A SYSTEM FOR THE STORAGE, RETRIEVAL AND ANALYSIS OF NUMERICAL DATA IN PALAEONTOLOGY

by IAN E. PENN

ABSTRACT. A statistical and data-management system (known as ASCOP) and a data-processing service are described. These meet present needs in the handling, storage and retrieval, and analysis of numerical data. The system is available, without cost to university users in the U.K. and to government departments at cost, since it is installed at the Atlas Laboratory, Chilton, Didcot, Berks.

THE use of numerical data in palaeontology has given rise to three principal problems which are capable of immediate solution.

After numerical data have been collected palaeontological activity is of two kinds. On the one hand time must be spent on routine mathematical work while, on the other, time has to be spent on scientific thought. Probably much unnecessary time is spent on the former and it is likely that increasingly complex calculations will aggravate this situation. Although electronic computing has speeded laborious calculation, much time has also been wasted by different palaeontologists repeating the programming efforts of their colleagues.

It is thought that the life expectation of a piece of work in palaeontology is about 20–40 years (Craig 1969, p. 317) and it would appear that the most stable part of a work is its 'raw' data. While museums store the specimens concerned and professional publications store the thought involved there is as yet no agreed store for the numerical information.

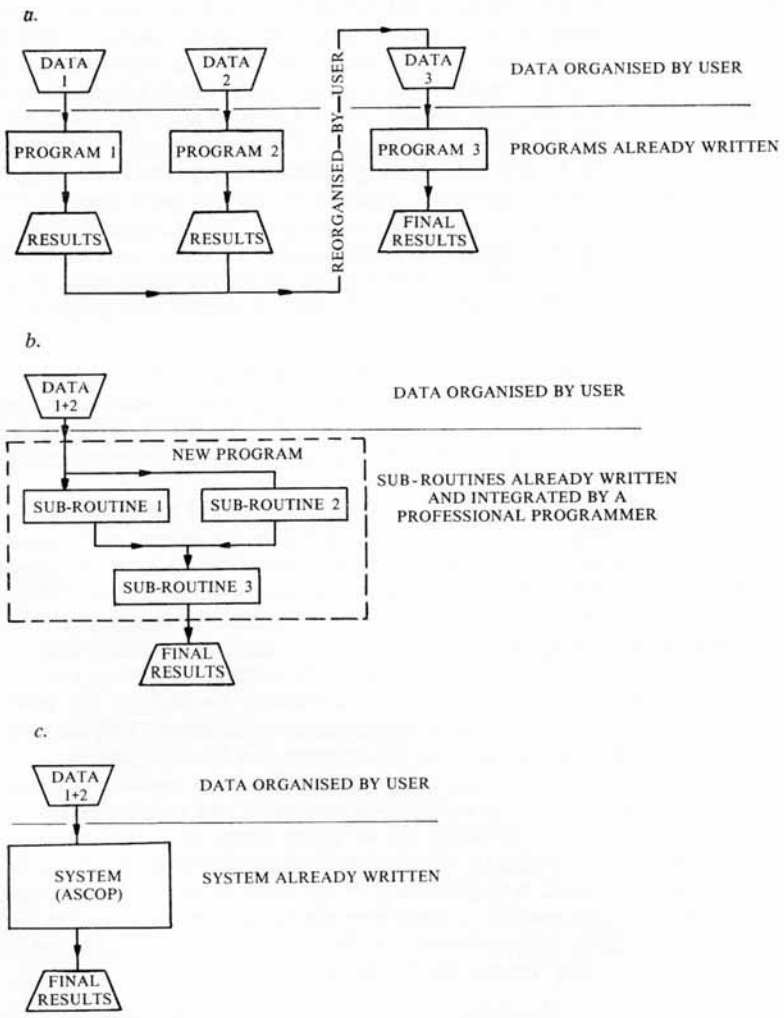
A third problem, that of choosing which statistical technique to use, involves scientific thought about which there may be no objective basis of agreement. Different statistical procedures may be considered relevant or irrelevant to similar situations by different palaeontologists. There is a need to accommodate legitimate differences of scientific opinion so as to simplify the problem of choice. Furthermore, new techniques should be made widely available as quickly as possible and not be the preserve of those who happen to possess an appropriate bit of computational hardware.

