2

Data analysis for archaeologists: the Institute of Archaeology packages

F. R. Hodson*

P. A. Tyers†

2.1 Introduction

The development of the IA packages was funded by a grant from the Science Based Archaeology Committee of the SERC (GR/D08937). Although designed initially for cemetery data, they attempt to provide analytical programs for most archaeological contexts. Some of the incorporated procedures (*e.g.* Principal Components, Canonical Variates and Cluster Analysis) are already available in standard packages such as SPSS, CLUSTAN or, more especially, GENSTAT. Other procedures (*e.g.* for Seriation) are absent from general packages, but have been previously programmed for archaeological use (*e.g.* Goldmann 1972, Kendall 1971 or Scollar 1988). However, a special purpose archaeological package is still required for the following reasons:

1. beside seriation, there are other archaeological configurations that require special treatment: for example, the construction of status tables.

2. general strategies for data analysis, such as Cluster Analysis, have a wide range of possible implementations and in spite of the complex and apparently comprehensive options offered by the standard packages, appropriate procedures are often elusive.

   For example, there is to our knowledge no other implementation of the specific suite of k-means procedures that is provided here and that seems the most appropriate for many archaeological contexts. Likewise, the Principal Coordinates aspect of the Principal Components model has wide archaeological relevance but is difficult for an archaeologist to implement and Correspondence Analysis is not yet regularly included in such packages.

* Department of Prehistory,
Institute of Archaeology,
31–34 Gordon Square,
London WC1H OPY

† Department of Prehistory,
Institute of Archaeology,
31-34 Gordon Square,
London WC1H OPY

The large standard packages are also clumsy to use: they tend to require complicated specifications as input and to produce unnecessary volumes of output. A major aim with these packages has been to provide, through carefully engineered default combinations, extremely simple commands for calling the various procedures and for receiving the minimum output necessary, carefully formatted. At the same time, if desired, a very wide range of options may be chosen by more detailed commands.

3. There are now many excellent statistical and graph packages available for microcomputers that archaeologists may use for the preliminary inspection of quantified data, but these cannot provide all the multivariate procedures included here. On the other hand, it is still difficult to install and support the large general purpose packages with multivariate procedures (that tend to be written in FORTRAN) on other than large central computer installations. By restricting the range of programs in this package to those of archaeological interest and by programming in C for UNIX, it is possible to install the IA packages on a mini computer, or even on a powerful micro.

This is not to suggest that users will not find the established small and large packages useful or even necessary, but the IA packages should be more generally convenient and comprehensive for their needs.

As currently implemented, IAP is seen as a first experimental package that will be revised and enlarged in response to user demand. The individual programs are assembled into two distinct sub-packages which require data in two different forms: as incidence ('presence') data (IAGRAVES) and as quantified data (IASTATS). This distinction emphasises the fundamental difference in treatment required for information of the present-or-absent kind and of the metric kind. The former may, as an expedient, be converted into numerical form (i.e. be coded as 1, 0 or some such), and, after standardisation, be treated as metric data, but this assumes that absence is as significant as presence: an assumption that is rarely realistic for archaeological data. This distinction also allows incidence data to be listed and stored economically, since the absence of a unit from a context does not need to be recorded, whereas all values of a metric variable (including zero) require to be specified.

2.2 IAGRAVES

Two specifically archaeological programs are involved: SERIATE for seriation and SOCISTAT for the investigation of status. A third suite of programs (JACCARD) combines more standard statistical procedures that are appropriate for incidence data. These programs and a comprehensive application are discussed in detail by Hodson (Hodson 1988, summarised in Hodson 1986). The following is a brief summary of some of the features.

2.2.1 Jaccard

These procedures allow incidence data to be tabulated, and cross-associations to be inspected and, if desired, analysed by appropriate multivariate techniques based on the

Jaccard coefficient (currently Single Link Cluster Analysis[1] and Principal Co-ordinates Analysis). This combination should have many applications, but in cemetery analysis provides a useful preliminary indication of major groups of related functional types and their likely sex-linkage. Such preliminary information will usually be needed before any further analysis.

