25. The Rozoy Numerical Ordination and Seriation program package for the analysis of nominal data matrices with MS-DOS Personal Computers

J.-G. Rozoy
26 rue du Petit Bois, F-08000 Charleville, France.
J. Bennink
Brahmslaan 6, 9603 CB Hoogezand, the Netherlands.
R.R. Newell
Biologisch Archaeologisch Instituut, Poststraat 6, 9712 ER Groningen, The Netherlands.
T.S. Constandse-Westermann
Lawickse Allee 66, 6707 AK Wageningen, the Netherlands.

25.1 Introduction

The *Rozoy Numerical Ordination and Seriation program package* records, sorts and then ordinates and/or seriates analytical units and their attributes, arranged in a two-dimensional ($N \times K$ = units $\times$ attributes) matrix of nominal data (presence/absence).

Theoretical considerations of these two techniques and their programmed execution have been treated elsewhere (Rozoy 1987; Newell *et al.* forthcoming; Rozoy *et al.* forthcoming) and in a WordPerfect file, which is included in the program package. Here it is sufficient to say that the program achieves the Wilkinson (1974) ordination by a mean barycentre ordering of both the attributes along the X-axis and the units along the Y-axis of the original data matrix. It leads to the detection of the presence or absence of discrete taxonomic entities within the static data. Ordination does not lead to the reliable resolution of a possible third dimension in the data, be it a clinal shift in chronology, in geographical location, or any social dimension or process. Seriation proceeds from a different premise, i.e. that differences in the first occurrence of the attributes will occur and that they will indicate a linear dimension in the data. By concentrating the attributes in the columns of the matrix, the best representation of this linear dimension can be obtained.

Because of the differing theoretical premises and operational constraints acting upon the linear ordination and the third-dimensional seriation approaches, both algorithms are included in the package. However, it must be stressed that the two cannot be utilized interchangeably. Following a successful ordination, the analyst would be well advised to run one or more subsequent seriation analyses on an ordinated matrix in order to check for unsuspected but potentially extant directional processes (dimensions) in the presumed static, two-dimensional data matrix. Conversely, when investigating temporal or other third-dimensional processes by means of the seriation approach, it would behove the analyst to execute an experimental run of the final seriated matrix with the ordination algorithm. In this way, some sense of the quantitative effect of the directional process being studied can be obtained.

25.2 The Rozoy Numerical Ordination and Seriation program package

The program can accommodate matrices as large as 140 individual analytical units (records) on the Y-axis and up to and including 140 discrete attributes, arranged in columns, on the X-axis. Both the ORDINATE and the SERIATE algorithm can be applied to any one of the following three types of input matrices:

1. coded *raw* unit data, lacking any previous archaeological ordering of the observations on the Y-axis and/or the attributes on the X-axis;
2. coded data of analytical units, all of which have been previously ordered on the Y-axis according to one or another method of ordination and/or seriation;
3. coded unit data, whose sequence of data-entry on the Y-axis follows an independent archaeological ordering, e.g. stratigraphy, C^{14}, spatial parameters, status hierarchy, etc.

Additionally, the SERIATE algorithm can be applied to a fourth type of input matrix:

4. coded unit data, whereby the first and the last N analytical units can be, or have been arbitrarily, FIXed. This symmetry is necessary because of the reversibility of every ordination and seriation.

25.3 System application availability

The MS-DOS TURBO PASCAL version of the *Rozoy Numerical Ordination and Seriation program package* is available in French and English as an executable MS-DOS '.EXE' file. It can be obtained from the authors on 5¼" floppy discs or on 3½" diskettes, formatted to 360 Kbytes or 720 Kbytes. APPLE IIe and MACINTOSH versions, in both French and English, are being completed and should be available soon.

25.4 Hardware/software requirements

The MS-DOS version requires an IBM compatible XT or AT machine with at least one disc drive and a minimum of 256 Kbytes memory as well as an on-line printer, having a minimum of four internal fonts with a character size range between 9 and 17 cpi.

The package software consists of one program file, one HELP file, one Config file, one document file containing description and instructions, two example files, one registration file and a self-executing INSTALL file, together comprising 170,000 bytes on a single disc. Finally, the program has been configured to accommodate the optional use of a Microsoft compatible mouse. The program is provided as a single, compiled MS-DOS '.EXE' file and can be run successfully by MS-DOS 3.0 and higher, without requiring the use of a TURBO PASCAL facility.