A computing system known as ASCOP (A Statistical COmputing Procedure), has been designed specifically to cope with management of data and statistical calculations. Because it is installed at the Atlas Computer Laboratory, Chilton, Didcot, Berks., it is widely available to those with no experience of or even access to electronic data-processing equipment. It is free of charge to university users in the U.K.: and to government departments at cost. The facility and service is sufficiently flexible to cope with the varied numerical needs of U.K. palaeontologists.

APPLICATION OF ASCOP

Three main approaches to statistical computation on electronic computers (text-fig. 1) have been applied in palaeontology. In the first, *the package of complete programs*

[Palaeontology, Vol. 14, Part 1, 1971, pp. 154–8.



TEXT-FIG. 1. To illustrate the three solutions in statistical computation by an example involving comparison of two different sets of data: (*a*) by a package of complete programs, (*b*) by a collection of sub-routines, (*c*) by a statistical system.

(text-fig. 1*a*), the user needs no programming knowledge but, assuming he can obtain and run the appropriate programs, he must know how each program will accept his data and organize his data compatibly. In the second, a *collection of sub-routines* enables a programmer to use a single piece of program in a number of places in his

program or even in different programs without having to rewrite it on each occasion. Given the services of a programmer, it is possible to integrate a package of sub-routines (text-fig. 1*b*) to perform many complex analyses. For the third, the *statistical system* approach (text-fig. 1*c*), no programming knowledge is required. Emphasis is placed on simplicity of presentation of the user's problem. Complex analyses, involving many stages, may be performed without the intervention of the user. In addition the user has available the expertise of the professional programmers who maintain the system.

Since the statistical system approach involves by far the least amount of non-palaeontological effort it must obviously be recommended to palaeontologists.

Full details of ASCOP, the particular system advocated here, are contained in the *ASCOP User Manual* (Cooper 1969*a*). It is possible by paraphrasing some of its contents here to give an indication of the ability of ASCOP to deal with palaeontological problems.

Data handling (Manual sections 4, 5, 6, 7 and 15). The Atlas Laboratory will process data presented on cards, paper tape, or magnetic tape provided these can be 'read' by its machine. Longhand notation is also accepted provided it is legible. As far as palaeontologists are concerned data recorded by the most complex of automatic methods is just as acceptable as data recorded on sheets of foolscap.

The basic form of ASCOP input is that of the conventional table arranged in rows and columns (i.e. the data matrix) in which the columns are termed variables and the rows are values of the variables (points); ASCOP terminology is given in parentheses. The variables are referred to by English letters, words, or combination of letters and numbers, supplied by the user while the points are numbered consecutively or carry numerical labels supplied by the user. Provision is made for missing data and for replicated variables (e.g. different values of the same variable for the same point perhaps obtained from successive measurements). Finally, in addition to 'raw' data, constant terms (coefficients), series of constant terms (parameters), and statistics such as the correlation or the variance–covariance matrices can be read directly.

Reference by means of the variable name and point label enables easy access to individual items and results in convenient organization and reorganization of data matrices. Thus matrices can be added to, or taken from, or combined with other matrices or with newly created ones whose content depends on the results of previous analyses. It is possible, from a single record of the data, to program and experiment with a complete numerical analysis without ever rearranging the original data by hand. Routine handling is done automatically and the time of the palaeontologist is spent almost entirely in designing 'experiments' with his data.

