PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

A Mathematical Approach to Seriation

D. G. Kendall

Phil. Trans. R. Soc. Lond. A 1970 **269**, 125-134
doi: 10.1098/rsta.1970.0091

Phil. Trans. Roy. Soc. Lond. A. **269**, 125–135 (1970) [125]
Printed in Great Britain

A mathematical approach to seriation

By D. G. Kendall, F.R.S.
Statistical Laboratory, University of Cambridge

We consider the problem of seriating tombs in a cemetery, using only the presence or absence (in each tomb) of objects carrying traits believed to be chronologically significant. This information can be summarized in the form of a matrix A of zeros (indicating absence) and ones (indicating presence), where each row of the matrix is associated with a tomb, and each column with a 'variety'. A mathematical investigation suggests that much (perhaps all) of the information relevant to seriation is stored in the matrix product AA', and accordingly this similarity-matrix is taken as the starting-point of a multidimensional scaling procedure. Experiments with artificial data arising from a linear structure show that a two-dimensional output in the form of a *horse-shoe* is to be expected. If we then read off the labels of the points representing the graves, starting at one 'vertex' of the horseshoe and working round it till the other is reached, we obtain an estimate of the serial order (or of the serial order reversed).

1. Introduction

Many of the papers presented to this Symposium have been concerned with questions of absolute chronology, but here we shall look at the less demanding matter of *seriation*; presented with an assemblage of prehistoric tombs we should like to know their correct temporal order, but we do *not* press for an absolute date for the assemblage, and we do *not* expect a shift of say ten places in the temporal order to represent a lapse of the same amount of time at all stages of the series.

It is clear that an absolute chronology cannot be constructed without some reference to external events (for example, changes in the Earth's magnetic field), and so it is not surprising that the natural sciences should be of direct assistance there. But seriation (as was realized by Flinders Petrie 1899) is, at least in some cases, possible by the use of internal evidence alone, and here the assistance provided by the natural sciences is only indirect (through the development of the high-speed computers which are necessary to perform the calculations).

The basis of 'internal' seriation methods is the crude but extremely powerful principle brought into action by Petrie in the context of Egyptian prehistory: *if we have a good chronologically significant typology for the objects found in the tombs, then the extent to which the contents of two tombs are similar will be an indication of how close together they should be placed in the temporal series.*

The 'if' here is rather a big one; how to construct a 'good' typology, and how to assess its 'chronological significance', are serious and difficult questions indeed, but they are primarily questions for the archaeologist rather than the mathematician, and we shall therefore not enter further into these matters beyond remarking that mathematical principles and computer practice do have some contributions to make to the problem of constructing a typology.

If a 'good' and 'chronologically significant' typology is taken as given, then we may summarize the internal evidence relevant to the matter of seriation by constructing a table of double entry, or *incidence matrix* as it is sometimes called, in the following way. The table is to have as many rows as there are tombs, and each row will be associated with a unique tomb; it is to have as many columns as there are 'varieties' in the typology, and each column will be associated with a unique variety. If we name the ith tomb and the jth variety this will uniquely determine a cell in the table, and the entry there (the (i, j)th element of the incidence matrix) is set equal to 1 (0) according to the excavation record; 1 means 'jth variety present in the ith tomb', and 0 means 'absent'.

Now suppose that by good luck the tombs happen to have been arranged in the correct chronological order; then we should expect the following pattern of 0s and 1s to be very roughly displayed by the incidence matrix for the excavation:

Pattern P. In every column, if it does not consist entirely of 0s, the 1s are bunched together.

(The assertion made in this last paragraph in fact describes what we mean by a 'chronologically significant typology'. The typology may be a good one in other respects, and yet the assertion may be false: it is easy to think of mechanisms which would bring this state of affairs about. When it is feared that the assertion *is* false, then the following methods are not recommended: *caveat emptor*!)

All we have to do, then, in order to seriate the tombs, is to rearrange the rows (the rows but *not* the columns) of the table in such a way that an approximation to pattern P is brought about.

