Comp. Appl. Arch. 13, 1985.

BAILING OUT THE INEXPERIENCED COMPUTER USER: SOME RECURRENT PROBLEMS

Paul Callow

Department of Archaeology, Downing Street, Cambridge CB2 3DZ

Introduction

Since 1970, when I wrote my first program, I have spent much of my free time advising students and staff of the Archaeology Department about computing matters and often providing the custom-built software required for their work. In a great many cases requests for help have come too late to avoid a considerable waste of effort on both sides. Common sense alone cannot be relied on to steer the investigator clear of elementary pitfalls, it seems, and careful guidance is often required long before the problems are recognised. Much has been written on the danger of a cookery book application of inappropriate statistical techniques, but many archaeologists are evidently unaware of the dangers awaiting them at a much earlier stage of their work. In the 1970s the emphasis in the literature (see Doran & Hodson 1975) was towards extolling the virtues of the computer to an entirely computer-illiterate audience. In the 1980s the spread of microcomputers and a welcome trend towards user-friendly software have encouraged the archaeologist-in-the-street to have a go, and created a market for new kinds of publication such as Richards & Ryan (1985). Nevertheless, this same popularisation has undoubtedly created fresh difficulties and a major educational gap exists.

There is scarcely the space here for detailed consideration of the individual cases that I have encountered, though I have taken specific aspects of some of them to illustrate a point. The thoughts that follow are offered as meriting consideration by anyone contemplating a computer-based project, irrespective of their particular line of research or the facilities available.

Four Deadly Sins that commonly recur are:

- ignorance
- misinformation
- self-delusion
- failure to anticipate

Of course what these amount to is that the soundness of an investigation and the ease with which it is carried out depend heavily on the quality of the research design.

Ignorance

There are no short cuts for avoiding mistakes made through ignorance. To be successful it is necessary to:

- know your objectives
- know your data
- know your facilities (hardware and software)
- know your own limitations

A clear set of goals is vital. Ideally, this should include contingency planning to cover the possibility of different outcomes at each stage of the analysis.

Failure to think carefully about the questions you are asking may well cause you to obtain the wrong advice from any specialists you consult. These are usually busy people who know little about your field of interest. They cannot be expected to guess what you are trying to do. Try their patience too far and too long, their eyes may glaze over and you may be sent away with less than carefully considered advice, simply to get rid of you.

If you are creating data files you should ask yourself whether you have all the information you need, before you start putting it on the machine. If you decide later that you simply must include one more observation for each of your cases this can prove very expensive, since in order to add it to the file each new item may have to be accompanied by an identifier to link it to the appropriate record. Adding a one-digit observation to a file with thousands of records is likely to involve in addition typing as many four-digit references. Much less effort is required to include it from the start (Fig. 1), quite apart from the probability that mismatches will occur due to typing errors. You may be able to get round this by entering the new data in strict sequence, but this often means a lot of extra preliminary work anyway. On the other hand, are you really going to require all that information? It is quite common to find that a lot of one's observations do not figure in the discussion of results and could have been eliminated from the outset. Also, people working against a strict deadline, towards a university degree for example, sometimes spend so much time creating a massive data bank that they have no time left for more than a superficial analysis, possibly carried out more expensively than the same work could have been done with a pocket calculator. So do you need a computer anyway?

If your ambitions outrun the capabilities of the available equipment and software you may find yourself wondering whether it was worth collecting all that data. Or you may have to hang around waiting for someone to write the software you need. Again, if you don't do your groundwork at the very beginning you may have to waste a lot of time transcribing or reformatting data for input to the machine. We shall return to this later. It is sound practice, by the way, to ensure that the input routines actually work for your data before you start filling in reams of coding sheets!

Finally, don't imagine that you are going to learn advanced statistics, master computing, or collect and analyse huge amounts of data and write up the results, let alone all of these, in very much of a hurry. My own long-term experience, and that of some other observers, is that to produce a PhD thesis in Archaeology with a reasonably heavy component of elaborate computer work takes perhaps a year longer than an otherwise equivalent piece of research, unless you already have relevant experience. Of course a lot depends upon the nature of the project and the aptitude of the student. Small-scale database work, using a well-documented package on a micro for the kind of study that could have been contemplated by traditional clerical methods in pre-computer days, requires little investment of effort for considerable gain. However, getting your own programs to run successfully, unless they are very straightforward, takes much longer than most people expect.