Data storage and retrieval (Manual section 14). The problem here is no different in kind from the data-handling problem mentioned in the preceding section. Storage time is simply longer. Using ASCOP, data can be stored on magnetic tape for any length of time. It is possible then for an individual who may be remote from computing facilities to build up his personal data bank. It is anticipated that widespread use of this facility would ease the demand for publication space. Reference could be made to the particular tape and an individual sufficiently interested in the content of a paper could have, on request, a print-out of the entire data on which it is based. He may even perform his own analysis of the data before receiving a print-out.

Data analysis (Manual sections 8, 11, 12, 13, 16, 17, 18, 20 and 21). Arithmetic operations such as the calculation of ratios, percentages and frequencies may be carried out and items may be plotted on any desired scale as histograms, scatter diagrams, or arranged as geographical and stratigraphical distributions.

Discontinuous variates can be treated by tabulation and associated tests of significance carried out, while continuous variates can be treated by any of a variety of techniques ranging from simple univariate analysis to component, factor and discriminant analysis. A list of some of the statistical operations available at the moment is in the Appendix to this paper.

Provision has been made in ASCOP for the writing of new sub-routines so that it is possible for the analysis section to be kept abreast of new developments in palaeontological statistical techniques: for example, the reduced major axis regression technique was inserted as a new sub-routine to accommodate palaeontological users. Were it used sufficiently frequently then it would be incorporated in the main body of the system.

SUMMARY AND CONCLUSIONS

The statistical system described is at present available and is adequate to meet present palaeontological needs in the handling, long-term storage and retrieval, and analysis of numerical data. The principal advantages to an individual using ASCOP are as follows:

1. Time is saved on the mechanical operations of statistical calculation and data handling.
2. There is access to powerful computational techniques as well as to a data store.
3. No previous knowledge of computing or local access to automatic data preparation techniques is required.
4. The entire service operates free of charge to U.K. university users and to government departments at cost.

The facilities and service described here are available at the present moment to any British palaeontologist. Were all to use them and each build up his personal data bank then there would be *de facto* a national data bank for numerical data in palaeontology. This could be used as a backing to formal publication. ASCOP, however, provides more than the ability to store data, it provides a range of statistical analyses. Use by all would facilitate a standardization of statistical technique but would also remain flexible enough for each to pursue his own method of analysis.

If numerical description becomes an integral part of palaeontological description it is envisaged that a system such as ASCOP would be integrated with a data bank dealing with other aspects of palaeontology. In the meantime it is for those who feel the need for such facilities to use them now; in due course a numerical bank of national proportions should emerge.

Acknowledgements. The writer acknowledges the help given by the staff of the Atlas Computer Laboratory at Chilton, Didcot, Berks. He is especially indebted to B. E. Cooper, the author of the ASCOP system. Dr. W. D. I. Rolfe, Hunterian Museum, University of Glasgow gave valuable advice and encouragement. This paper is published by permission of the Director of the Institute of Geological Sciences.

APPENDIX

The following is a list of the statistical operations of common palaeontological application which ASCOP currently (1970) performs:

Univariate analysis of each variable

minimum value
maximum value
mean
variance
standard deviation
skewness
kurtosis
fitting of a normal distribution
chi-squared test of normality
certain analyses of variance

Bivariate and multivariate analyses of any combination of variables

correlation and variance-covariance matrices
simple regression
multiple regression
components analysis
factor analysis
discriminant analysis

REFERENCES

COOPER, B. E. 1967. ASCOP—a statistical computing procedure. *Jl R. statist. Soc.* Series C (Applied Statistics), **16** (2), 100–10.

—— 1969*a*. *ASCOP User Manual.* Available from the National Computing Centre, Quay House, Quay Street, Manchester 3, England.

—— 1969*b*. In MILTON, R. C. and NELDER, J. A. (eds.), *The continuing development of a statistical system* in *Statistical Computation*, 295–315. Academic Press.

CRAIG, G. Y. 1969. Communication in geology. *Scott. J. Geol.* **5** (4), 305–21.

IAN E. PENN
Department of Palaeontology
Institute of Geological Sciences
Exhibition Road
London, S.W. 7

Typescript received 6 June 1970