If the number of tombs (= rows) is small, say 6 or 7, then this may be a fairly simple task. but in practice the number of rows may be anything from the order of 50 (as in the example to be presented here) to 500 (as in the example studied by Petrie). The number of possible rearrangments for the rows is then so much more than astronomically vast that it is utterly impossible to test all the rearrangements, whether by hand, or with the aid of high-speed computers now available, or with the aid of the fastest computer logically thinkable, even if we were to run it for the conventionally accepted 'age of the universe'. The task is made the more formidable by the fact that we do not expect pattern P to be exactly attainable—if it were, and uniquely so, a short cut to the solution might then be possible.

In an earlier attack on this problem (Kendall 1963) I expressed it in statistical terms and showed how (at the cost of introducing a number of rather artificial assumptions) one could set up a 'maximum likelihood' procedure for estimating the desired rearrangement of the rows. This procedure associates with every possible permutation (the mathematician's word for 'rearrangement') a figure of demerit S, and the rule then is: *find the permutation for which* S *is least*. It is quite impossible to follow such a rule because one cannot examine the value of S for every permutation; there are just too many of them. A modification of the rule such as (1°) take a 'random walk' through the set of all permutations for so long as time allows, and accept the permutation thus encountered having the smallest S value, or (2°) proceed more systematically through the permutations, using some campanological algorithm, and choose the permutation as before, might be worth exploring; the report by Hole & Shaw (1967) would be of considerable assistance in such a venture. Here, however, we wish to describe an entirely different method of attack, and to show how it has been successfully applied to a full-scale archaeological problem. The details of the mathematical argument have been set out elsewhere (Kendall 1969*a*, *b*) and will only be summarized here. The full-scale application of the method has not been reported on before.

Before entering into a description of the method one final remark is needed, to the effect that no unique 'solution' is to be expected. In the first place, if any rearrangement yields a close approximation to pattern P then we have only to place the tombs in reverse order and we at once obtain another acceptable rearrangement. Usually, of course, one of two such mutually reverse rearrangements would at once be rejected on 'external grounds', say because it placed some 'obviously' early tomb at the late end of the series, etc. Secondly, and this is more important, no perfect realization of pattern P is to be expected, and therefore each 'good' rearrangement will have a large number of equally good 'neighbours', generated from it say by a few exchanges of consecutive rows. Of course such small displacements in the serial order will not be of archaeological significance, either, but some quite major displacements of particular tombs may also have little effect on the degree of success with which pattern P is attained, and the displacements may be such as to affect crucially the subsequent chronological decisions. Evidently therefore a method, to be really satisfactory, must allow for the possibility of more or less independent repeat analyses, so that we can see whether a surprising feature of one 'solution' is in fact common to all 'solutions', or at any rate, to most of them. We shall mention two ways in which this replication can be achieved (see §§4*b*, and 4*f* below), although there is evidently a limit beyond which one cannot go unless there is a further stock of data to draw upon.

2. The method

We first quote a fact about matrices of zeros and ones which was established elsewhere (Kendall 1969*a*). Suppose that the matrix A is such that at least one rearrangement of the rows will produce pattern P *exactly*. (We then say that we have in A a row-scrambled 'Petrie matrix', or equivalently that the given matrix A is *petrifiable*.) It can then be shown that all the information contained in A and relevant to the problem of row-rearranging A, in order to exhibit the pattern P, is contained in the derived matrix AA' whose (i, j)th element is

the number of varieties common to the ith and jth tombs.

(This new matrix is 'square', that is, it has as many rows as it has columns, and each row and each column alike is uniquely associated with a tomb.)

Now, according to Petrie's basic principle, the (i, j)th element of AA' will be large to the extent that it is reasonable to bring the ith and jth tombs close together in the serial order. In other words, in AA' we have a *similarity matrix for tombs*, where similarity means degree of closeness in the serial chronology. If therefore we know of any procedure which will recover a linear order for objects known to be capable of being linearly arranged, and for which we have available a similarity matrix of the type described, it will be natural to apply that procedure in the archaeological context using AA' as a starting-point. This is intuitively plausible in the archaeological situation when we do not expect exact petrifaction to be possible, and it is true in a much more substantial sense when A really is a row-scrambled Petrie matrix, for then AA' contains *all* the information relevant to a serial chronology.