Misinformation

A little help from others can save a lot of time, but as already mentioned, unless you can clearly state what you are after you won't necessarily get it. In the end it is up to you to make sure that the procedures you are planning to use are appropriate. Read the small print. An apparently ideal database package

may have significant limitations, for example, on the number of characters per field or the number of fields in each record (see also Moffett, this volume). It may be advisable to exercise caution when approached by someone waving a listing of his or her favourite program and desperate to find another application for it. It is a good idea to get a second opinion, and if there is disagreement . . . Well as a matter of principle you should make sure you understand the arguments well enough to evaluate them, shouldn't you? An example that comes to mind concernes the fitting of a set of observations to a model. A pseudo-random number generator had been used to produce thousands of values based on the desired distribution. These were then used in the production of cell counts with which those derived from the archaeological case were compared. However, the properties of the model distribution in fact permitted direct calculation of the values needed, without the slight degree of approximation implicit in the Monte Carlo method. The computer simulation was not only unnecessary but even undesirable.

Self-delusion or folie de grandeur

Though some of my earlier comments might just as well have come under this heading, I wish to concentrate on one particular area. All too frequently people fail to appreciate just how widespread is human error. More years ago than I care to remember, I was involved, in a very junior capacity, in the computerisation of a large payroll. Employees tend to object if their monthly salary cheque is short by a few pounds. To avoid a riot I and another junior employee were made to carry out a painstaking validation of the personal data held by the machine. Even though the paper tapes had been verified by being punched twice over, the results were salutory indeed. Archaeological data is more forgiving, but archaeologists are also less rigorous about procedure.

On many occasions, when someone has shown me the results of some complicated analysis and complained that they seem very odd, failure to guard against the most elementary errors has proved responsible. A typical case might arise when the number of variables, or conceivably the format of the input data, has been incorrectly specified, so that part of the information has been excluded from the computation or has been misread. Again, in a retrieval operation the logic of a complex selection instruction may not tally exactly with the user's intentions, and so give misleading results. The moral is that one should always, as a matter of routine, confirm that critical parameters have been set correctly before giving credence to the end-product. This applies just as much when the latter appears to fulfil expectations!

The point of my last remarks is that human error may invariably be expected to be present, and should be planned for accordingly. People almost always underestimate the frequency with which they make mistakes in data collection and entry. Thus on many occasions I have taken a few pages of a file pronounced clean by its owner, and called it over against the manuscript version, with mortifying consequences. Data validation is often an expensive business and it is tempting to neglect it, but you do so at your peril. It is cheaper to find the errors at an early stage than to have to re-run all your calculations several times over.

Precision in measurement is pointless if large-scale errors are allowed to stand. For instance, the use of some types of vernier callipers, with their small window, sometimes results in a misreading of the cm figure. It is all too easy to over-concentrate on the mm and smaller values. This is just the kind of error that can be trapped by the computer, either because the measurement is

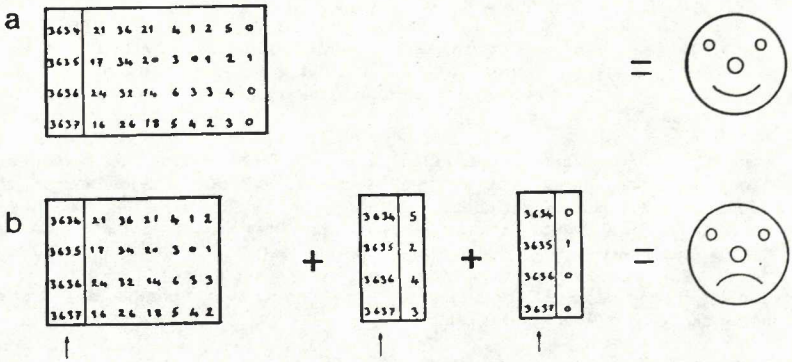

Figure 1: Adding data to a file in an unnecessarily piecemeal fashion makes extra work!

