

## Computers in research—promise and performance

By F. Yates

Delivered to The British Computer Society in London on 26 September 1961.

It is the custom in presidential addresses, whether inaugural or valedictory, to make a broad survey of some field of interest to members. Three years ago our first President, Dr. Wilkes, gave us an address entitled “The Second Decade of Computer Development” (Wilkes, 1958). Perhaps, therefore, as my address is valedictory, it will be appropriate for me to attempt a review of some facet of what has been achieved in the first decade.

I have purposely confined myself in the main to a single field of application, that of research statistics, with which I am particularly familiar. As you will see, our progress, or lack of it, in this field, has some general morals which are applicable in many other fields.

Dr. Wilkes made the important point that “The true significance of any important technological advance lies not in the extent to which it enables existing practice to be simplified or cheapened, but rather in the extent to which it makes things possible which were hitherto impossible. Digital computers are no exception to this rule, whether they are used in scientific research, in engineering design, or in business administration.”

Unquestionably the most obvious thing made possible by computers is the solution of problems involving heavy numerical computation, and it is in the solution of such problems, both those arising in pure research and in engineering design, that computers have achieved their most striking successes. In crystallography, for example, computers have made possible the determination of the structure of large complex molecules such as the myoglobin molecule, described by Dr. Kendrew two years ago at the first Conference of this Society at Cambridge (Kendrew *et al.*, 1958). They are now an essential tool in aircraft design, and enable many points of design to be resolved which previously required the construction of prototypes, a slow and expensive business, and one not without hazards. The development of atomic energy has been greatly aided by them.

Here are clear cases of making the impossible possible. But even in fields of research where the computational tasks can be, and have been, dealt with by desk calculators, computers have introduced, or are potentially capable of introducing, three very real new features. The first is the speed with which the required calculations can be effected; the second is that a much more thorough job can be done; and the third is that a good deal of methodological technique can be wholly remitted to the machine.

Often—but not, of course, by any means always—speed is of great importance in research. I am not here thinking of the obvious cases where the conclusions of some piece of applied research are immediately required as a basis for action, but rather those many instances where the research worker cannot proceed further with his investigations, whether these are in the realm of thought or of further experiment and observation, until he knows the results of his previous work. Even with unlimited computing staff, there is a very definite limit to the speed at which computations can be performed by human computers. And in most research establishments computing staff is by no means unlimited, nor would scientists be prepared to face the organizational problems if it were. The research worker, consequently, frequently finds himself in the position of having to lay aside an investigation while the necessary computations are performed. Although these fallow periods are sometimes of value, they frequently result in a transfer of interest to some new problem, so that when the results are obtained they in turn are laid aside.

As to thoroughness, those responsible for analyses of large masses of data are only too well aware of the imperfections that frequently have to be tolerated; such imperfections as inadequate editing of the data, inadequate checking of the computations, use of unsatisfactory approximate methods, failure to utilize more than a small fraction of all the data available, and failure to explore many lines of analysis that suggest themselves.

Relegation of computational methodology is equally important. Much time is at present spent in learning computational techniques of various kinds. These techniques are frequently involved, in part because of the need to economize computational effort, and of little intellectual interest. Large numbers of biologists, for example, find that they have to learn not only the principles of experimental design but also the computational techniques required to analyse the numerical results of experiments when they have been performed. These are tiresome enough, in the more complicated types of design, for the non-mathematically minded person even when everything goes according to plan. They become very much more tiresome when, through some misfortune, the results are incomplete. All this type of work can be relegated to a computer which has been properly programmed for the task. Instead, therefore, of many scientists having to learn the techniques, they

need be understood only by the few who have to prepare the necessary programs.

This lack of understanding of the methodology is not in general harmful. After all, knowledge of the methods of solving differential equations, for example, does not lead to any greater understanding of the solutions. All that the person using the results has to know is that these are in fact the solution of his equations. In many ways, also, computers simplify the computational methodology. The wide applicability of iterative methods is an example of this.

This relegation of computational methodology is likely to be of importance in many fields, since it will enable the workers concerned to devote their attention to the essentials of their subject. It is certainly of importance in statistics. The development of modern statistics has demanded more and more complex computational processes, and much of the time of practising statisticians is taken up with learning and applying such processes. But this is not their real task. They should be examining and assessing the end products of these computational processes. In so far as their energies are absorbed by computational processes, either directly or through having to train and supervise computational staff—I am thinking here of the application of known processes, not the evolution of new ones—they are not performing their real function, or at least not the function that is expected of most statisticians by those who employ them.

