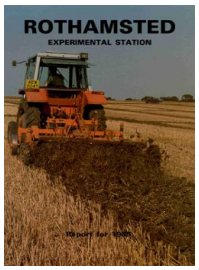
ROTHAMSTED RESEARCH

Rothamsted Experimental Station Report for 1985

Full Table of Content



The Development of Statistical Computing at Rothamsted

J. C. Gower

The Development of Statistical Computing at Rothamsted, J. C. Gower (1986) Rothamsted Experimental Station Report For 1985, pp 221 - 235 **- DOI: https://doi.org/10.23637/ERADOC-1-34074**

The development of statistical computing at Rothamsted

J. C. GOWER

Abstract

An account is given of the development of statistical computing at Rothamsted. It is concerned mainly with the period from 1954, when the first electronic computer was delivered, until the present. Initially many specialized programs were written but it was soon realized that, for efficiency, general-purpose programs each unifying many statistical techniques were required. The development of these programs was gradual and required corresponding developments in statistical theory. Now the bulk of statistical work, not only for Rothamsted but also for the AFRS as a whole, is covered by a few programs, notably Genstat which has an international market. Further developments of these programs are required to make them more accessible to scientists not well-versed in statistics and to take advantage of technological advances.

Introduction

Applied statistics involves extensive calculation which always seems to have been limited by the computational facilities available; theoretical advances in statistics often stem from observations made during the course of calculation. Thus statisticians have long had a deep involvement with computing and nowadays computers may be regarded as their laboratory instruments. Around the turn of the century the Biometric school, associated with the names of Karl Pearson, Galton and Weldon, relied heavily on Brunsviga calculators with which were computed the astonishingly extensive Biometrika Tables for Statisticians; Brunsvigas were still in use at Rothamsted well into the 1960s. A letter from Student (W. S. Gosset, whose employer, Guinness, required a pseudonym) dated September 1919, in reply to one from R. A. Fisher on his appointment to Rothamsted, advises on suitable calculating machines, mentioning the Triumphator (an improved Brunsviga), the Millionaire (an improved Tait) and the Burroughs adder on which Student comments 'unless you've a horrid lot of tots to do I fancy that is rather expensive'. Fisher did not buy the Triumphator but he early bought a Millionaire for the Statistics Department at a cost said to exceed £200, a considerable investment at the time, but commensurate with the importance of statistical calculation. Yates bought two more in 1939 and was still using one in 1980.

At Rothamsted, the 1920s and 1930s saw a rapid development in experimental designs and associated methods for their analysis. During this period an increasing number of experiments was analysed (115 in 1934 rising to 437 in 1951). Also in the period, the first edition of *Statistical tables for biological, agricultural and medical research* (Fisher and Yates, 1938) was published with the work shared between the Galton Laboratory and Rothamsted. Under Frank Yates, the 1930s saw the beginning of the use of sample surveys in agriculture whose analysis by the post-war years had become a substantial departmental commitment. The results were put on to punched cards and analysed externally at a Hollerith Bureau (Hollerith being part of the British Tabulating Machine Company, a British counterpart of IBM). In the 1948 Annual Report we find the statement:

> 'During the year the department has had the use of a sorter-counter and arrangements have been completed for the installation of a rolling total tabulator and the replacement of the sorter-counter by a sorter'.

This equipment was delivered in June 1949 when it was found that

> 'having a machine on the spot under the direct control of the research workers has resulted

in a far more enterprising and flexible approach to punch-card work than was the case when all the tabulation had to be carried out at a separate bureau'.

This claim was indeed borne out when in 1950 work was reported using this equipment (by then slightly modified) for the analysis of surveys, the analysis of replicated experiments, multivariate analysis and distributional problems. The equipment had been designed for accounting work and considerable ingenuity had clearly been exercised to cope with this range of statistical problems, many of which involved multiplication, not a built-in facility of punched-card tabulators. As was to be expected, the availability of the equipment was a stimulus to the development of statistical methodology in previously neglected fields. The equipment was already becoming overloaded and by 1951 a reproducer-summary punch had also been acquired that allowed more flexibility by providing simple facilities analagous to the storing on file of intermediate results. Nevertheless by 1952 the limitations were being felt—'One stumbling-block is the very heavy computing involved—punched card machines are valuable in this field, but are by no means a satisfactory solution to the problem'.