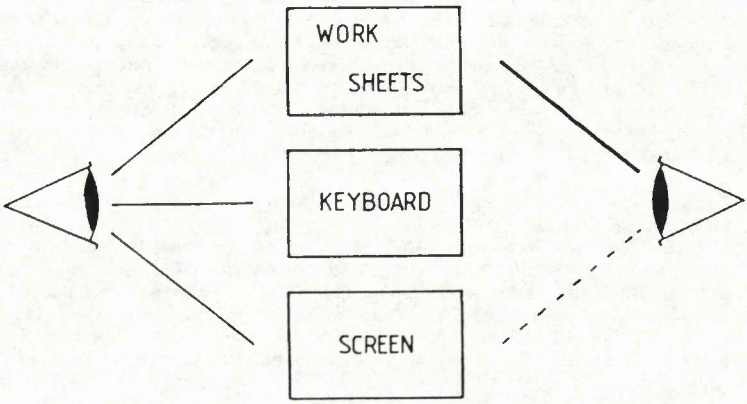

Figure 2: The eye movements of (left) novice and (right) experienced typists during data entry at a terminal.

obviously extreme, or because there is enough redundancy in the data to show that its relationship to other observations on the same object is deviant. The value of redundant information for data validation cannot be emphasised too strongly. Remember that when the computer processes information electronically, of the eight bits usually employed to represent a character one is there just to provide a check that the others are correctly set.

Failure to anticipate

It is always a good idea to explore the available software before you have collected a lot of data. You can then think your way through the proposed analysis and try to identify potential problem areas. Consider, for instance, the case of data with a strong hierarchical structure of observations, both qualitative and quantitative, and some conditional on others. At a simple analytical level this can usually be handled without difficulty by a good database package. On the other hand, if you want to do anything very elaborate by way of statistical analysis further software may well be required. As a rule, however, statistical packages are very feeble-minded about the structural complexity that they can handle. You may find it difficult to make them do precisely what you want. In the last couple of years the situation has improved slightly, but not as much as some software documentation might appear to suggest. It can be helpful to know their limitations at the beginning, when you are planning your research strategy.

One frequently encountered problem in archaeology arises over the use of multivariate statistics on data of mixed type, which by definition violate the assumptions implicit in otherwise routine procedures such as principal components or discriminant analysis. More appropriate methods may not be available in the packages offered by your installation, or may only be accessible through a combination of several such packages. Again, the clustering routines provided by CLUSTAN, SPSS, etc. cannot handle mixed data adequately. I know of no commercially-produced program that calculates the requisite similarity coefficients, for example Gower's, with correct handling of missing data. If anyone else does, please let me know! Anyone wishing to do this type of analysis often has to make special efforts to get the software.

A more common experience of this software gap is the reformatting that may occasionally be required to get the output from one package into a suitable state for reading into another. Computationally, this is a simple operation and requires minimal investment in programming skills, if you are prepared to do it yourself, but the need to do so should be identified before it arises.

The importance of planning may also be considered in connection with data entry from, say, a terminal. The keyboard skills of the person who is going to do the job should be taken into account when designing software and coding sheets. Are the data to be entered on a QWERTY keyboard or, if the data are numeric, is there a numeric keypad on the machine? Interactive data capture programs have considerable appeal, particularly those that reproduce the layout of the original on the screen. They are well suited to the continual eye movements between work, keyboard and screen that are typical of the inexperienced typist, who is likely to have to carry out a lot of correction (Fig. 2). A more experienced keyboard operator tends to look almost all the time at the work itself, and makes relatively few mistakes, so that elaborate software may be redundant or even inefficient. In fact a simple text processing program can be extremely effective for numeric data, as the use of the tabulation key, or space, comma, etc., to be transformed later into a tab, avoids the need

for careful formatting during input. However, the worksheets should be drawn up with this in mind. The time saved on a large amount of data can be very considerable.

Good pro forma design may contribute just as much as software to ensuring rapid and accurate data entry, as well as reducing the likelihood of error during the data collection phase. The very simple record sheet depicted in Figure 3 highlights some of the principles involved. Nothing about the original layout (top) helps the underpaid/bored/inexperienced finds assistant/volunteer helper to decide which of the elements must be circled or to ensure that all the required information is collected. Therefore this is a poor design, irrespective of whether a computer is to be used or not. In its hierarchical structure and differences in lettering, the revised layout (bottom) reflects the logic behind the data collection process and the scheme of analysis envisaged by the excavator. Clear distinction is made between attributes and attribute states, so that it is relatively easy to tell at a glance if anything has been omitted. A further refinement is intended to save time later, anticipating the software to be used in data entry and processing. This is the use of numeric codes (circled at the same time as the keywords). If suitable input routines are provided only these codes need be entered. Thanks to the particular scheme used, the computer can be made to check that they are read in ascending order, as a limited form of automatic data checking. In this case the data structure is implicit in the coded values. There are other ways of designing codes, as discussed by Richards and Ryan (1985, chapter 5). Many further improvements could probably be dreamt up by anyone prepared to spend time on this example.