### Reactions of Statisticians to Computers

Initially many statisticians evinced extreme reluctance to attempt to utilize computers. This surprised me greatly. To me it has always seemed that a general-purpose computer is the computing tool of which statisticians might have dreamed all the years they have been working desk calculators. In part this reluctance stems from the feeling that only by actually working over the figures could a statistician elucidate their meaning, a feeling summed up in the remark by an eminent statistician: "I may be old-fashioned, but I do like to see the figures I am analysing."

This criticism has some substance, for in the course of analysis one does, of course, become very familiar with the data one is analysing, and with the results emerging from the analysis. Desk computing is also a restful process, and during its progress fruitful ideas often emerge. As Sir Ronald A. Fisher—who, as many of you know, is the outstanding mathematical statistician of our generation—once said to me, when discussing whether someone should have an assistant for computation: "Does she know enough to use an assistant? Most of the statistics I have learnt, I have learnt on a calculating machine."

Nevertheless in real life, faced with the analysis of a large body of data, or indeed with the routine analysis of much smaller bodies of data, such as the results of experiments, most statisticians, if they are in a position to do so, get their computations done for them by

assistants of various kinds. What is essential is that statisticians, when freed from the discipline of having to do the computations themselves, should train themselves to study the results of computations very thoroughly.

Two other reasons for this reluctance to turn to computers which are not so commonly mentioned are, I think, first that many skills which have been acquired are rendered obsolete and new skills have to be learned, and second that considerable disruption of existing computing organizations is inevitable. Some statisticians, I suspect, have let themselves become mere human computers, or directors of teams of computers, and would feel lost without computations of the kind with which they are familiar to fill their days.

This initial resistance, I am glad to say, is now breaking down, and statisticians are looking more and more to computers to solve their very real computing problems. If they are not to be disappointed we must, I think, do some hard thinking and harder work to enable computers to give the service that is to be expected of them, and of which I am convinced they are capable.

### Present Performance

In considering the statistical computations of a research nature that are at present being performed on electronic computers, we may adopt the useful sub-division of computations that were impossible by conventional methods (desk calculators or punched-card machines), computations which were barely possible and somewhat rarely undertaken, and computations which were commonly done by conventional methods, though often somewhat sketchily and often with excessive delay.

Useful progress has been made in the first class. I may mention in particular simulation studies, large-scale analysis of time series, problems of population growth, epidemic spread, prey-predator relationships, etc.; also problems in genetics such as the effects of in-breeding, linkage and selection. Monte Carlo methods have been used for the generation of statistical distributions which are intractable mathematically. On the fringe of statistics there is the important development of linear programming.

In the barely possible class there has been a fair amount of work requiring multivariate techniques (multiple regression, principal components, discriminant functions, and factor analysis).

As yet, however, computers appear to have been relatively little used for computations which were commonly done by conventional methods, even in fields where computers have now shown their superiority over these methods, such as the analysis of surveys, the analysis of experiments, and the fitting of probit lines and planes. I am not here considering descriptive and administrative statistics, mainly large-scale censuses and surveys: there has been excellent progress in this type of data processing, for example by the U.S. Bureau of the Census (Daly and Eckler, 1961), and electronic data-

processing is now being adopted for similar work in many other countries, e.g. for our own 1961 Census.

It is, of course, very difficult to assess exactly how much work previously done by conventional methods is now being done on computers. I did, however, recently have the opportunity of seeing an advance copy of a set of abstracts of available statistical programs, mainly on American machines, prepared by J. D. Alanen, G. M. Andrew, F. C. Leone and A. S. Qureishi. In general the impression gained from this list is that as yet only limited fields of application have been covered, and even in these only the simpler cases were dealt with. There are, for example, a large number of programs for regression and multivariate analysis, and a considerable number for the analysis of particular types of experiment, though many of these latter programs are clumsy and will not deal adequately with variants that are commonly encountered. There does not appear to be any good general program for dealing with the many complexities met with in the analysis of variance. There is no program for the fitting of probit planes.