It is not surprising therefore to find in 1953 that Healy and Rees attended a programming course for electronic machines held at the Mathematical Laboratory at Cambridge, where there happened to be a prototype computer, the Elliott-NRDC 401, which in 1954, on the advice of a Visiting Group, was moved to the Statistics Department at Rothamsted. This article is mainly concerned with subsequent developments.

The first computer

The national picture in computing in 1954 was that important prototype computers were working in research laboratories at Cambridge University (EDSAC), the National Physical Laboratory (Pilot ACE) and at Manchester University. These machines had been running for several years and in spite of Professor Hartree's optimistic statement that two EDSACs would satisfy all the nation's scientific computing requirements for the foreseeable future, the first generation of commercially produced machines was just coming on to the market. The Elliott-NRDC 401 was itself the prototype for a moderately successful range of commercial machines. It was the first computer to be associated primarily with agricultural research and with statistics (formally 50% of its time was available to NRDC, but little, if any, of this was taken up). Isolated statistical computer programs had been written elsewhere but usually in universities and for particular research projects. Although such uses of computers were of interest the Statistics Department was also faced with the problem of how best to cope with a large and increasing amount of statistical computation of a more routine nature.

The limitations of the early computers are perhaps not familiar to the younger generation. All differed, but the 401 was in many ways typical. It could support only one user at a time, it had no compilers so that all programming was in machine code, it was supplied with no software, it had a tiny 'fast access' store (five words), backing store was switched by electro-mechanical relays, input was through a very slow and unreliable paper-tape reader, output was on to an electric typewriter. Instructions and data held on a rotating disc could be read only when they passed the reading-heads and efficient programs therefore required 'optimal programming' to ensure that the successive instructions passed the reading-heads just at the right moment, otherwise the disc would have to waste a rotation, or part of one. Because there was a single user, often the programmer himself, programs could be partially controlled from the hand-switches on the computer console. This allowed programs to be stopped and started or, as an aid to finding faults, instructions could be obeyed step-by-step (a device now commonly available in a somewhat more sophisticated form in present-day operating systems), or convergence of an algorithm could be identified by visual inspection of oscilloscope monitors and the hand-switches used to change the course of a program or initiate output.

The machine had a 32-bit word length, five fast registers (two of which formed a double-length accumulator), seven 'immediate' access tracks, each of 128 words, and 16 further 128-word tracks, any one of which could be selected by a relay. This gave a total store of 2944 words plus five fast registers, but 128 of these were reserved for the initial orders which allowed programs and integers to be read and integer results to be printed; it also contained a division subroutine. These initial orders were originally written in Cambridge but after arrival at Rothamsted, they were substantially improved by D. H. Rees. They certainly needed improvement, for the original integer input required numbers to be written backwards and the division subroutine gave the wrong answers. In 1957 Rees added three further fast registers and in 1961, eight further relay-selected tracks. For further information see Healy (1957) and Yates and Rees (1958).

All this was very primitive, and tiny, compared to the capacity and reliability of a cheap modern microcomputer, but it was a major advance on the existing punched-card equipment because it allowed flexible programs to be written for a wide variety of statistical computations. As well as essential subroutines for basic mathematical operations (divide, square-root, log, decimal reading, floating point operations etc.) the first year saw the beginnings of the development of a library of programs for standard analyses. Already a program for analysing randomized block experiments was in use (Healy) and work had started on programs for factorial experiments and Latin squares. The possibility of using the 401 for survey analysis was also being considered, but the lack of a card-reader was a major obstacle. Several research investigations were in progress including what must have been one of the first uses of canonical variate analysis. This multivariate technique, thoroughly understood since before the war, involves inverses and eigenvectors of matrices, barely possible by hand but now feasible. The problem under investigation was to try to discriminate between men and great apes by using teeth measurements, in the hope of throwing some light on the nature of australopithecine fossils. Yates sums up the first nine months experience with the 401 and gives what turned out to be a remarkably accurate forecast of the future, as follows:

> 'Having an electronic machine on the spot has made all the difference to developing its applications to research statistical problems. In this respect our experience is exactly parallel with our experience with Hollerith equipment, where we found that it was only by having equipment on the spot, so that research workers could themselves use it, test out different methods and examine the results as they were obtained, that we could exploit its full potentialities.
>
> The introduction of electronic methods of computation will make available for regular use statistical methods which at present are scarcely used because of the heavy numerical work involved. This in turn is likely to lead to major developments in method. It will also facilitate and speed up the routine analyses which are at present done on desk machines, but which are of a sufficiently standard type to be programmed electronically, and enable a much more thorough preliminary examination of the data to be made (to check for gross errors, inconsistencies, etc.) than is at present customary or possible.'

A detailed account of work done on the 401 over its nine years of life cannot be given here. Apart from the basic support software developed, and research projects that directly generated about 60 scientific papers, programs for general purpose analysis were being developed apace. These programs may be grouped into six main areas:

1. Analysis of designed experiments
2. Analysis of surveys
3. Curve and distribution fitting
4. Assay
5. Multivariate analysis
6. Multiple regression

ROTHAMSTED REPORT FOR 1985, PART 2

The development was much the same in all areas. An initial series of programs for very restricted purposes was produced, which were later improved and combined and by the early 1960s these were leading to the development of a few more general programs each of which could cope with a range of statistical computations. The remainder of this section describes this developmental sequence for the headings given above.

Analysis of experiments. A randomized block program was written in 1954 and one for the 3^3 single replicate design in 1955. By 1955 it was already clear that all programs for analysing experiments should adopt the same conventions for presenting the data and for deriving new variates needed for analysis. Thus in 1955 a program named GIED (General Input for Experimental Designs) was written and thereafter this was appended to all new programs for the analysis of experiments. Also in 1955, a Latin squares program was written, an improved version of which was made available in January 1956, followed by an improved randomized block program in February (Healy). The analysis of split-plot experiments waited until October 1956 followed in November by a program for the 2^n series of designs (Yates). This program was a breakthrough, as it could analyse experiments with confounding, partial confounding, fractional replication and even those laid out in 8×8 quasi Latin square form. The methods of analysis differed in many ways from those used on hand calculators, working in terms of deviations from block means, using efficiency factors to adjust for unequal information and using pseudo-yields to check the pattern of confounding. The analysis depended critically on the careful distinction between treatment factors and local controls, such as blocks, rows and columns, thus anticipating later developments to be described below.

All of these programs permitted covariance analysis and handled missing plots using a simple but effective algorithm developed by Healy and Westmacott (1957). A further feature was that, as well as the basic tables of estimates of treatment effects, with their standard deviations and associated analyses of variance, tables of residuals too were given. For the first time, these allowed certain checks to be done, for an exceptionally large residual or systematic pattern of residuals draws attention to what might be errors in the data or unusual field patterns—both of which require further investigation. A fuller account of this work is given by Yates, Healy and Lipton (1957).

Meanwhile GIED had been improved to extend its facilities and to make it what would now be termed 'user-friendly' and it appeared in its Mark 3 version in February 1961. Other programs were written, notably the Multiple Orthogonal Classifications program (Gower in 1958) that covered many of the more simple factorial designs and, in 1962, the first attempt at a General Experiments Program (Yates, Gower and Simpson, 1963), briefly described at the end of the next section.

The effect of these programs on routine analysis is shown in Fig. 1, where the number of experiments (variates) analysed rose from 419(834) in 1955 to 3383(18 054) in 1964. Not only were more experiments analysed, many from NAAS (the forerunner of ADAS) and the National Institute of Agricultural Botany, but there was also an increase from an average of about two variates per experiment to over five variates per experiment.