2.2.2 Seriate

In spite of voluminous literature on archaeological seriation, there is no accepted 'best' mechanical approach but a range of possible alternatives. IAGRAVES provides a means to experiment with many of these alternatives, but also allows the user to call by a simple command a default sequence of currently preferred operations leading to an overall result:

1. elimination of types that occur less than twice and graves that include less than two cross-associated types
2. diagonalisation of the data matrix by one dimensional correspondence analysis
3. minimisation of the spread of column incidences by reordering rows
4. engineering a regular profile for the appearance of types

This default combination results from extensive experimentation with trial data. Stage (b) follows Ihm's algorithm (Ihm 1981, Ihm 1983) for diagonalising a data (not a similarity) matrix. This is now widely accepted as a substitute for the erratic AXIS algorithm used to derive an approximate result in many earlier programs (*cf.* Wilkinson 1974, Graham *et al.* 1976). This strategy itself provides a seriation that may be considered appropriate on theoretical assumptions and may be taken as a final result.

However, following the work of Kendall (Kendall 1963), Doran (Doran 1971) and Wilkinson (Wilkinson 1974) and after extensive trials (Hodson 1988), it has seemed advisable to 'improve' a stage 2 ordering by invoking the concentration principle as stage 3: i.e., by seeking to concentrate incidences in the columns (types) of the data matrix. Two criteria of concentration are available for minimisation in this package: Doran's, which simply sums the ranges of column incidences, and Wilkinson's, which attempts to allow for the density as well as the range of column incidences, and invokes a gamma function. Empirically, Doran's criterion has seemed more generally appropriate, and is taken as the default option.

The third default choice (stage 4) orders types (and if necessary reorders rows) to preserve a constantly falling leading edge to the appearance of new types. In other words, a type is located by its first appearance rather than by the mid-point of its currency. Some practitioners may not wish to use this default option but it has obvious practical and possibly some theoretical advantages: it simplifies the study and interpretation of results (not unimportant with large data sets), but it also

[1]The choice of the single link method for clustering requires some justification, since although esteemed theoretically, this procedure has usually proved disappointing in practice. However, the performance of any clustering method is highly dependent both on the nature of the data and on the similarity coefficient through which it operates. Combined with the Jaccard coefficient, the single link method has certainly proved effective for clustering functional types in graves and so, pending further results, it is included here as a standard procedure.

accommodates the assumption that the frequency of type occurrences is likely to be very skewed: a rapid circulation of any new style followed by a slow fall-off, rather than a symmetric Gaussian model which is sometimes assumed.

Because of the uncertain direction of any seriation *a priori*, the default provides for stages 3 and 4 to be performed on both the initial stage 2 result and its reverse.

As with the SOCISTAT package (discussed below), auxiliary data (listed with but not controlling the result) may be used with SERIATE.

Abundance data

Although most cemetery seriations are performed on incidence ('presence') data, seriation may be desired for assemblages described by counts. Such quantified data should be processed by the IASTATS package. A Correspondence Analysis, taking the order for contexts and types on the first axis, would seem a suitable option.

2.2.3 Socistat

Methods for producing 'Status Tables' from cemetery data are even less standardised than seriation and the SOCISTAT implementation is necessarily experimental. Although a standard default procedure may be called, this program should preferably be controlled by declaring a combination of options to suit a given sample of graves.

The program requires as data a list of functional types occurring in graves (both heraldic and utilitarian types may be regarded as functional). For each type a status index is calculated (the mean number of different functional types in graves where the type occurs), and this index is used to order graves. Clearly, such indices are very sensitive to the size of the data set and thus the program is designed for relatively large numbers of graves. By default, a grave is ordered first by the highest status type which it contains, second by the sum of status indices for its remaining types. An alternative approach, ranking graves according to the sum of status indices for their constituent types, is available (Hodson 1977).

Where possible, the presumed sex-linkage of types (derived perhaps from JACCARD) should also be listed so that the graves may be initially split into male, female, uncertain or mixed male and female.

Published results using this program may be found in Brenan (Brenan 1985) and Hodson (Hodson 1986, Hodson 1988).