Again, to take an example in the barely possible class, anyone who has had to interpret multiple classifications of data knows that the marginal means, or for quantal (yes or no) data the marginal percentages, can be very misleading if the numbers of observations in the different sub-classes are disproportionate. The following tables illustrate this point for a very simple case.

NO. OF OBSERVATIONS				NO. OF POSITIVE RESULTS			
	A <sub>1</sub>	A <sub>2</sub>	Total		A <sub>1</sub>	A <sub>2</sub>	Total
B <sub>1</sub>	100	400	500	B <sub>1</sub>	5	40	45
B <sub>2</sub>	400	100	500	B <sub>2</sub>	60	30	90
Total	500	500	1,000	Total	65	70	135

PERCENT POSITIVE			
	A <sub>1</sub>	A <sub>2</sub>	All
B <sub>1</sub>	5	10	9
B <sub>2</sub>	15	30	18
All	13	14	13½

The tables have been constructed so that the percent positive for A<sub>2</sub> is twice that for A<sub>1</sub> at each level of factor B, and the percent positive for B<sub>2</sub> is three times that for B<sub>1</sub> at each level of factor A. With the disproportionate numbers of observations shown, the marginal percentage for A<sub>2</sub> is almost the same as that for A<sub>1</sub>, and the marginal percentage for B<sub>2</sub> is only twice that for B<sub>1</sub>.

The methods by which disturbances due to the effects of cross-classifications can be eliminated from the marginal values are well known, but for quantitative data the computations are heavy, and for quantal data quite intolerable for tables of any size. Yet no program

for dealing with quantal data appears to exist in America. And so far as I know none exists in this country except on our own machine at Rothamsted.

It is interesting to note that at a paper I and Healy gave at the 33rd Session of the International Statistical Institute in Paris (Yates and Healy, 1961), Professor W. G. Cochran of Harvard University stated that they had felt the lack of adequate general statistical programs on their university computer. This confirms the impression gained from the American survey.

I also received very helpful information from the university computer units in this country. In general the amount of statistical computation (not including problems in statistical physics such as bubble-chamber analysis) is small in amount and restricted in scope. The London University computer probably covers the widest field. There 11% of computing time (290 hours) was spent on statistics in 1960–61, and 15% (318 hours) in 1959–60. For Cambridge the figure for 1960–61 was only 2.5%. This is in part a reflection of the fact that several of the Statistics Departments of London University have interested themselves in computer applications, whereas this has not occurred with the Statistical Laboratory at Cambridge.

There is no doubt that many research workers in universities require a statistical computing service, associated with advice on what computations to perform. If the university has a computer they naturally hope that such a service will be available. Are they getting it? It would seem that at most universities they are not.

There appear to be two reasons. The first is machine availability. University computers are mostly heavily loaded, and if a research worker has to get in a queue, or personally go on the machine at some unearthly hour of the night, it may be easier and even quicker to use conventional methods. It is also undoubtedly right that scarce computer resources should in the first instance be used on problems requiring heavy computing which cannot be dealt with by conventional methods.

The second, and I think much more important, reason is the absence of a powerful battery of general programs which are capable of performing, with a minimum of fuss and bother, all computations of the types commonly required. I believe that the construction of such a battery of programs must be primarily the responsibility of computer-minded statisticians, since only statisticians really know the intricacies of the analyses that have to be performed. In the following sections I will consider in a little more detail the features that such a battery of programs should possess, if a good computing service is to be provided.

### Specialized Autocodes

I think we are only now at the stage of learning how to instruct computers to perform relatively small jobs with efficiency. Our experience at Rothamsted in the analysis of replicated experiments has shown that, once they are so instructed, they are eminently suitable for

even small self-contained jobs of this type, for which desk calculators are particularly suitable. The essential point, in work of this kind, is that the computer should do the whole of the job, including preliminary computations on the data, and presentation of the final results in acceptable form. It must accept data without rearrangement in the most easily punchable form. It must be capable of handling without special programming the multiplicity of variants in design and requirements of analysis that are met with in real life. It must be capable of dealing with awkward contingencies, such as missing data.

Before we had our computer we were analysing about 400 experiments a year. Now we are analysing some 3,000 to 4,000 a year. The number of variate analyses (which are a better measure of the total computing effort) has increased from about 800 to 11,000. This has been achieved with little increase in staff, and in addition we can give a much speedier service.

