TWO SOFTWARE PACKAGES FOR ARCHAEOLOGICAL QUANTITATIVE DATA ANALYSIS

S.G.H.Daniels
1, Gwendrock Villas, Fernleigh Road, Wadebridge, Cornwall

The two analytical packages AQUA and ARCHON share many features and capabilities. Both are designed primarily for use in projects with medium to large quantities of data, where many different analytical questions may be asked about the same data base. Various data storage forms, reasonably close to those in which archaeologists record information, are provided, together with facilities for addition to and alteration of the data base, and extraction of subsets of data. Output is intended to be readily understood by archaeologists unfamiliar with computers and entities are identifiable by full names rather than coded abbreviations. Graphic output via pen plotter or line printer is normally available in addition to numerical output. Both packages are modular in construction so as to make expansion for further applications rapid and comparatively painless.

AQUA, which is the property of the Centre for Nigerian Cultural Studies, Ahmadu Bello University, Zaria, Nigeria, is a 12,000-statement package in FORTRAN. It was written for use on a CDC CYBER-72 and for compilation by the University of Minnesota FORTRAN compiler (MNF), but conforms to ANSI standards as far as possible. Graphic output was designed for a CALCOMP 563 pen plotter and requires the availability of the CALCOMP PLOT library. The full package contains a shell of control language segments which execute programs, and handle files and error conditions under the operating system NOS 1.1.0-430B. However the

FORTRAN programs can be modified for use independently or within a similar shell designed for another operating system. A 180-page manual contains user instructions, algebraic descriptions of algorithms, details of data storage conventions and necessary information for conversion to another computing environment.

While I was writing AQUA the system was under-utilised and economy of time and storage space was therefore largely sacrificed to the then over-riding priority of producing usable results as soon as possible. In the earliest stages the package was intended for an ICL 1900 computer with no remote terminals. Successive modifications, of concept and detail, adapted it to the CDC CYBER, through FTN and MNF to ANSI FORTRAN, and to use from a remote terminal. The final package shows fossilised traces of this process, including an unnecessary rigidity, much inelegant programming, and an interactive character which is only partial, in that interaction is with the control language shell, not with the individual programs.

ARCHON is now in process of development. It is currently about the same size as AQUA, but not yet of the same power. It is written for a Tandy TRS-80 Model 1 micro-computer, with at least 32K RAM, at least 2 mini-floppy disk drives, and a 120-column line printer. The graphic portions make use of a Houston Instruments HI-PLOT flat bed pen plotter and HI-PAD digitizing pad, communication being by RS-232-C interface. The package is written in Radio-Shack Level 2 BASIC with some machine code segments, though portions will soon be rewritten in Radio-Shack FORTRAN for greater speed. Operation is fully interactive and 'friendly'. Processing large bodies of data can be very slow, but overall time from formulating the problem to obtaining the results can compare favourably with time-sharing on a mainframe where scheduling is not under the user's control. Considerable attention has been paid to economy of time and storage space, making it possible to tackle realistically large quantities of data using 48K RAM.

Some features, though common to the two packages, have been improved in ARCHON. Nearly all input and output, as well as frequent mathematical operations, are handled by pre-written subroutines which can be inserted into new programs as required. Addition of new programs is therefore faster and less error-prone. Intelligibility has also been inproved, with output ready for immediate inclusion in a coherent analytical project report. File handling is more flexible and more data storage forms are catered for. ARCHON is my own property and portions of it can be made available on a commercial basis from autumn 1980.

Familiar capabilities of one or both packages include basic univariate statistics, cross tabulations, regression and principal components, while graphic options cover, amongst others, histograms, bivariate scattergrams, 3-pole graphs, distribution and location maps, contour plots and isometric and perspective views of surfaces. There are also a range of non-standard analytical techniques developed for archaeological work and as yet unpublished, partly published or unfamiliar. These include:-

<u>Lentifer analysis</u>. This is a multi-dimensional scaling procedure designed to recover underlying variables of which observed data are approximate lenticular functions. The procedure may be applied to any data in which individual objects have been classified as members of two or more groups (whether assemblages or observational categories), and to square or rectangular matrices. It is metric and uses no intermediate difference

coefficient, being applicable directly to raw or proportional frequencies and presence/absence data. Solution for each dimension does not affect those already found for previous dimensions. The algorithm, which is iterative and comparatively rapid, attempts to minimise within-group variance. I have not met solutions which were non-global minima, though I can not demonstrate this as a theoretical rule. The output is a configuration of points in the solution space, each point standing for a group.

Ramal analysis. This is intended for the recovery of branching trees from a matrix of interpoint distances, the distances being given by:-

$$d_{ij} = - \log_2 c_{ij}$$

where $\underline{c}$ is the proportion of elements in the same state in two systems. The tree found differs from most in that there is in general no point on the tree from which all the terminal points are equidistant. The difference matrix generates a ramal matrix, in which the element in row $\underline{i}$ and column $\underline{j}$ is the estimated network distance from an arbitrary root to the point at which the paths to $\underline{i}$ and $\underline{j}$ diverge. A series of ramifications then replaces this by a ramified form (a ramal matrix which meets the ultrametric constraints required for a real tree), while attempting to minimise the sum of squared differences between initial ramal matrix and ramified form. This procedure is repeated iteratively until stable estimates are obtained for the ramified form. Ramifications can be very lengthy for large data, but iterations are usually very few. The solution is not unique, in that it is given in terms of an arbitrary root, and an infinity of equally well fitting solutions correspond to different positions of the root. A unique rooted solution can be obtained by requiring that the variance of the distances from root to terminal points be minimal.

Delta technique. Intended for classical one-dimensional seriation, this algorithm operates on a difference or similarity matrix and requires only addition and subtraction. For matrices up to about 12 x 12 it is perfectly feasible with a pocket calculator. With the units arranged in a given arbitrary order a set of estimates δ_{ij}, for the distance between *i* and *j*, are obtained using only those units preceding *i* and following *j*. If every δ is positive the order is *consistent* and is accepted as a solution. If any δ is negative, it is taken to imply that the two assemblages concerned are in the wrong order. In this case the pair of units with the absolutely smallest δ is transposed and the procedure repeated. It is arithmetically possible for a given set of data to have no consistent order, or more than one consistent order, but I have not encountered this with real archaeological data and metric differences, where a consistent order (apparently unique) has always been reached.

Mapping of character space into real space-time. Given the reduced-space output of some multi-dimensional scaling technique, the aim is to show how this space relates to the geographical space or space-time from which the assemblages or objects were drawn. A vector function is obtained which gives the approximate location of a point in character space as a function of its location in real space. In AQUA the elements of this function are polynomials of arbitrary degree, while in ARCHON they are a piece-wise approximation by splines. For any given region of the real space, the function gives a region of a particular size in the character space, and the ratio of the latter to the former will vary across the real space.

This ratio exists at a point in the real space, and it is this which is plotted to show the relationship between the two spaces. Regions of the real space in which the ratio is low are those in which the assemblages are generally similar. Regions where the ratio is high are those where adjacent assemblages are comparatively different. The degree of the function, taken together with the number of assemblages and the amount of reduction in the character space, results in more or less 'smoothing' of the variation in ratio values. With a function of suitable degree regions of cultural similarity are revealed as low-valued basins separated by high-valued ridges of dissimilarity.

Ramal analysis is treated fully in Daniels (1968), and I hope soon to be able to publish concise but full details of the other techniques available in AQUA and ARCHON

Daniels, S.G.H. Ramal analysis and evolutionary trees. W.Afr.J.Archaeol., 8. 1978 (In press)