The *Rozoy Numerical Ordination and Seriation program package* consists of three linked program modes; data entry and editing, matrix analysis, and evaluation and presentation of results. The execution of the program package is illustrated by the flow-diagram in Fig. 25.1.

The program permits the analyst to view the ordinated/seriated data-matrix interactively on the monitor screen and to inspect the Matrix Evaluation. The Matrix Evaluation records the total number of cells enclosing the ordinated/seriated matrix, the number of empty cells within the enclosed ordinated/seriated matrix, and the number of occupied cells within the enclosed ordinated/seriated matrix. The enclosed ordinated/seriated matrix is the sum of all the cells between the first appearance of an attribute and its last incidence in each column. It includes both observed (presence) data and negative (absence) data cells interspersed between them. Rozoy's Measure of Concentration (Newell *et al.* forthcoming; Rozoy 1987; Rozoy *et al.* forthcoming) is a percentage expression of the number of occupied cells divided by the total number of cells enclosing the ordinated/seriated matrix. It measures the success of the ordination or seriation to include within the enclosed matrix only those empty cells which are the result of normally expected absence of attributes, described above, and absences caused by sample bias or stochastic variation. Because the Measure is calculated relative to the enclosed matrix, which can attain a maximum of 100%, it is independent of the size of the total data-matrix (units × attributes). Calculating the Measure of Concentration in this fashion produces indices which are mutually comparable and can be used to assess different ordinations/seriations of the same matrix as well as between different matrices.

25.5 Program description

Job-control is realized by a hierarchy of Main Menu and Sub-Menu windows, each of which is supported by its own pull-down HELP screen. At critical nodes the program provides the opportunity to check, correct, change, transpose (move), delete or add to the list of nominal attributes and/or the list of analytical units being ordinated or seriated. The File I/O option permits the user to Save and/or Change the name of all New input files, to Load and further analyze extant files and to Change the name and description then Save such mutated files. This robust program package also features a suite of internal checks and interactive job-control functions to ensure consistency and conformity between the computer and the analyst.

The program has a SETUP option giving the user the opportunity to specify and save the predetermined directory and the printer codes. The saved configuration will be called each time the program is used. Furthermore there is an interactive option for entering the data of a new data matrix, to move specific units or attributes by hand, to insert new units or attributes or to delete same.

The options designed to ordinate and to seriate the data both have the possibility to choose the number of iterative runs to be made in search for the ideally ordinated or seriated matrix. While there is room for 999 runs, in practice a small matrix, e.g. 5 × 5, requires no more than 5 runs to achieve the optimal result. A medium-sized matrix, e.g. 20 × 16, achieves maximum automatic ordination/seriation within 100 runs. A large matrix, e.g. 103 × 137, will require at least 400 runs, depending upon its complexity. Once the desired number of runs has been entered, the program presents another window, in which the ordination/seriation process can be monitored. For each successive run the Run number and the Measure of Concentration are presented. When a specific run yields a higher Measure of Concentration than any previous run, the ordinated/seriated data-matrix of that more successful run is automatically saved and used to over-write the formerly highest matrix.

In the case of an ordination analysis, each successive run starts with a pseudo-randomized sequence of the units and the attributes. In the case of a seriation analysis each successive run starts from the foregoing result and is expected to produce a cumulatively better approximation of the final machine-generated optimally compacted matrix. If the machine produces the same result in two successive seriation runs, the program generates a pseudo-randomized sequence and continues with the iterative analysis.

The program package can generate the following suite of on-line printer outputs:

1. A Matrix Evaluation, described above.
2. An Ordinated/Seriated Matrix of the ranked attributes and ordered analytical units, in which the intersects of co-incidence are filled with a square (■). There is also the option of having all intermediate empty column cells between filled intersects marked with a vertical dash (|), denoting the total space of the enclosed ordinated/seriated matrix within the total analytical structure of *x* attributes by *y* analytical units.
3. An Ordinated/Seriated List of the Attributes by their 20-character (max.) alphanumeric codes and by Attribute number.
4. An Ordinated/Seriated List of the Graves/Sites by their 20-character (max.) unit labels and by Unit number.
5. A List of the Graves/Sites with their characteristic Attributes, with tabulated frequencies.