Even now, however, we are not able to undertake the analysis of some of the less commonly-used types of design, and the analysis of others requires special fiddles which are both time-consuming and tricky. We still require a program for the analysis of experiments which is sufficiently general to cope with exceptional designs, and with special analyses.

The way to a program of this type has been made clear by the construction of a similar general program for the analysis of surveys. This program has an interesting history. We have been analysing surveys on the computer since 1956. Initially each analysis was specially programmed. This required considerable programming effort, which often resulted in considerable delays. Moreover, supplementary analyses, which are often required after the results of the first analysis have been examined, required further programming, with consequent further delay. A simple but limited general program for constructing basic tables for surveys using stratified random sampling was written by H. R. Simpson in 1958. In late 1959, in the course of preparing a new edition of *Sampling Methods* (Yates, 1960), I amused myself by drawing up a general autocode scheme for performing all the types of calculation which were indicated as necessary in other parts of the book. At the outset I said to myself "Let us assume we have a fine large fast computer", and consequently suffered from no machine inhibitions. At first I thought that this work had no relevance to our own small slow computer, but on further consideration I concluded I was wrong. The outcome was our present General Survey Program (Yates and Simpson, 1960, 1961).

This program has completely revolutionized the analysis of surveys on our computer. The actual tabulations and analyses required are specified by simple instructions in numerical form, and experience has shown that it can be used by workers who have no knowledge of programming. During the first eight months of 1961 some 500 hours (27% of all productive computing time) have been devoted to analysis of surveys

from many varied sources, and there is a large unsatisfied demand for such analyses, so much so that we are hoping to avail ourselves of a very generous offer by Elliott Brothers of a 402, which we shall use mainly for survey analyses over the next two years.

We are not alone in thinking that the key to successful survey analyses on a computer is the existence of a good general survey program. There is the Autostat program of Leeds University (Douglas and Mitchell, 1960). Other programs for the analysis of simple surveys, such as those used in market research, were described at the Harrogate Conference of the Society (Cook, 1960; Gosden, 1960). Ferranti are planning to write one for the Orion. (This we should otherwise have to do ourselves, as a first priority for our own Orion.) London University have recently written one for their Mercury.

I believe this represents a major break-through, though the idea of specialized autocodes of this nature is not, of course, new. Matrix interpretive routines, which are specialized autocodes for matrix work, have been in existence for many computers for some years, probably because (a) computers were early used for matrix computations, and (b) a condensed notation for such computations existed in matrix algebra; this provided a ready-made classification of the required operations, which was not available in survey analysis.

Attention has also been diverted in recent years from specialized autocodes by the development of general autocodes. Although these have immensely simplified the task of programming, particularly for problems of the mathematical type, they do not in my opinion obviate the need for programs designed specifically to deal with recurrent problems, all of one general type. For these a general program, using a specialized autocode, is the answer. It would be completely frustrating, for example, if a fresh program had to be written, even in autocode, every time a new variant of experimental analysis was required.

A general autocode can of course be used to prepare specialized autocodes, but this is a secondary matter: the main work in constructing a specialized autocode is in planning what it shall do, and how. A general autocode can also be used to supplement specialized autocodes, by the insertion of specially-programmed operations at specified points. By this means the generality of a specialized autocode can be greatly extended without excessive elaboration. In our own general survey program special functions can be included by inserting the appropriate specially-programmed or library routines. The former could be written in autocode if it were available.

### Standardization of Specialized Autocodes

There is a further question concerning specialized autocodes, that of their standardization for different machines. There are two issues here, ease of construction and convenience to users. Once an acceptable specialized autocode for a particular class of problem

has been written it is obviously considerably easier, if a detailed specification of the program is available, to use this as a basis for the construction of a similar program on a second machine, than to start afresh from first principles. If the specialized autocode has been written in some general autocode common to the two machines, then it can of course be transferred without further work. If a utilizable exact specification is available, such as in Algol, then, subject to machine limitations, an exact replica can be produced. (I say utilizable here, because, given the order code and other relevant details of the first machine, the program is itself an exact specification, but one which is extremely difficult to interpret.)

I think, however, that there are practical limitations to this idealistic approach. Many specialized autocodes require considerable machine capacity, and are designed to take full advantage of the resources of the machine for which they were written. They are therefore only immediately transferable to a machine of substantially larger capacity, in which case it may be better to write a more powerful autocode.