Now such a procedure does exist: the multi-dimensional scaling (MDSCAL) program of Shepard and Kruskal (Kruskal 1964). This starts with a 'ranked'-similarity matrix and produces as output a geometrical configuration. To illustrate the idea in a simple case,[†] suppose that we wish to make a map of a country and that we have available information of the following sort: an automobile handbook supplies a table showing the distance in miles by the most convenient route linking each distinct pair from among N towns T_1, $T_2, \ldots, T_N$. We do not wish to use such detailed information, and so we extract from the given table the following simpler one:

between T_a and T_b the road distance is smallest;
between T_c and T_d the road distance is next-smallest;
... ...
between T_y and T_z the road distance is greatest.

[†] I am indebted to Mr A. D. McLaren for this example.

On being presented with the information contained in this new table, the MDSCAL program will compute the coordinates of the towns $T_1, T_2, \ldots, T_N$ in a synthesized map, starting from an entirely arbitrary (random) initial map, and will draw the final map on tracing paper. The synthesized map is constructed so that as far as possible the *ordering* of the town-pairs (T_a, T_b), $(T_c, T_d), \ldots, (T_y, T_z)$ is the same no matter whether we use the road distances recommended by the automobile handbook, or the direct distances on the synthesized map constructed by the computer. Notice that the computer does *not* know the actual handbook distances; it only knows the ordering contained in the second table, telling it which is the closest pair, which is the next closest pair, and so on, down to the most remote pair.

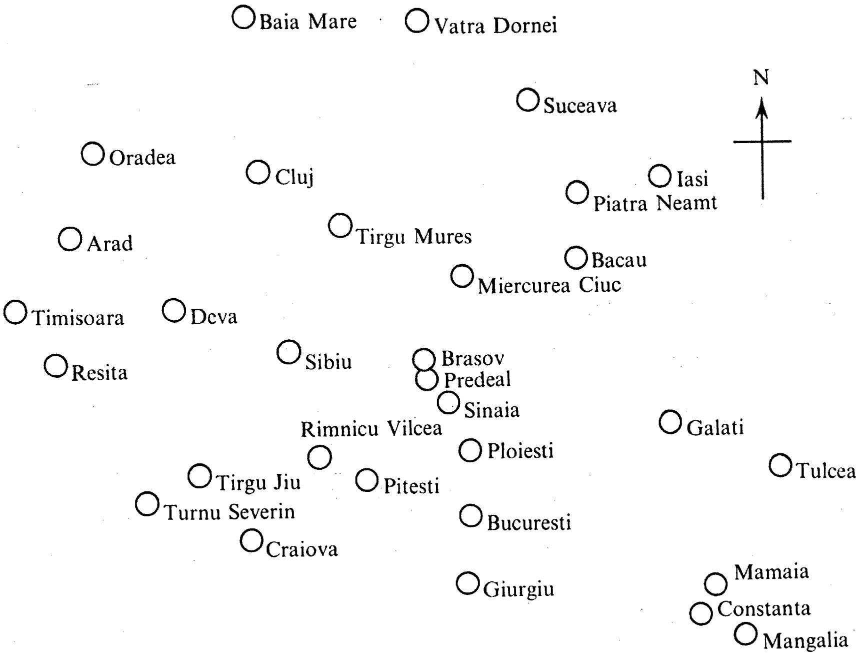

FIGURE 1. A MDSCAL map of Romania.

It is a matter of experience that the maps so constructed can be astonishingly accurate, even when the country is so mountainous that the most convenient road route is very far from being direct. For example, in figure 1 the reader will see a map of Romania constructed in this way, and he may care to reach for an atlas and form an opinion of its 'qualitative' accuracy.

Here we have a situation in which the objects (the towns) have an unknown (or, as here, suppressed) two-dimensional structure, and the computer reconstructs this from appropriate similarity data. But what happens if the underlying structure is one-dimensional, as in a chronological problem? We can best answer this question by a further illustration.

Let us take a system of 51 objects labelled by the numbers 1,2,...,51; we are to think of them as capable of being arranged along a straight line (which might for example represent the time-axis), the jth object being located at the point with coordinate j. We shall call the objects $O_1, O_2, \ldots, O_{51}$. If we told the computer the actual distance from each O_i to each O_j it would not be very surprising if it managed to recognize the linear character of the data and to produce a fair graphical representation of it, but we shall tell the computer much less than this; in fact we shall simply provide it with the following information.