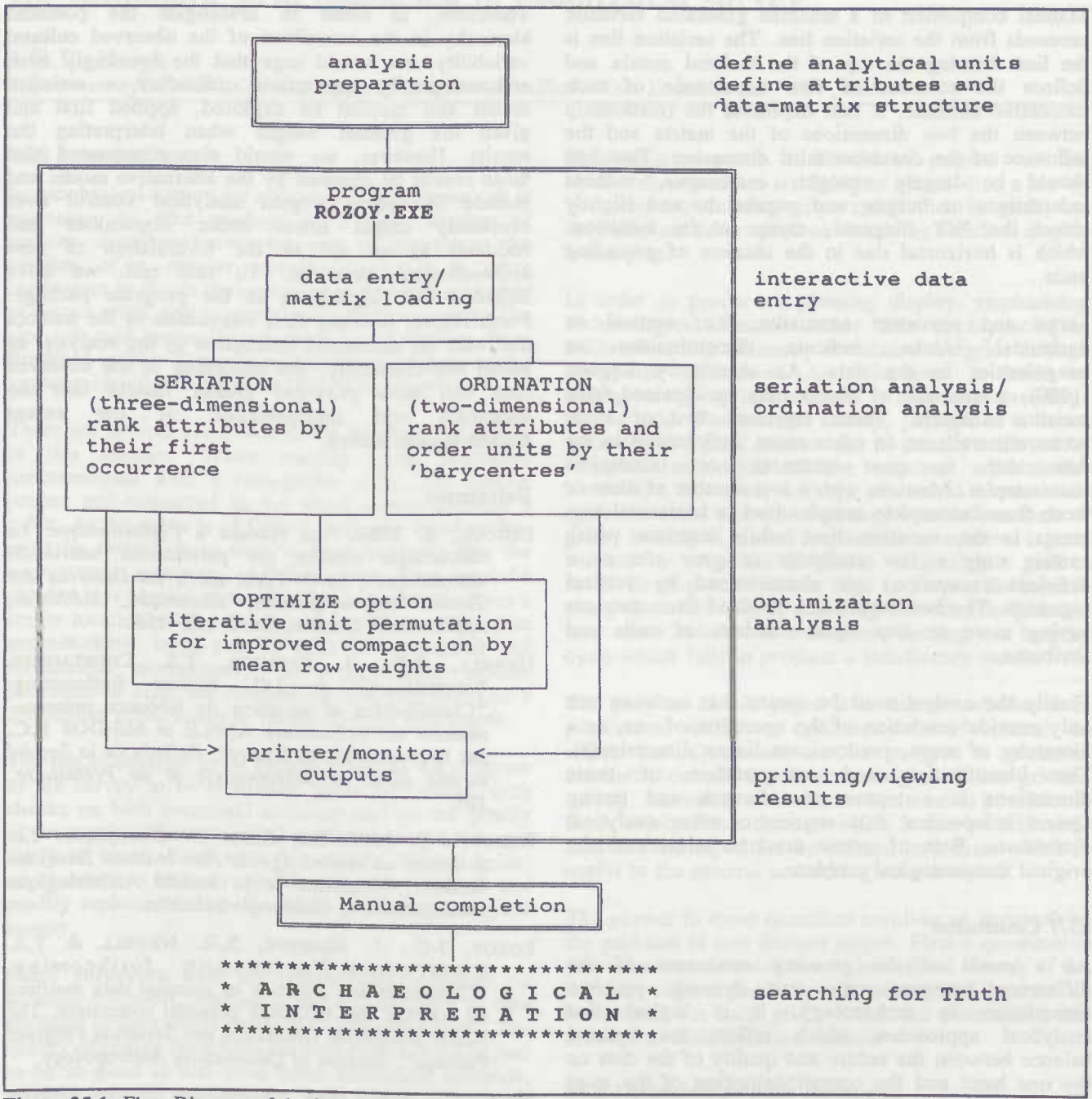

Figure 25.1: Flow-Diagram of the *Rozoy Numerical Ordination and Seriation program package*.

6. A List of the Attributes with the Graves/Sites in which they occur, with tabulated frequencies.

Finally the program contains two example files.

25.6 Manual completion of the matrix

The iterative nature of both the ordination and the seriation algorithms may not lead to the optimal solution. Therefore it is necessary to complete the analysis manually, with the support of the program, which facilitates executing the necessary permutations of rows and transpositions of columns, while calculating the Measure of Concentration after each change. In this way it is possible to design and examine several dozens of alternatives in one session and reach the optimal solution step by step.

