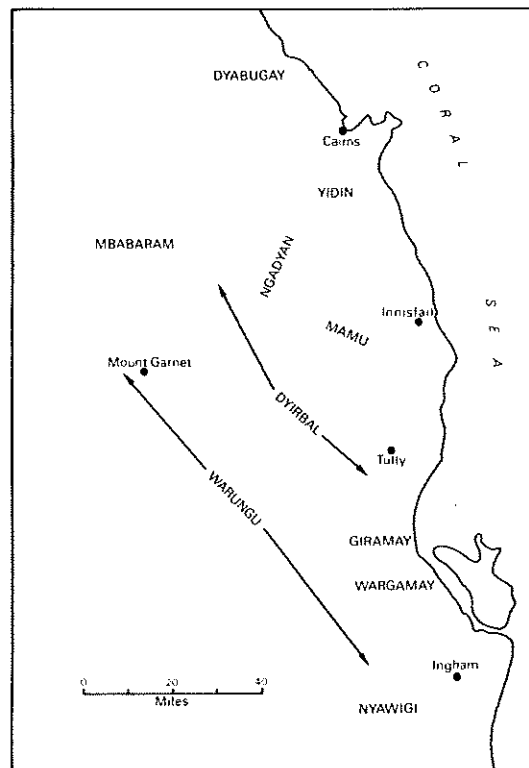


Fig. 1. Geographical distribution of the languages (based on Dixon (1970)).

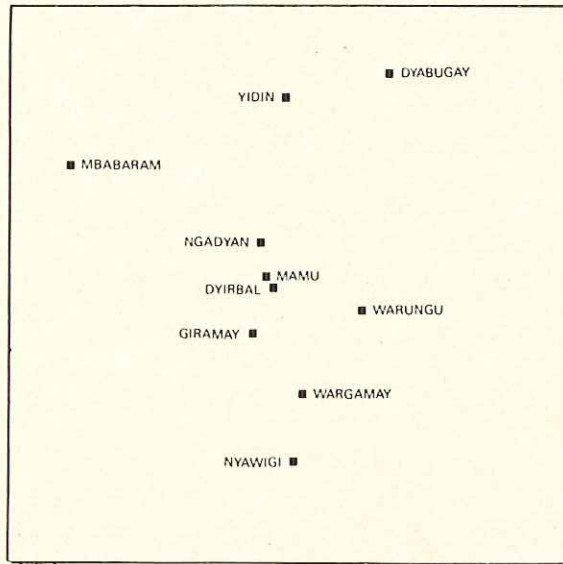


Fig. 2. Configuration produced by multidimensional scaling of data in Table 1.

with Fig. 1,  $t = 2$  dimensions were used, although the stress value of 10 per cent is typically regarded as only fair. (For these data, the best one-dimensional configuration has a stress value of 30 per cent whilst for the best three-dimensional configuration the value is less than 5 per cent which is usually regarded as satisfactory.) Nevertheless most of the languages in Fig. 2 are very near their actual geographical positions. The only serious exception is Warungu which is on the 'wrong side' of Dyirbal and Giramay, too far eastwards. (A configuration with Warungu in its 'correct' position is obtained from certain initial configurations but it is associated with a stress value of 12 per cent and so represents only a local minimum.)

In applying multidimensional scaling to four groups of languages in the Philippines, Africa and North America, Black (1976) also found that two-dimensional configurations produced from lexicostatistical percentages correspond well with the geographical location of each member of the groups. In that study the relationships best represented spatially, rather than hierarchically, appear to be associated with the process of dialectal differentiation (the localised geographical spread of linguistic innovations by which a language can gradually break up into dialects over a period of time).

It is likely that the same phenomenon occurs in the Australian data and that the relationships between the very similar varieties Ngadyan, Mamu, Dyirbal, Giramay, Wargamay and Warungu are largely of spatial origin, whilst the relationships between them and the other languages and amongst the other languages are more hierarchical in nature. However, since borrowing is usually heaviest between adjacent languages and Dixon's data consist of both borrowings and cognates, the spatial structure has probably been enhanced with the result that the two-dimensional configuration corresponds very closely to the geography.

## 5. Concluding remark

Recently Carroll (1976) has developed hybrid models for dissimilarity or similarity data in which hierarchical ('tree') structure and continuous spatial ('cline') structure are combined. Geometric representations of the data are produced using both clustering techniques and multidimensional scaling. Such analysis of lexicostatistical and other taxonomic data promise to provide useful insight into the nature of the evolutionary processes which gave rise to such data.

## Acknowledgements

The authors are grateful to Joseph B. Kruskal for introducing them to multidimensional scaling and encouraging their interests in lexicostatistics. They also thank R. M. W. Dixon for the use of his data.

## References

- BLACK, P. (1976) Multidimensional scaling applied to linguistic relationships. *Cahiers de l'Institut de Linguistique de Louvain* 3, 43-92.
- CARROLL, J. D. (1976) Spatial, non-spatial and hybrid models for scaling. *Psychometrika* 41, 439-463.
- CORMACK, R. M. (1971) A review of classification. *J. R. Statist. Soc. A* 134, 321-367.
- DIXON, R. M. W. (1970) Languages in the Cairns rainforest region. *Pacific Linguistics* 13, 651-687.
- GOWER, J. C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53, 325-338.
- HYMES, D. H. (1960) Lexicostatistics so far. *Current Anthropology* 1, 3-44.
- KRUSKAL, J. B. (1964a) Multidimensional scaling by optimising goodness of fit to a nonmetric hypothesis. *Psychometrika* 29, 1-27.
- KRUSKAL, J. B. (1964b) Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29, 115-129.
- SHEPARD, R. N. (1962a) The analysis of proximities: multidimensional scaling with an unknown distance function I. *Psychometrika* 27, 125-140.
- SHEPARD, R. N. (1962b) The analysis of proximities: multidimensional scaling with an unknown distance function II. *Psychometrika* 27, 219-246.
- SHEPARD, R. N. (1974) Representation of structure in similarity data: problems and prospects. *Psychometrika* 39, 373-421.
- TORGERSON, W. S. (1958). *Theory and Methods of Scaling*. Wiley, New York.