2.2.4 Sample inputs and outputs

A sample data set, listing all attribute types (2AR, RFT etc, associated with the named contexts (01:01, 01:02 etc, and suitable as input to all parts of the IASTATS package, has the following format:

```
01:01  2AR RFT 0
01:02  NRG 0
01:03  COF SCV DAG PSH 0
03:01  2AR SCV PSH CTR PRI 0
05:01  COF NRG PSH PRI CTR LAR 0
06:03  2AR COF 0
06:04  0
```

```
06:06 COF 0
13:01 COF SCV PRI CTR 0
13:03 2AR COF NRG AMB 0
13:04 0
13:05 COF SCV 0
13:06 COF SCV 0
13:09 COF 1FN 0
13:10 2FS 0
13:12 COF 0
13:13 0
```

With the data set in the file `demo.dat` the UNIX command:

```
seriset demo.dat | socistat
```

would produce a status table using the default combinations—the program `seriset` is a pre-processor. Tables of statistics and other data which may be generated are here omitted.

```
 1  06:04  ..............
 2  13:04  ..............
 3  13:13  ..............
 4  13:10  *.............
 5  06:06  ..*...........
 6  13:12  ..*...........
 7  01:01  .*.*..........
 8  06:03  ..**..........
 9  13:05  ..*.*.........
10  13:06  ..*.*.........
11  01:02  ........*.....
12  13:09  ..*....*......
13  13:03  ..**..*.*.....
14  01:03  ..*.**...*....
15  13:01  ..*.*.....**..
16  03:01  ...**....***..
17  05:01  ..*.....*****.
18  27:01  ..*....**.****

           2RC2SDA1NPPCL2
           FFOACAMFRSRTAF
           STFRVGBNGHIRRT
```

With the addition of sex-linkage and auxiliary data, read from additional data files, the resulting table contains additional information:

```
 1  06:04  ? I?  B      .............
 2  13:04  ? C          .............
 3  13:13  ? C??        .............
 4  13:06  H1C??m   M   ..*.*.........
 5  13:05  H C      M   ..*.*.........
 6  01:03  H1I      M   ..*.*.*......*
 7  13:01  H2C? m   M   ..***..*......
 8  01:01  H I  s   F   .*........*...
 9  06:03  H1I? wB  F   ..*.......*...
10  01:02  H2I      F   ............*.
11  13:09  H2C??    F   ..*........*..
12  13:03  H2C?     F   ..*..*....*.*.
13  13:10    C  w   U   *.............
14  13:12  LTC??m   U   ..*...........
15  06:06  LTI  sB  U   ..*...........
16  03:01    I? wG  X   ...**..*..*..*
17  05:01    I  wY  X   ..**...**...**
18  27:01  ? I   Y  X   ..**...***.**.

                                mmMMmMFFFf

          D RQ CC       2RCCSADPL221NP
          A IU OO       FFOTCMARAFAFRS
          T TA FL       STFRVBGIRTRNGH
          E EL FO
            I IU
            T NR
            Y
```

The columns of data above the auxiliary data names (DATE, RITE, QUALITY etc) list the auxiliary data states associated with each context.

The sex link of each type is shown above the type name using these codes

M	diagnostically male
m	preferentially male
F	diagnostically female
f	preferentially female
0	not sex-linked

and the probable sex linkage of each context is noted beside the auxiliary data

M	male
F	female
U	uncertain
X	mixed

2.3 IASTATS

This package attempts to provide a comprehensive range of more or less standard procedures for the analysis of quantified data. The following are included: Principal Components/Coordinates analysis, Correspondence analysis, Canonical Variates analysis, and Cluster Analysis (k-means and single-linkage).

IASTATS does not at present provide procedures for non-metric multi-dimensional scaling nor for many of the possible but not necessarily appropriate versions of Cluster Analysis. Special purpose methods of spatial analysis are not supplied, although the k-means program may be so used with x, y co-ordinates treated as scores on two variables (*cf.* Kintigh & Ammermann 1982). Database facilities as such also are not provided: data must be input in a standard format, but this should be derivable from a data base via standard data base management systems. Missing data and a range of transformations (logs, square-roots etc) are provided.

The main procedures implemented are as follows:

2.3.1 Principal Components and Principal Coordinates

Principal Components analysis (PCA) is a mathematically respectable, well-tried method for revealing structure within multivariate data (Doran & Hodson 1975, p. 190). Correlated information from the original variables is combined to form a series of new summary variables, the components, arranged in descending order of importance (i.e., in terms of the amount of variation summarised). The components may be treated as coordinate axes against which either the units of the analysis or the original variables may be plotted as a scatter diagram.

