MODEL STUDIES IN SERIATION TECHNIQUES

Ian Graham
Patricia Galloway
Irwin Scollar

Rheinisches Landesmuseum
Bonn

In this paper we wish to report the results of a comparative study of computer seriation methods. In an attempt to avoid the circularity inherent in the use of real data for the testing of such methods, we chose to devise a method for generating a variety of model data for the basis of our study. Because cemeteries present ideal qualities for seriation, we have chosen to make our model in terms of graves and artifact types. We have concerned ourselves solely with those methods which seek to produce a chronological ordering from the information contained in the incidence matrix. The following methods have been studied:

Algorithms which operate directly on the incidence matrix

AXIS: A program written by Wilkinson (1974) and based upon an algorithm suggested by Goldmann (1971), AXIS offers a simple and intuitively obvious method of four steps.

(1) Calculate the mean positions of the 1's in the columns of the incidence matrix.
(2) Order the columns according to these means.
(3) Calculate the mean positions for the 1's in the rows.
(4) Order the rows according to these means.

This cycle of operations is repeated until no move is made in either of steps 2 or 4.

POLISH: A program also written by Wilkinson (1974). POLISH attempts to maximise a score over the columns of the incidence matrix. This score represents the concentration of the 1's in the columns and has the form

$$\text{Score} = \sum_i V(R_i - N_i) - V(R_i).$$

R_i is the range of the 1's in Column i
N_i is the number of 1's in column i
V is the logarithmic gamma function

This score is derived by the maximum likelihood principle from the assumption that all graves are intrinsically equally rich.

Algorithms which operate upon a similarity matrix

From the incidence matrix it is possible to calculate many different matrices which represent in some way the similarity between pairs of graves. Of these we have considered only two. The first is a simple matching coefficient which expresses the number of types the two graves have in common. The second attempts to compensate for differences in grave richness by dividing each value of the simple matching coefficient by the total number of types in the two graves (Sokal and Sneath, 1963). These two similarity coefficients utilise only the links between pairs of graves. In an attempt to make use of the information provided by links among three or more graves, various authors have proposed modifications to the similarity matrix (Kendall, 1971; Sibson, 1972; Wilkinson, 1974). We have explored those suggested by Kendall and Sibson.

Kendall defines what he calls the 'circle product' of the similarity matrix by the following equations:

$$(SoS)_{ij} = \sum_k \min (Sik,Sjk)$$

SoS is the new similarity matrix
S is the original similarity matrix
i and j label two rows in the similarity matrix
k labels columns in this matrix

What we call here the 'Sibson transformation' is a slightly modified form of the Kendall rank correlation co-efficient (Kendall's τ) between pairs of rows in the similarity matrix. The method yields a matrix whose elements are computed in the following way. For each pair of rows i and j in the original similarity matrix, we consider all the possible pairs of elements (Sik, Sil) (Sjk, Sjl) where k and l are column indices. For each agreement in the rankings of these pairs, we add 1 to the i,j element of the transformed matrix. ('Agreement in ranking' refers to the situation where <u>either</u> Sik>Sil and Sjk>Sjl <u>or</u> Sik<Sil and Sjk<Sjl).

We have considered two scaling methods which operate upon the similarity matrix in all of the forms so far discussed.

MDSCAL: Written by Kruskal (1964a and 1964b), based on an algorithm by Shepard (1963), the multidimensional scaling program attempts to construct an n-dimensional configuration in which distances between points in the configuration are related to similarities in the source matrix by a monotonic-decreasing function. This program is available from the Bell Telephone Laboratories; we have used versions 5MS and 6MP.

LOCSCAL: This program, written by Wilkinson (1974), is very similar in principle to MDSCAL, except in that separate monotone regressions are performed for each row of the similarity matrix. The output from the LOCSCAL and MDSCAL programs can be obtained in the form of two- or three-dimensional configurations. The three-dimensional configurations may be plotted and displayed as stereo pairs, with desired links drawn between the plotted points. Normally we have obtained a seriation from both sorts of configuration by taking the order along the first principal axis.

<u>The model</u>

In view of the methods which we wished to test, our model was required to produce only an incidence matrix with a known chronological order. To make the model realistic, the incidence matrix must reproduce features seen in real incidence matrices. But from an unordered incidence matrix very little information is available; without a chronological order, we are limited to the examination of the variation of number of types in graves and number of occurrences of types. We must, of course, use this data in the building of our model. To make the model look real, however, we must hypothesise certain external information. For example, the number of times that an artifact type appears in a matrix is a function of both its commonness and its lifetime.