The manual approach to ordination proceeds from the notion that the maximum symmetry and compaction of the data in an ordinated matrix provide a model against which one can search visually for alternative and better solutions to the program generated result. These may consist of the simple transposition of attributes and/or graves/sites to obtain a higher Measure of Concentration. They may also include a condensation or dispersal of missing or unobservable data in the matrix. Alternatively, the operation, albeit secondary, of an underlying dimensional process in the ordinated matrix may also produce insights for the realization of a mini-max ordination, which is different from that generated by the program automatically. Clearly, these supplementary and/or complementary approaches should be explored before the analyst simply accepts a program-generated result. In principle and practice, the same is true for program-generated seriation results.

Manual completion of a machine generated seriation proceeds from the seriation line. The seriation line is the line forming the top of the seriated matrix and defines the moment of first occurrence of each successive attribute. It best expresses the relationship between the two dimensions of the matrix and the influence of the *dominant* third dimension. That line should be largely straight, continuous, without indentations or bulges, and parallel to and slightly above the X/Y diagonal, except at its initiation, which is horizontal due to the absence of preceding units.

Large and persistent anomalies, i.e. vertical or horizontal blocks, indicate discontinuities or irregularities in the data. As shown by Legoux (1980), a structure of blocks may be obtained from seriation analysis. These represent two or more taxonomic units or, in other cases, deficiencies in the data due to poor matrices or incomplete sites/samples. Matrices with a low number of units or those from incomplete samples lead to horizontal segments in the seriation line, while matrices which contain only a few attributes (a poor site or a deficient excavation) are characterized by vertical segments. The best results are obtained from matrices having more or less equal numbers of units and attributes.

Finally the analyst must be aware that seriation can only provide resolution of the operation of one, or a hierarchy of more, predominant linear dimension(s). The identification and interpretation of those dimensions is a matter of inference and testing against independent data sources or other analytical dimensions. Both of course must be relevant to the original archaeological problem.

25.7 Conclusion

As a result of the growing awareness of the difference between static and dynamic problem formulation in archaeology, it is logical that analytical approaches which reflect the optimal balance between the nature and quality of the data on the one hand and the operationalization of the most relevant underlying assumptions and principles, on the other hand, be developed. At the same time, our increasing sensitivity to the polythetic and systematic nature of cultural variability should caution us against relying too heavily upon the simple binary choice of either the one approach or the other, in the expectation that it alone will provide complete or even maximum resolution of the inherent patterning.

Therefore, in order to investigate the potential hierarchy in the causalities of the observed cultural variability, we would urge that the seemingly most archaeologically appropriate ordination or seriation model and method be explored, applied first and given the greatest weight when interpreting the results. However, we would also recommend that those results be checked by the alternative model and method in order to gain analytical control over eventually extant lower order dimensions and residuals as an aid to the formulation of new archaeological questions. To that end, we have included both algorithms in the program package. Furthermore, it is the firm conviction of the authors that with the successful conclusion of the analysis, as tested experimentally, the patterning in the observed data will have provided greater insight into the diagnosis and interpretation of the extant archaeological record.

References

LEGOUX, R. 1980. "Le recours à l'informatique: La chronologie relative par permutation matricielle automatique", in P. Périn (ed.), *La Datation des Tombes Mérovingiennes, Historique, Méthodes, Applications*, Genève, Droz: 138–155.

NEWELL, R.R., J. BENNINK, T.S. CONSTANDSE-WESTERMANN & J-G. ROZOY forthcoming. "Classification et sériation de tableaux présence-absence sur ordinateurs APPLE et MS-DOS P.C. par le programme de Rozoy", *Bulletin de la Société Royale Belge d'Anthropologie et de Préhistoire*, 101.

ROZOY, J-G. 1987. *Les Celtes en Champagne. Les Ardennes au second Age du Fer: le Mont Troté, les Rouliers*, Mémoires de la Société Archéologique Champenoise 4, Charleville-Mézières.

ROZOY, J-G., J. BENNINK, R.R. NEWELL & T.S. CONSTANDSE-WESTERMANN forthcoming. "Ordination and seriation of nominal data matrices on APPLE and MS-DOS personal computers: The Rozoy Numerical Ordination and Seriation Program Package", *Journal of Quantitative Anthropology*.

WILKINSON, E.M. 1974. "Techniques of data analysis seriation theory", *Technische und Naturwissenschaftliche Beiträge zur Feldarchäologie*, Köln, Rheinland Verband: 1–142.