The pairs of objects may be arranged into 9 groups as follows:

(i) the first group contains the most closely related pairs of objects: these are (O_1,O_2), (O_1,O_3), (O_1,O_4), (O_2,O_3), (O_2,O_4), $(O_2,O_5), \ldots, (O_{50},O_{51})$;

(ii) the second group contains the next most closely related pairs of objects: these are (O_1,O_5), (O_1, O_6), (O_1,O_7), (O_2,O_6), (O_2,O_7), $(O_2,O_8), \ldots, (O_{47},O_{51})$;

... ...

(viii) the eighth group contains the next-but-most-remote pairs of objects; these are (O_1,O_{23}), (O_1,O_{24}), (O_1,O_{25}), (O_2,O_{24}), (O_2,O_{25}), $(O_2,O_{26}), \ldots, (O_{29},O_{51})$;

(ix) the last group contains the most remotely related objects; these are (O_1,O_{26}), $(O_1,O_{27}), \ldots, (O_1,O_{51})$, (O_2,O_{27}), $(O_2,O_{28}), \ldots, (O_2,O_{51}), \ldots, (O_{26},O_{51})$.

In fact we know (but the computer does *not*) that the distances for the pairs in the first group are 1, 2, or 3; that those for the pairs in the second group are 4, 5, or 6, ...; that those for the pairs in the eighth group are 22, 23, or 24; and that those for the pairs in the ninth group are 25, 26, 27, ..., or 50 (this being the largest possible distance, namely that from O_1 to O_{51}). We are therefore making things harder for the computer in three different ways; we are concealing the actual distances; we are lumping together distances 1, 2, and 3, distances 4, 5, and 6, and so on: and we are lumping together as 'very large' *all* distances in excess of 24.

Again, *we* know that the label j of the object called O_j in fact specifies its position on the line. The computer is aware that O_j is labelled j (the objects have to retain distinct labels or they will become confused within the computer), but it does *not* know that j is a positional coordinate for O_j; it does not know this (i) because we did not tell it that this is so, and (ii) because it is too stupid to guess that it might be so.

It might be thought, then, that the computer really has very little information about the original structure of the system of objects presented to it. However we may regard the information supplied as a ranked-similarity matrix; we then call in the MDSCAL program, and allot an arbitrary (random) initial configuration, and after 45 iterations we obtain the plot shown in figure 2*a*, where the points representing the objects are given their original object-labels, and all object-pairs within the first group are linked by straight lines. It will be seen that the original order has been reconstructed very well, but the slightly surprising feature of the plot is its *horse-shoe* shape, which is even more striking after 5 further iterations (figure 2*b*). This is a consequence of our deliberate 'blurring' of the large distances, and has been found to be characteristic of such situations. We may conjecture, therefore, that if we were to repeat this procedure, *taking the tombs as 'objects' and telling the machine which pairs of tombs have most varieties in common, which pairs have the next largest number of varieties in common, and so on, until we come to the (very many) pairs of tombs that have the least number (zero) of varieties in common, then we would again obtain a horse-shoe output, and could obtain a serial chronology by reading off the tomb labels as one proceeds around the horse-shoe from one vertex to the other.* (Of course we should obtain either a chronology or an anti-chronology, according to the twofold choice of the vertex from which we start).

We have now formulated a method for obtaining a serial chronology, and it only remains to find a suitable test excavation and to put it into practice.