It is obviously convenient to the user if specialized autocodes for the same class of work on different machines adopt the same language. Premature standardization of language would, however, in my view, be a mistake. We have all seen the difficulties of constructing a satisfactory common language for general autocodes. In the initial stages of evolving specialized autocodes there is much to be said for allowing the writers of these autocodes freedom to evolve what they consider the best solution, while encouraging them to adopt common language conventions in so far as these fit in with their own ideas.

### Sequential Analysis

In research statistics it is often impossible to specify at the outset the whole of the analysis that will be required on a body of data. The analysis must in fact proceed step by step, the exact nature of the next step being determined after examination of the results of the previous step. This presents considerable problems of organization between different statistical programs, since the results at each step must be stored (clearly on magnetic tape, if available) and indexed in such a manner that the required items can be specified as data for the program performing the next step.

So far as I know, little has been done on this aspect of statistical analysis on computers. It is an important next step, which demands attention in the immediate future.

### Data Editing

As yet, also, little has been done in research statistics on the general problem of the preliminary editing of data before analysis. This is a job which is vitally important, even when the amount of data that has to be handled is quite modest. It is a job for which com-

puters are theoretically eminently suited, since once appropriately instructed a computer will perform any required tests on each item of data as it is read in, and can draw attention to anomalies, reject suspicious items, or even in some cases make the appropriate corrections. An excellent example of what is required is provided by the provision for error correction of time series data in the powerful general program for the analysis of time series (BOMM) described by Sir Edward Bullard at a recent meeting of the Society.

Here again is a task for statistical programmers in the immediate future. Provision for sophisticated data editing should, for example, form part of any good general survey program.

### Specialized versus General Computer Installations

On this issue I cannot do better than quote some passages from the paper by myself and Healy already referred to (Yates and Healy, 1961):

"The size of computer installation best suited to the needs of research statistics raises many difficult problems. The whole tendency in modern computer development is towards large and expensive machines with very high computing speeds, backed up by elaborate and equally expensive peripheral equipment. Such installations involve not only large capital expenditure, but also proportionally heavy annual maintenance costs. On the other hand, the cost per unit of computation, if the computer is fully utilized, is substantially lower, and, what is more important in many research jobs, much more powerful general programs and autocodes can be developed, and the handling of large masses of data and intermediate results is greatly simplified.

"There are great advantages to statisticians in having close contact with and control over the computational processes. Such control is clearly easiest if there is a computer on the spot. But, equally clearly, even the largest statistics departments [in research establishments] can only make effective use of a small computer, if this is employed solely on their own work. If larger computers are to be employed, either a computer in a given locality must be used co-operatively with workers in other fields, such as engineers, physicists, crystallographers, etc., or a computer installation devoted wholly to research statistics must draw work from a large area.

"The former is the solution adopted by universities in Great Britain, but from our experience at Rothamsted we have come to the conclusion that there are great advantages in having a computer installation oriented to a particular field of work. The successful application of a computer to a problem requires not only that the necessary programs shall be available but also that there shall be somebody who has a thorough knowledge of these programs and the details of their use. Some problems may of course have to be specially programmed, but the percentage of such jobs [in statistics] is in fact small. To be used efficiently, therefore, there must be a team of experts who have a thorough knowledge of the

programs relevant to their subject. Otherwise it is necessary for every person who has a problem in a particular field to acquaint himself with the details of the programs available in this field on the computer he proposes to use, or alternatively to program his work directly.

"A computer which is used for many fields of application cannot build up such a team of experts unless the staff is very large, for the experts must not only be acquainted with the available programs but must also know the appropriate methods to apply to particular problems if they are to give efficient service. At Rothamsted all our programmers are statisticians, though not all our statisticians are programmers.

"Research statistics is probably an extreme case because a very large number of research workers come up against statistical problems from time to time, but not with sufficient frequency to justify familiarizing themselves with the relevant techniques of statistics, even on desk computers. It is certainly absurd that such workers should be asked to undertake the presentation of their problems directly to electronic computers. A small number of them may wish to do so, but the majority will have neither the inclination nor the time."