Another kind of layout problem is exemplified by Figure 4, in which the data is binary: present-absent; yes-no; positive-negative; etc. At Cambridge the programmer's pads sold by the Computing Service are very popular with students as a low-cost source of gridded sheets convenient for data coding. A recurring mistake is to pack the data far too tightly (Fig. 4a) with the result that the slightest distraction can cause the typist to make an error. It also makes data entry very tiring. Spacing the data as in (Fig. 4b) gives only slight improvement, as it is still easy to lose one's place and, moreover, if there are a lot of observations each record may be spread out over many lines. The compromise employed in (Fig. 4c) groups the binary codes into discrete packages of constant length so that it is easy for the eye to follow the line during data entry, while for the inexperienced typist the regular rhythm, thumb and three fingers, helps to increase accuracy. The three-digit groups can be split up by the computer either when the data is first read, by format control, or later in the program, if free format is used. For sparse binary data, provided that there are no missing values requiring a third coded value, the scheme shown in (Fig. 4d) can be very efficient, although it requires careful preparatory work. Here a number is assigned to each attribute, and is entered if that attribute is in the 1 state. This approach obviously requires inclusion of a simple decoding routine in the software.

Conclusions

I hope that the preceding remarks make clear the importance of not only logic but lateral thinking in planning an investigation. These are no less important if you are working with a paper and pencil, of course. On the other hand, because of the scale of much computer-aided research and the readiness with which totally bogus results may be accepted after a heavy number-crunching session, the cost of early mistakes can be that much higher. The value of seeking expert advice cannot be overstressed, though you must always take care

EREHWON 1982 FIND No 357

SQUARE A3 LAYER 14 FEATURE 2

POTTERY WHOLE (BROKEN) RIM. BASE (BODY) COLOUR BUFF RED (GREY) BLACK. TEMPER (SHELL) SAND .....

FLINT TYPE TOOL CORE FLAKE. BUTT CORTICAL PLAIN DIHEDRAL FACETTED MISSING. WHOLE BROKEN .....

METAL BRONZE IRON .....

POTTERY	CONDITION	111 whole	(112 broken)			
		SHERD	121 rim	122 base	(123 body)	
	COLOUR	131 buff	132 red	(133 grey)	134 black	
	TEMPER	(141 shell)	142 sand			
FLINT	CONDITION	211 whole	212 broken			
	TYPE	221 tool	222 core	223 flake		
	BUTT	231 plain	232 cortical	233 dihedral	234 facetted	
		235 missing				

Figure 3: Two versions of a simplified record sheet for small finds from an imaginary excavation: (top) poorly structured and prone to error; (bottom) redesigned to facilitate accurate field recordi and data entry.

to understand the reasoning behind it, as the final responsibility is yours. Don't delay, however. Some proposals can be identified as not suitable for mathematical or computational reasons even before data collection begins. Better to find out then than later!

References

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

RICHARDS, J.D. & RYAN, N.S. 1985 Data Processing in Archaeology. Cambridge University Press.

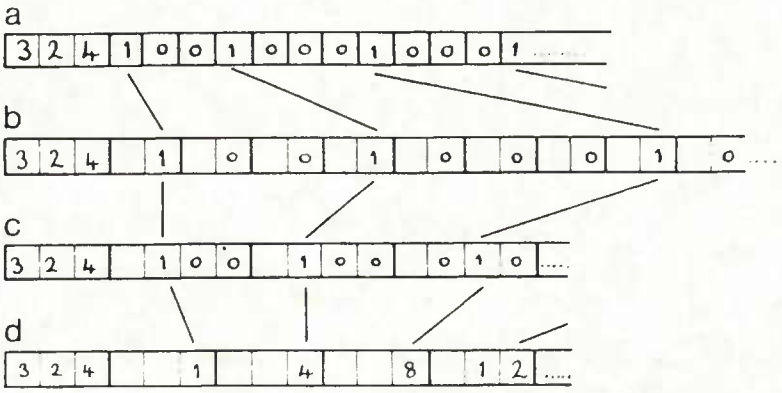

Figure 4: Some alternative ways of laying out sparse, numerically coded binary data for input.