Analysis of surveys. The development for surveys is similar but differs because survey designs are less standardized than those for experiments. It was not until the end of 1956 that a primitive card reader became available. By July 1957 one of our regular surveys, the Survey of Fertilizer Practice (SFP) was being analysed on the 401 (Simpson). The story then goes—October 1957 (Amendments), April 1958 (Analysis), December 1958 (Modifications), March 1959 (Fertilizer Analysis), July 1959 (Fertilizer Rates), December 1960 (Fertilizer Analysis, No. 2). The survey then biennial (now annual) had an altogether smoother ride for the SFP analysis of 1959—October 1959 (Initial Analysis), December 1959

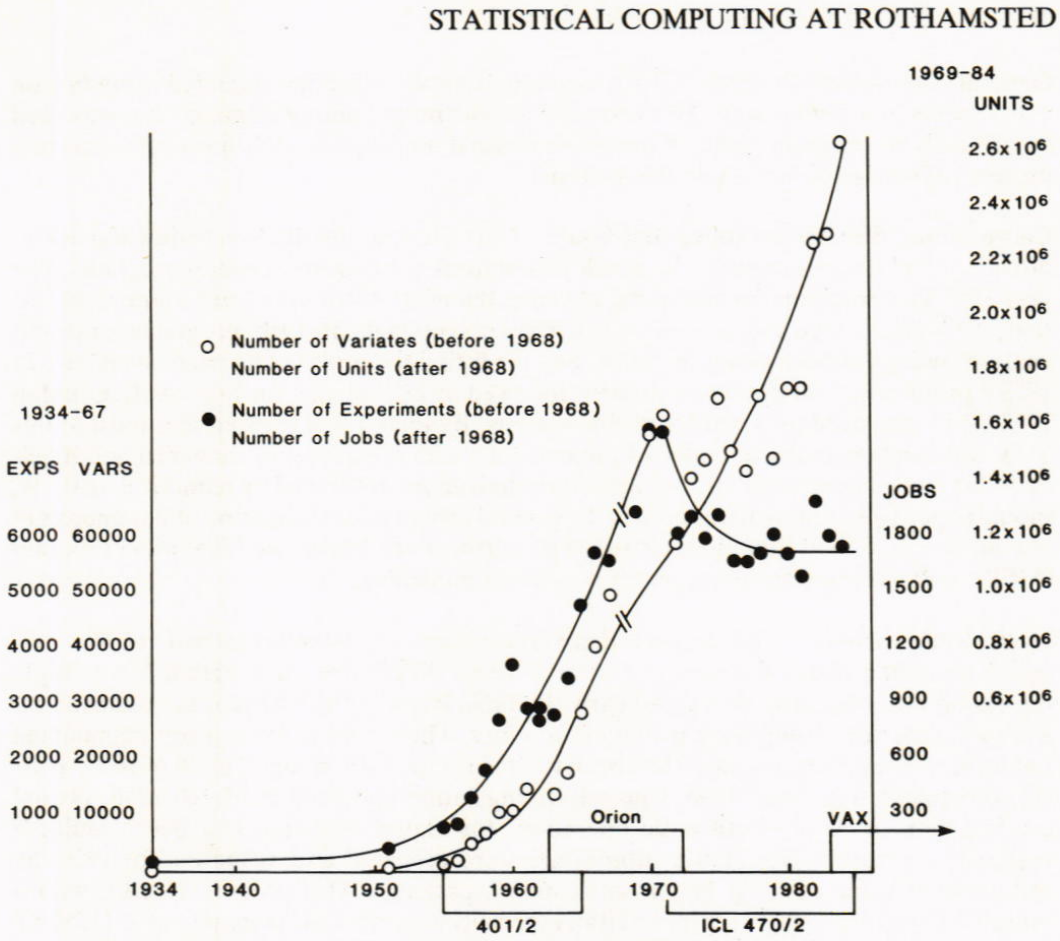


FIG. 1. Numbers of routine analyses in the Statistics Department, 1934–1984.

(Crop Analysis), January 1960 (Fertilizer Analysis). Clearly a lot had been learned, and a survey of the incidence of cattle disease was quickly analysed. By 1958 we have the first attempt at a more general survey program, one for Stratified Random Sampling. (Simpson)

Further surveys continued to be analysed. In 1959 Frank Yates was revising his book, *Sampling methods for censuses and surveys,* for its third edition during the course of which he wrote a new chapter on the use of electronic computers for survey work, which described a general system for the specification of survey analyses. He then thought that the 401 was too limited for what was required but this proved a somewhat gloomy view and by 1960 the General Survey Program (GSP) was written, which in a revised form, and now termed the Rothamsted General Survey Program (RGSP), continues to provide analyses for all our surveys. This work is described by Simpson (1961) and Yates and Simpson (1960, 1961). GSP (and RGSP) works in two parts: the first part reads in the sample data unit-by-unit, allowing for almost any generality of design, checks the data, stores them on file and forms basic multiway tables; the second part is essentially a table manipulation language operating on the tables produced by the first part.