The output from PCA includes

1. Information about the importance of each component relative to the total variation within the sample. This is expressed as a percentage and cumulative percentage of total variation. The method is most helpful when a relatively large percentage (say 70%–80%) is accounted for by the first two or three components. The resulting scatter plots of component 1 against 2 and 1 against 3 will then provide a comprehensive summary.

2. A table listing the contribution of original variables to the successive components. This should help to interpret the components archaeologically.

3. A table listing the relationship of the units to the new components. This provides the coordinate values for plotting units against components.

Principal Coordinates Analysis or PCO (Doran & Hodson 1975, p. 194) is closely related to Principal Components Analysis but locates units on principal axes not from raw scores but from a matrix of distances or similarities between the units. This allows units described by presence data (as in IAGRAVES) to be plotted against principal axes via a matrix of suitably transformed JACCARD coefficients.

2.3.2 K-means Cluster Analysis

This program subdivides the sample of units into k clusters (2, 3, 4, ...) up to a specified or a default-estimated maximum (Doran & Hodson 1975, p. 180). For

each subdivision a statistical measure of clustering is maximised and recorded. This measures the concentration of units around cluster centres (the *k-means*) relative to the total dispersion of the sample. It thus reflects the intuitive notion of a cluster—a group of units displaying internal cohesion and external isolation. The successive values of the clustering criterion as k is increased may be plotted and in favourable circumstances (for example, where a marked shoulder in the curve is seen) taken to indicate an appropriate number of clusters to interpret. Randomisation of the raw data by column permutation may also help to decide on a suitable k.

A general problem with k-means programs is the difficulty of reaching a subdivision for any k that may be considered optimal. The specific algorithm programmed here (the Singleton-Kautz Algorithm) attempts to achieve this sequentially by moving back and forth between different levels of clustering until an optimal result is achieved for each.

Cluster statistics

This program **cs** provides comprehensive statistics for the preferred k-means result.

Clustering in discriminant space

For certain kinds of metric data, Mahalanobis (discriminant) distance may be preferred to Euclidean distance. This compensates for correlations between variables and weights variables according to their ability to discriminate between clusters. The **disc** program makes it possible to modify a provisional k-means result using Mahalanobis distance. Even if no change in cluster membership is suggested, this supplementary program may prove rewarding since it provides for units to be plotted in discriminant space (against canonical variates, equivalent to principal components derived from discriminant functions). As for Principal Components, the archaeological significance of the canonical variates should be interpretable from the table of associated weights for raw variables.

For examples of this k-means plus canonical variates approach, see Hodson (Hodson 1971), Doran and Hodson (Doran & Hodson 1975, Figs. 9.14, 9.17, 9.20) and Brown (Brown 1982). For the use of this k-means program in spatial analysis as originally proposed by J. Doran (see Kintigh & Ammermann 1982).

2.3.3 Correspondence Analysis

This procedure is related to Principal Components Analysis but summarises data using a Chi-square related metric rather than Euclidean distance (Greenacre 1984). Thus, when the units of analysis are assemblages, the size of assemblages as well as proportionate counts of variables is taken into account. A distinct attraction of this approach is the close relationship preserved between units and variables when the final result is presented. Following Benzecri (Benzecri 1973) the method has been much favoured in France (*cf.* Leredde & Djindjian 1980). IASTATS by no means exploits the full potential of this approach and simply provides a starting-point.

2.3.4 Other IASTATS procedures

IASTATS also provided some standard univariate procedures for summarising and displaying data (*bstats, hplot*) and to assist in formatting and checking data and results (*rdata, dataset, staterror, statusage*). Further programs allow for the transformation of raw data (*std, tform*) and for missing value substitution (*missing*).

Since Single Link Cluster Analysis and a program drawing dendrograms are already implemented in IAGRAVES, they are included in IASTATS (*slca*) where they may be used with quantified data: for example, to cluster correlation coefficients if they have been calculated from appropriate variables. However, k-means cluster analysis would normally be the first choice.