Variables in the model

The first requirement of a cemetery is dead bodies. We must, therefore, consider the variation of the population with time. On the suggestion of Mme. van Haeke of the Musée Royal in Brussels, we have tried to model Merovingian cemeteries, and we have taken as our population function a Gaussian with the form

$$e - \frac{(TIME-650)^2}{40,000}$$

This population function produced a peak at 650 A.D. between the arbitrary 400-800 A.D. limits that we set for all the models. None of these dates should be construed in any way as actual historic dates. For each model, dates for the graves were selected randomly in the interval between 400 and 800, with a time density proportional to the population.

The second variable is grave diversity, the relative number of objects liable to appear in a given grave. The distribution of diversity over the graves may be estimated from an unordered incidence matrix. We have also generated models in which all graves have equal diversity.

For artifacts the situation is a little more complicated, since an artifact type does not represent a point in time, as a grave may be considered to do, but rather a range of years over which it was in use. For simplicity's sake, we have assumed that the probability of an artifact type appearing in a grave is constant during the lifetime of the artifact and zero beyond these limits. We distribute the centre dates of the artifact types randomly between a time approximately equal to half the longest artifact lifetime before the beginning of the cemetery and half the longest artifact lifetime after the end of the cemetery.

In real cemeteries, the number of times that an artifact type appears is determined by the product of its lifetime and its commonness. We may estimate lifetime and commonness either separately or by estimating their product function. We have adopted both of these procedures in different models.

The model program

The model program is a simple FORTRAN IV program in which the different variables are computed by FORTRAN function subroutines. The program proceeds through each of the variables described above, setting up a table for each. Taking the graves one at a time, the program searches to find those artifact types whoch occur at the time the grave was made. A probability for the artifact's appearance in the grave is then calculated from the formula

$$P_{ij} = W D_i C_j$$

P_{ij} is the probability of the j^{th} artifact appearing in the i^{th} grave
D_i is the diversity of grave i
C_j is the commonness of artifact type j
W is a constant weighting factor

	CLOVIS	CLOTHAR	DAGOBERT	GUNTRAM
POPULATION	$e^{\frac{-(CDATE-TIME)^2}{D^2}}$	unchanged	unchanged	unchanged
GRAVE DATES	same function as above, randomly sampled	unchanged	unchanged	unchanged
GRAVE DIVERSITY	spike function (all graves have equal diversity)	unchanged	piecewise linear: 1 at 0 to .2 at 40, with linear decrease to 0 at 200	unchanged
ARTIFACT COMMONNESS	spike function (all types are equally common)	derived	derived	derived
ARTIFACT LIFETIME	piecewise linear: 0 at 0 to 1 at 20, with linear decrease from 1 to 0 between 80 and 240.	unchanged	unchanged	doubled
ARTIFACT PRODUCT	not used	2^{-K} (20) K = 1 to 200	unchanged	2^{-K} (40) K = 1 to 200
ARTIFACT CENTRE DATE	linearly random between 300-900	unchanged	unchanged	linearly random between 200-1000
MATRIX SIZE	57 X 48	47 X 44	56 X 73	39 X 60

TABLE I

A random number between 0 and 1 is generated. If this number is less than the probability P_{ij}, the artifact type is assumed to appear in the grave. When each of the possible artifact types has been considered for a particular grave, the artifact type codes which appear in that grave are written out in a form suitable for input to the seriation programs.

The model cemeteries

Out of sixty-eight models generated by the above program, we chose four to investigate in detail. These models were named after Merovingian kings, and detailed information on the functions used in their generation is given in Table 1.

CLOVIS: The simplest of the four models, Clovis has equal diversity for all graves and equal commonness for all artifact types.

CLOTHAR: For this model, artifact commonness was allowed to vary through the introduction of the product function.

DAGOBERT: In Dagobert, both grave diversity and artifact commonness were varied.

GUNTRAM: Similar to Dagobert, Guntram differs by the fact the artifact lifetimes were doubled.

Experimental procedures

The data for each model was run through our READIN program, which removes from the incidence matrix all those graves which contain less than two types and all those types which occur less than twice in the matrix. The output of this program is in column- and row-coded form (Kendall, 1971). Ten seriations, each starting with a different random order, were made with AXIS. These seriations were then used as starting orders for ten runs of POLISH. As a comparison, POLISH was run on one hundred random starting orders and on the correct order as it came from the model program. MDSCAL and LOCSCAL were run in two dimensions on similarity matrices generated with both similarity coefficients and on matrices with up to three applications of the Kendall circle product or the Sibson transformation. Seriations were produced from all of the two-dimensional configurations by projection onto the first principal axis. For all seriations, the Kendall rank correlation with the correct order was calculated.