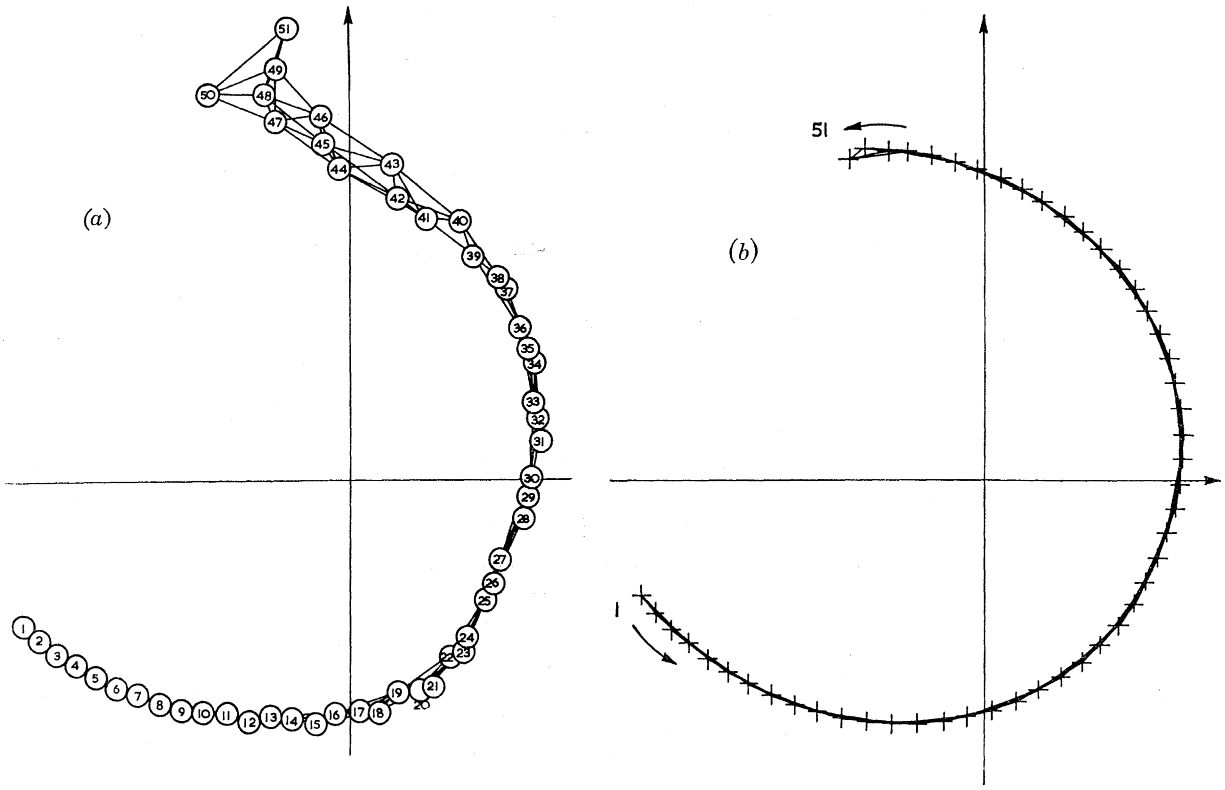

FIGURE 2. A MDSCAL reconstruction of strictly linear data; (*a*) 45 iterations, (*b*) 50 iterations.

3. THE APPLICATION

The method described in the preceding section was published (Kendall 1969*a*, *b*) before it had been tested on large-scale genuine data. It is now possible to report on its performance in constructing a serial chronology for the La Tène cemetery at Münsingen–Rain for which a very carefully worked-out typology of the fibulae, anklets, bracelets, etc. is available (Hodson 1968). Dr Hodson selected a set of 59 tombs for analysis, and supplied an incidence matrix A of 59 rows (tombs) scored for 70 columns (varieties). He had previously worked out a serial chronology himself, but this was concealed during the analysis by a random encoding of the tomb- and variety-labels. The possibility of subsequently comparing the computer seriation with that of Hodson was of course one of the motives for selecting this material for analysis. Another was the fact that the Münsingen–Rain cemetery is long and narrow, and that the interments there may very well have started at one end and proceeded along it; thus there is also the attractive possibility of comparing the computer's seriation with the serial ordering of the tombs along the major axis of the cemetery.

On computing AA' it was found that the pairs of tombs, when ranked according to the number of varieties held in common, were arranged as follows:

1 pair with 7 varieties in common;
4 pairs with 6 varieties in common;
7 pairs with 5 varieties in common;
7 pairs with 4 varieties in common;

38 pairs with 3 varieties in common;
93 pairs with 2 varieties in common;
247 pairs with 1 variety in common;
1314 pairs with no varieties in common.