I do not, of course, wish to suggest that university computer installations are inherently unsuitable for research statistical computations, only that in the past there have been inadequate computer and specialist resources for their full development. To some extent, possibly to a large extent, computer resources will improve in the next decade, though even so we must remember that demands for heavy computing work in other fields are likely to increase greatly, and universities may (rightly or wrongly) feel that statistical work should be accorded relatively low priority. But even if computer resources are adequate there remains the question of specialist resources. Here I think the answer is that any university which wishes to provide a good statistical computing service on its computer should look to its statistical department to provide the necessary experts, who should prepare an adequate battery of general programs and advise on their use, and on the preparation of data for the machine. I would not expect that all university statistical departments would have the resources, or would wish to devote the effort, to this type of work. Nor need they do so. For a statistical department or a research worker at one university can always make use of a computer which specializes in research statistics at another university or at a research station such as Rothamsted.

If this thesis of specialist computer installations, or specialist groups associated with a very large installation, is accepted it necessarily implies channelling work from many sources to one centre. This presents considerable problems of communication and organization. The computing speed of the new installation at Rothamsted, for example, will be of the order of 25–50 times that of the existing installation, and the overall speed on the types of work at present undertaken will be of the order

of 10 times. Our present system of handling computer problems will therefore require considerable revision. At present we do all the detailed work of presenting the problems to the computer; if ten times the volume of work has to be handled and all this were still to be done, a considerable increase in staff would be required, even with greatly improved general programs and a powerful autocode.

To avoid this we plan to make the use of the new computer a much more co-operative venture. The majority of users, particularly statisticians at other research stations, will be expected to write their own control instructions for jobs which can be dealt with by standard general programs; for special problems they will be encouraged to write their own programs in autocode; and they will be expected to take full responsibility for examining the results emerging from the machine. Data preparation will also be decentralized as far as possible. We expect that this co-operative use of the computer will appeal to statisticians at other research centres, since many of them would like to play a more active part in the development of computer applications in research statistics.

We shall, of course, not insist that all work is sent to the department in a form suitable for direct presentation to the computer. It would be unreasonable to ask workers with isolated problems to attempt to familiarize themselves with the intricacies of autocodes and standard programs.

## Conclusion

In this address I have dealt mainly with the application of computers to research statistics, because this is the field of which I have direct personal experience. I am certain, however, that the conclusions reached are applicable to all those fields of application in which there is at present a substantial volume of desk computing, and in which many variants of the same basic analytical processes are required.

In engineering design, in particular, there must be many small and medium sized problems which can be more effectively solved by a computer. Dr. Douglas (1959) made this point in a paper to the Cambridge Conference on Computers in British Universities: "When carrying out studies in local engineering firms I have been struck by the number of straightforward numerical problems, arising in the drawing office, which are at present inadequately investigated by hand techniques. Very often this leads to costly decisions being made on inadequate data. Many such problems are suitable for a computer, but there is no knowledge, among the relevant staff, of the potentialities of these machines, and consequently the problems do not reach the machines even though they may be available locally." To remedy this the Computing Laboratory of Leeds University "has taken over from the pure mathematics department the teaching of computation to engineers, physicists, chemists and general scientists at the undergraduate

level" with the aim of giving them "a reasonable grounding in numerical analysis and also to give them some understanding of how and when a computer can assist them in their work."

This, I believe, is only part of the battle. If engineers,

in particular, are not to be disappointed in their hopes, we must develop teams of engineering experts who have ready access to a computer, and who will make it their job to develop the necessary engineering programs and give advice and assistance in their use.