Results

We were pleasantly surprised at the robustness of the methods studied, but our use of model data, with its corresponding control over the actual order of the graves, made it possible for us to observe some pitfalls.

AXIS: AXIS is capable of producing very good results. In ten random starts for each model, there was always one AXIS run which produced a Kendall correlation coefficient of over .8. Unfortunately it would be impossible with real data to tell which AXIS result was the best. The program can indeed produce seriations in which large segments are transposed or reversed. It is thus very dangerous to use AXIS on its own as a seriation method. It is a very fast program, and its best use is as a quick method for producing starting orders for other programs.

POLISH: POLISH seems to work very well when used with starting orders from AXIS. It tends to improve a good AXIS order, and the order with the highest POLISH score proved to have the best correlation coefficient with the correct order, in all the models except Guntram. POLISH can also be used to improve any other order, as from MDSCAL or LOCSCAL. POLISH on one hundred random starts was in no case better than the best POLISH on an AXIS order.

Of the two scaling methods tested, LOCSCAL produced results which were in general better than those from MDSCAL. More circle products than one tend to make the results worse, especially for MDSCAL. The two dimensional configurations in such cases often reduce to a tight group with a few scattered stragglers. Sibson transformations do not have the disastrous effects of circle products, but neither do they improve matters substantially. The most consistent results that we had from similarity matrix algorithms were obtained with the use of local scaling on the original similarity matrix.

Even these results were not so good generally as those obtained from POLISH on AXIS orders. In the two cases where we tried POLISH on MDSCAL output as an experiment, the seriations were improved, and the final results were also better than the AXIS plus POLISH results for the same model. This technique is obviously worthy of further exploration.

Afterthoughts

Our study has been necessarily limited by the availability of program and computer time. The project did in fact consume over three hours of c.p.u. time on an IBM 370/168. We have been able to examine only a limited number of models, but our results have suggested that a model study would be a desirable preliminary to any large and important seriation project. We have also found the most satisfactory seriation method for our problem.

References:

Goldman, K. 1971 — Some archaeological criteria for chronological seriation. in MAHS 396-400

Kendall, D.G. 1971 — A mathematical approach to seriation. Phil.Trans.Roy.Soc. London (A), 269, 125-135.
and Seriation from abundance matrices. in MAHS 215-252.

Kruskal, J.B. 1964 — Multi-dimensional scaling by optimizing goodness of fit to a non-metric hypothesis. Psychometrika 29, 1-27
and Non-metric multi-dimensional scaling: A numerical method. Psychometrika, 29,115-129

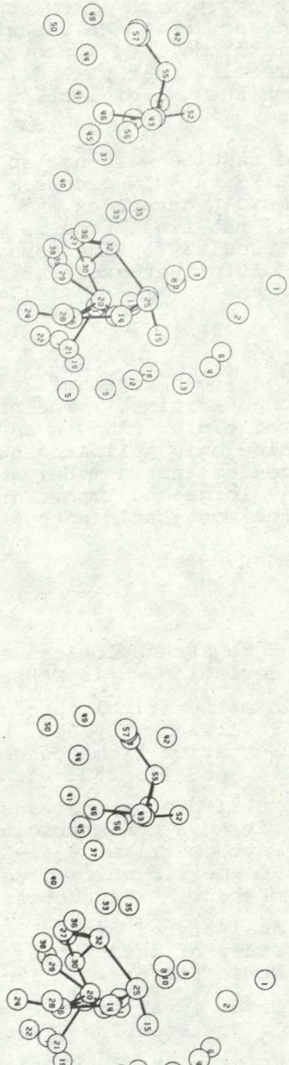
Shepard, R.N. 1963 — Analysis of proximities as a technique for the study of information processing in man. Human Factors, 5, 19-34

Sibson, R. 1972 — Order invariant methods for data analysis. J. Roy.Stat.Soc., Series B, 34, 311-349.

Sokal, R.R. Sneath, P.H.A. 1963 — Principles of numerical taxonomy. Freeman: San Francisco and London.

Wilkinson 1974 — Techniques of data analysis, Seriation theory. Archaeo-Physika, Bonn, 5.

MAHS is: Mathematics in the archaeological and historical sciences. Edinburgh University Press.



CLOVIS: stereo pair. MDSCAL with no circle product and 3-D display with ORTEP.