This information, together with the assignment of each one of the 1711 pairs to the appropriate one of the 8 groups (0 varieties in common, 1 variety in common, ..., 7 varieties in common) was supplied to the computer as the effective input for the MDSCAL program, and *no further information than this was used in the seriation procedure*. The initial two-dimensional configuration for the 59 points representing the 59 tombs was prescribed by a randomizing mechanism, and the program was then run for 50 iterations. The resulting plot is shown in figure 3. Here each point (= tomb) is shown as a small circle containing a number-label. If two tombs hold 2 or more varieties in common they are regarded as 'strongly linked' and the points representing these tombs on the plot are linked by a straight line (drawn by the computer).

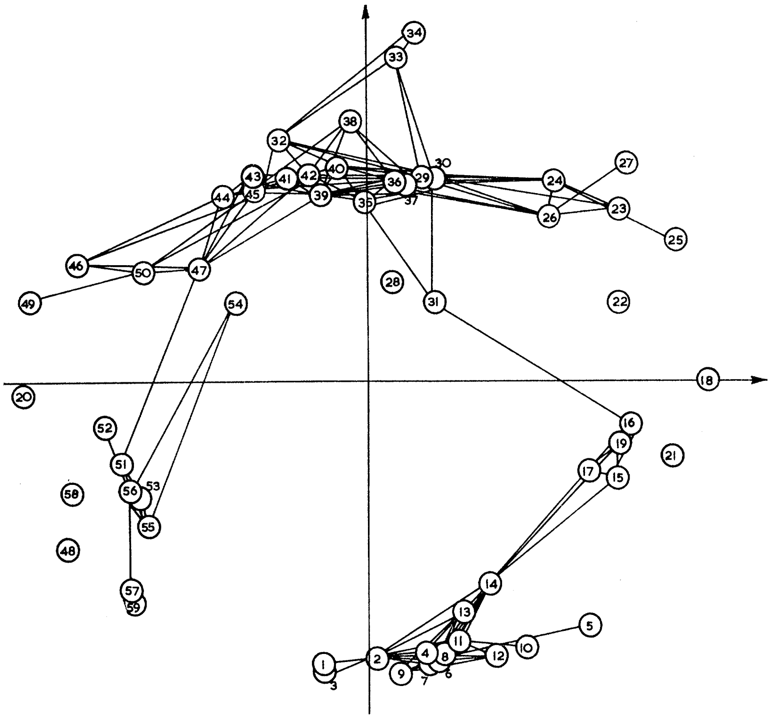

FIGURE 3. The MDSCAL output for the Münsingen–Rain data, using AA' as a similarity-matrix. (The labels show the serial positions of the tombs in Hodson's chronology.)

The horse-shoe form of the plot stands out well, and a seriation can be obtained from it by reading off the labels identifying the graves, as seen from the 'centre of gravity' of the whole array (the point of intersection of the two perpendicular lines); thus, reading anti-clockwise, we obtain the seriation:

1, 3, 2, 9, 4, 7, 6, 8, 11, 13,
12, 10, 14, 5, 17, 15, 21, 19, 16, 18,
22, 25, 23, 27, 26, 24, 31, 30, 29, 28,
37, 36, 34, 33, 35, 38, 40, 39, 42, 32,
41, 43, 45, 44, 47, 54, 50, 46, 49, 20,
52, 51, 58, 56, 53, 48, 55, 57, 59.

Here (and in figure 3) the labels have been *decoded* and are *the serial positions which were assigned to the tombs by Hodson*. This has been done to enable the reader to compare the computer's seriation with that of Hodson with the minimum of trouble. The agreement between the two is so good as to be somewhat startling.