2.3.5 Sample command sequences

The following annotated UNIX shell file performs a sequence of analyses on a data set in a file DATA. Note that the names of the files holding the unit and variable names are specified by the shell variable **NAMES** and inserted into all the command lines. Lines preceeded by the hash (#) are comments.

```
#
# statistical analysis of data set DATA
# (unit and variable names in UNAMES & VNAMES)

# set unit and variable names for all programs
NAMES="-U UNAMES -V VNAMES"

# first the raw data format
rdata $NAMES DATA

# now the basic statistics
#  -c  correlation matrix
#  -f  write correlation matrix to file
bstats $NAMES -c -f Cmatrix DATA

# at this point you can try a single linkage clustering
# of the correlation matrix of units created by bstats
#  - in this case the 'units' are the variables
#    so use a different -U value
slca -U VNAMES Cmatrix

# do principal components analysis on standardised data set
#  -S  standardise data set before analysis
pca $NAMES -S DATA

# k-means analysis on standardised data set (using default value for k)
#  -S  standardise data set before analysis
#  -f  write solutions from 2 to k to file
kmns $NAMES -S -f Kmns.out DATA
```

```
# choose k value and extract new data set with clustering data
# run it through disc to reallocate units between clusters
# (for this run k = 4)
# cs
#  -k  solution to extract
# disc
#  -f  write new solution to file
cs -k 4 Kmns.out | disc $NAMES -f Clustered

# get basic statistics on the clustered data set
bsc $NAMES Clustered
```

References

BENZECRI, J. P. 1973. *L'Analyse des données. I. La Taxinomie, II. L'Analyse des correspondances.* Dunod, Paris.

BRENAN, J. 1985. "Assessing social status in the Anglo-Saxon cemetery at Sleaford", *Bulletin of the Institute of Archaeology, London University,* 21/22: 124–131.

BROWN, M. A. 1982. "Swords and sequence in the British Bronze Age", *Archaeologia,* 107: 1–42.

DORAN, J. 1971. "Computer analysis of the La Tène cemetery at Münsingen-Rain". *in* Hodson, F. R. *et al.*, (eds.), *Mathematics in the Archaeological and Historical Sciences,* pp. 422–431.

DORAN, J. & F. R. HODSON 1975. *Mathematics and computers in Archaeology.* Edinburgh University Press, Edinburgh.

GOLDMANN, K. 1972. "Zwei Methoden chronologischer Gruppierung", *Acta Praehistorica et Archaeologica,* 3: 1–34.

GRAHAM, I., P. GALLOWAY, & I. SCOLLAR 1976. "Model Studies in Computer Seriation", *Journal of Archaeological Science,* 3: 1–30.

GREENACRE, M. J. 1984. *Theory and applications of correspondence analysis.* Academic Press, London.

HODSON, F. R. 1971. "Numerical typology and prehistoric archaeology". *in* Hodson, F. R. *et al.*, (eds.), *Mathematics in the Archaeological and Historical Sciences,* pp. 30–45.

HODSON, F. R. 1977. "Quantifying Hallstatt: some initial results", *American Antiquity,* xlii: 394–412.

HODSON, F. R. 1986. "Hallstatt: dry bones and flesh", *Proceedings of the British Academy,* LXXXI: 29–43.

HODSON, F. R. 1988. *Hallstatt, the Ramsauer Graves: Quantification and analysis.* Society of Antiquaries, London.

IHM, P. 1981. "The Gaussian model in chronological seriation", *in X Congreso. U.I.S.P.P. Commission iv*, pp. 108–124, Mexico.

IHM, P. 1983. "Korrespondenzanalyse und Seriation", *Archaeologische Informationen*, 6: 8–21.

KENDALL, D. G. 1963. "A statistcal approach to Flinders Petrie's sequence dating", *Bull. I.S.I. 34th session*, pp. 657–680.

KENDALL, D. G. 1971. "Seriation from abundance matrices". *in* Hodson, F. R. *et al.*, (eds.), *Mathematics in the Archaeological and Historical Sciences*, pp. 215–252.

KINTIGH, K. W. & A. J. AMMERMANN 1982. "Heuristic approaches to spatial analysis in Archaeology", *American Archaeology*, 47 (1): 31–63.

LEREDDE, H. & F. DJINDJIAN 1980. "Le traitement automatique des données en archéeologie", *Dossiers de L'Archeologie*, 42: 52–69.

SCOLLAR, I. 1988. *The Bonn seriation and archaeological statistics package: version 3.1.* Remagen.

WILKINSON, E. M. 1974. "Techniques of data analysis: seriation theory", *Archaeo-Physika*, 5: 3–142.