## References

- COOK, R. L. (1960). "Market Surveys with a Small Computer," *The Computer Journal*, Vol. 3, p. 140.
- DALY, J. F., and ECKLER, A. R. (1961). "Applications of Electronic Equipment to Statistical Data Processing in the U.S. Bureau of the Census," *Bull. Inst. Int. Statist.* (in the press).
- DOUGLAS, A. S. (1959). "The State of the Art—(b) Computers in British Universities," *The Computer Journal*, Vol. 2, p. 100.
- DOUGLAS, A. S., and MITCHELL, A. J. (1960). "AUTOSTAT: A Language for Statistical Data Processing," *The Computer Journal*, Vol. 3, p. 61.
- GOSDEN, J. A. (1960). "Market Research Applications on LEO," *The Computer Journal*, Vol. 3, p. 142.
- KENDREW, J. C., BODO, G., DINTZIS, H. M., PARRISH, R. G., and WYCKOFF, H. (1958). "A Three-Dimensional Model of the Myoglobin Molecule obtained by X-ray Analysis," *Nature*, Vol. 181, p. 662.
- WILKES, M. V. (1958). "Presidential Address: The Second Decade of Computer Development," *The Computer Journal*, Vol. 1, p. 98.
- YATES, F. (1960). *Sampling Methods for Censuses and Surveys* (3rd edn.), London: Griffin.
- YATES, F., and HEALY, M. J. R. (1961). "Electronic Computation and Data Processing for Research Statistics," *Bull. Inst. Int. Statist.* (in the press).
- YATES, F., and SIMPSON, H. R. (1960). "A General Program for the Analysis of Surveys," *The Computer Journal*, Vol. 3, p. 136.
- YATES, F., and SIMPSON, H. R. (1961). "The Analysis of Surveys: Processing and Printing of the Basic Tables," *The Computer Journal*, Vol. 4, p. 20.

## Correspondence

To the Editor,  
*The Computer Journal*.

Dear Sir,

Professor F. L. Bauer of the University of Mainz (private communication) has pointed out that a technique analogous to that we have described for the Givens process (*The Computer Journal*, Vol. 4, p. 177) may be applied to the Householder process.

The typical stage in the Householder process is the reduction of a matrix  $A_r$ , which is tri-diagonal in its first  $(r-1)$  rows and columns, to  $A_{r-1}$ , which is tri-diagonal in its first  $r$  rows and columns.  $A_{r-1}$  is defined by the relation

$$A_{r-1} = (I - 2w_r w_r^T) A_r (I - 2w_r w_r^T).$$

It has been shown (*The Computer Journal*, Vol. 3, p. 23) that if we define  $p_r$ ,  $K_r$  and  $q_r$  by the relations

$$\begin{pmatrix} A_r w_r & p_r \\ w_r^T p_r & K_r \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \end{pmatrix}$$

then

$$A_{r-1} = A_r - 2q_r w_r^T - 2w_r q_r^T.$$

The method suggested is that during the computation of  $A_{r-1}$ , which involves the reading of  $A_r$  from the backing store and its replacement by  $A_{r-1}$ ,  $w_{r+1}$  and  $p_{r+1}$  should also be computed.  $A_{r-1}$  differs from  $A_r$  only in rows and columns  $r$  to  $n$  and as soon as the  $(r-1)$  row of  $A_{r-1}$  is known,  $w_{r+1}$  can be computed. As each element of  $A_{r-1}$  is determined it can be used to compute the corresponding contribution to  $A_{r-1} w_{r+1}$ . When the calculation of  $A_{r-1}$

and its entry on the backing store is complete,  $w_{r+1}$  and  $p_{r+1}$  will be determined and hence  $K_{r+1}$  and  $q_{r+1}$  may be computed without further reference to the backing store. This technique halves the number of references to the backing store. (The gain is not so great as in the Givens process precisely because Householder's method is already very efficient).

It is usual to store only the upper triangle of each  $A_i$  and when this is done the computed  $(i, j)$  element of  $A_{r-1}$  must be used to determine the relevant contributions both to  $(A_{r-1} w_{r+1})_i$  and  $(A_{r-1} w_{r+1})_j$ . The working store must be able to accommodate the five vectors  $w_r$ ,  $q_r$ ,  $w_{r+1}$ ,  $p_{r+1}$  and the current row of  $A_r$  which is being transformed into a row of  $A_{r-1}$ . Further, if we are to realize the full accuracy obtainable by accumulating the inner products involved in  $p_{r+1}$  then each of its elements must be held in double-precision, since these inner products are built up piecemeal as  $A_{r-1}$  is computed. The requirement is then effectively that for six single-precision vectors.

Johansen (*J. Assoc. Comp. Mach.*, 1961, p. 331) has shown that the working storage requirements for the Givens process which we described can be reduced slightly. If transfers to and from the backing store are of  $k$  words at a time, then the requirement can be reduced from 4 vectors to 3 vectors plus  $k$  words. In a similar way the working storage for the Householder method can be reduced from 5 vectors to 4 vectors plus  $k$  words.

8 December 1961  
Oxford University Computing Lab.  
National Physical Laboratory.

Yours faithfully,  
J. S. Rollett.  
J. H. Wilkinson.