As a second check on the method, figure 4 was prepared. This is a genuine map of the

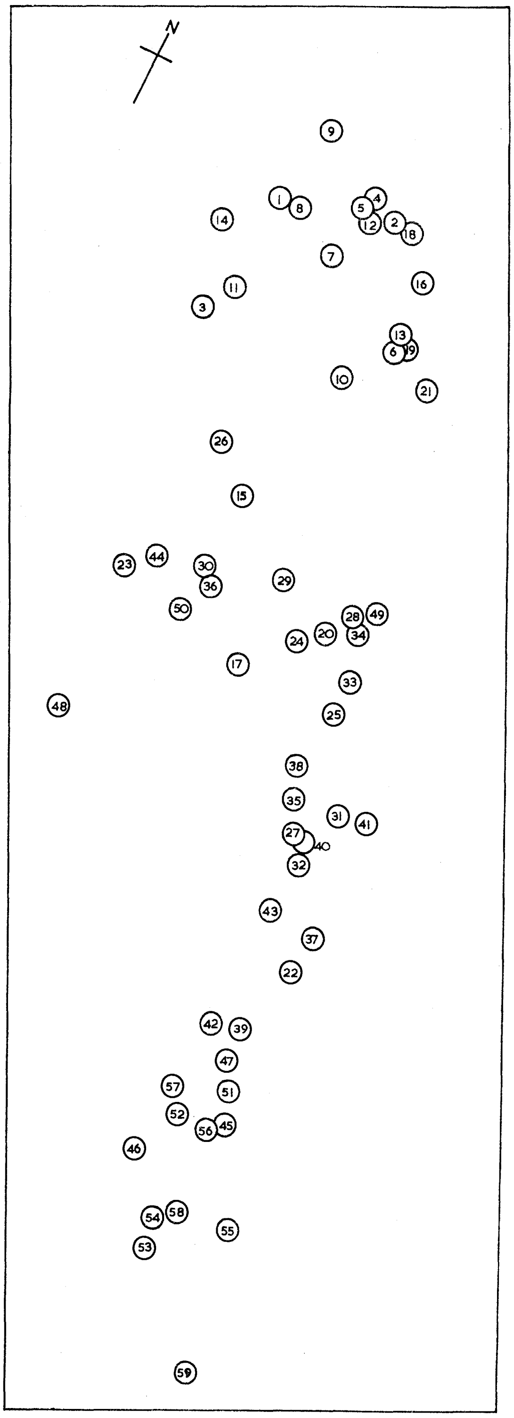

FIGURE 4. A true map of the Münsingen–Rain cemetery, in which the tombs (shown as circles) carry labels indicating their serial positions in the computer chronology derived from the 'horse-shoe' in figure 3.

Münsingen–Rain cemetery and here the tombs are represented by small circles located at their actual geographical positions. This time the labels carried by the tombs and shown within the circles indicate their positions in the *computer's* chronology, so that the general increase in label as one scans the cemetery from one end to the other shows that, using only the assignment of the 1711 pairs to the 8 groups, we have to some extent succeeded in *reconstructing the topography of the cemetery*.

4. Discussion

It is possible that the Münsingen–Rain data was unusually favourable as a test for the method, but the success of the latter with this data at least justifies similar experiments with other suitable material. Some technical points will be noted in conclusion; these are of less than general interest and the reader may choose to omit them.

(*a*) As a third check on the method the incidence matrix for the Münsingen data was row-rearranged in the manner dictated by the computer, and it was found that in all varieties save one (in relation to one single tomb) pattern P was closely adhered to.

(*b*) It has several times been mentioned that the MDSCAL program starts from a random initial position. This can be varied at will, and so one obtains a number of different computer seriations which it is of interest to compare. Five such seriations were obtained for the Münsingen data and showed strong concordance. This is a fourth check on the method.

(*c*) A fifth check, implicit but not commented on above, was the fact that the plot for the Münsingen data turned out to be of the horse-shoe form.

(*d*) We do not yet know how to tell from the gross statistics of the incidence matrix whether we have a seriation problem that will 'come out' by the present method, or not. The development of such a 'rule of thumb' is urgently needed.

(*e*) It may be possible to 'unbend' the horseshoe into linear form by reducing the weight given to the highly dissimilar pairs (Conan Doyle 1892). (I owe this suggestion to Dr R. M. Needham.)

(*f*) A further check on the method would be obtained by using not all pairs of tombs, but various randomly chosen subsets of these.

(*g*) Frequently one finds that 'varieties' are multiply represented in tombs. From such data one can always obtain an incidence-matrix by scoring 1 and 0 for presence and absence, as usual, but information is being lost and so it is of interest that the theorem about AA' [Kendall 1969*a*) which forms the basis of the present method can be extended as follows: in the case of multiple representation, the information relevant to seriation is contained in the numbers c_{ij} calculated from

$$c_{ij} = \sum_h \min\,(n_h(i),\, n_h(j)),$$

where $n_h(i)$ denotes the number of representatives of the hth variety in the ith grave. Here 'seriation' refers to the restoration of

Pattern Q. In every column (i.e. for each variety), the number of representatives per grave increases to a maximum, and then decreases.

This is the natural generalization of pattern P when multiple representations occurs. The HORSHU program (= the Shepard–Kruskal MDSCAL program supplemented by the further procedures briefly outlined in the present paper) could thus be used without essential change even in the case of multiple representation, and its performance in this role is now being studied.

(*h*) Another use for HORSHU is to verify that material believed to be temporally homogeneous is in fact so, for an occurrence of the horse-shoe pattern would be evidence against such homogeneity.

(*i*) The HORSHU program as it exists at present cannot handle more than 90 tombs, and larger cemeteries must be broken down and seriated separately, and the partial seriations must then be combined. The design of such piecemeal analyses is also being studied.

(*j*) The general (0,1) method applies whenever we have data relating to two types of entity:

'*happenings*' (e.g. tombs) located precisely at a *point* of time;

'*fashions*' (e.g. varieties) which are in vogue for *periods* of time.

A score of 1/0 in a cell of the incidence matrix implies that some named 'fashion' was/was not in vogue at the instant of some named 'happening'. The possibility of achieving (approximately) pattern P amounts to this: that for each 'fashion', the 'happenings' in which that 'fashion' was manifested can be brought (approximately) together by a single serial rearrangement of the 'happenings'. Now identify a 'happening' with a *point*-mutant of a virus, and identify a 'fashion' with a *deletion*-mutant for the same virus. When such viruses are bred together in a host cell, recombination (reappearance of the 'wild' type) will occur if and only if the 'place' of the point-mutation lies outside the 'interval' of the deletion-mutation on the (linear) virus. Thus the record of such recombination experiments leads to an incidence matrix A, and on interpreting this suitably (recombination corresponds to 'absence', no recombination corresponds to 'presence') we can employ the HORSHU program to reconstruct the linear structure of the virus, as established first by Benzer (1959 and 1961), *just as if it were a cemetery*. 'Time', in this use of HORSHU, corresponds to location on the linear virus. This program has been successfully carried out on data supplied by the M.R.C. Molecular Biology Laboratory, Cambridge; it will be reported on elsewhere. (See also Fulkerson & Gross (1965) for another mathematical approach to genetic seriation.)

In conclusion I should like to thank J. B. Kruskal and Jessie McWilliams who provided me with a version of MDSCAL on the Titan computer in the Mathematical Laboratory, Cambridge; the Director of that Laboratory for computing facilities; Sarah Prior and Mary Brooks for computing assistance; A. D. McLaren for numerous helpful discussions of novel uses of the MDSCAL program and for suggesting the 'towns' example; and F. R. Hodson who supplied the Münsingen–Rain data and who has been a never-failing source of enlightenment and encouragement.

References (Kendall)

Benzer, S. 1959 *Proc. natn. Acad. Sci. U.S.A.* **45**, 1607.
Benzer, S. 1961 *Proc. natn. Acad. Sci. U.S.A.* **47**, 403.
Doyle, A. Conan 1892 (Feb.) *Strand Magazine*. London.
Fulkerson, D. R. & Gross, O. A. 1965 *Pacific J. Math.* **15**, 835.
Hodson, F. R. 1968 *The La Tène Cemetery at Münsingen-Rain* (Bern).
Hole, F. & Shaw, M. 1967 *Rice Univ. Studies* **53**, no. 3.
Kendall, D. G. 1963 *Bull. Int. Statist. Inst.* **40**, 657.
Kendall, D. G. 1969*a* *Pacific. J. Math.* **28**, 565.
Kendall, D. G. 1969*b* *World Archaeology* **1**, 68.
Kruskal, J. B. 1964 *Psychometrika* **29**, 1, 28.
Petrie, W. M. Flinders 1899 *J. Anthrop. Inst.* **29**, 295.