In late 1961, Elliott Automation gave the department an Elliott 402 (the commercial development of the 401) and this was used exclusively for the first part of GSP analyses, until March 1965, when the machine was transferred to Watford Technical College. The success of GSP suggested that a similar approach could be adopted for experiments and hence the

General Experiments Program (GEP), mentioned above, which was intended to operate on experiments in a similar way. This provided a basic programming language that was used occasionally to program some of the more unusual analyses—it did, however, require a precise knowledge of how to do the analysis.

Curve fitting, distribution fitting and assay. Curve fitting, distribution fitting and assays provide yet another example of the development from specialized to general programs. The year 1957 saw programs for fitting the negative binomial distribution, exponential regression, the logistic curve and, in 1960, Chebychev polynomials. In 1956, programs for probit analysis, using desk-calculator methods, and for fitting the probit plane were available. In 1958 a probit assay program was written, followed in 1959 by one for logit analysis and in 1962 by an improved program for probit analysis. By then it was realized that most of this work was fundamentally a matter of minimizing a sum-of-squares or maximizing a likelihood, so that general methods for function-optimization, developed by numerical analysts, should cope. This approach was used in the general program for Estimation of Parameters in Maximum Likelihood Equations (Ross), a precursor of the Maximum Likelihood Program (MLP), now the standard program for non-linear modelling.

Multivariate analysis. The story for multivariate analysis has similarities with the developments described above but also has some marked differences. A program for multiple regression was, of course, developed early (by 1956, if not before). Most other multivariate analyses were done using basic matrix subroutines. Thus, even in 1955, a program named Latent Roots and Vectors (Slow) had been written for use with the apes' teeth project; a fast version appeared in April 1956. Choleski triangulation appeared in March 1956, pivotal condensation in April—both basic operations for matrix inversion and hence multiple regression. Several other matrix subroutines were prepared and at the end of 1956 the collection of some of these into a small matrix package, AUTOMAT (Healy), with a primitive control language, represented an especially important development. AUTOMAT was a subroutine package that allowed many of the classical multivariate analyses to be concisely programmed.

Work on classification, in the sense of forming classes, was new and stimulated the development of several inter-connecting programs (Gower in 1961). Basically, one program evaluated a similarity matrix for up to 128 taxa according to a general coefficient of similarity (Gower, 1971) and punched it out in coded form on to paper tape. This had to be done at least twice to check for and hence eliminate punching errors. Further programs read this tape and operated on the similarity matrix to give hierarchical classifications, summaries etc. Thus in multivariate analysis there were few special-purpose programs but methods were provided for their easy construction.

The 401 was switched off in July 1965 and the machine was taken to the Science Museum, South Kensington.

After the 401

Much was learned from the work done on the 401. Perhaps the main lesson was—'to survive, unify'. The first steps towards unification had already been taken: GIED, The General Survey Program, The General Experiments Program, AUTOMAT, Maximum Likelihood Estimation by optimization, an integrated system of Classification Programs and an integrated system of Multivariate Programs. Towards the end of the life of the 401, serious thought was given to how best to use the next computer, a Ferranti (later ICT, later ICL) Orion computer that was eventually delivered in October 1964, but did not become operational until March 1965. Because all existing programs had perforce been written in the 401

machine code, there was no question of transferring them to the new machine. Yates (1966) contains a discussion that illustrates the thinking at that time.

The possibility of unifying our programs into a few general programs was actively pursued. There were several types of unification to be considered. Firstly there was unification of special programs into more general programs. Our original programs had a variety of different input conventions which only added to the initial chaos, later rectified, of the 401 having different input and output codes—clearly data-presentation conventions needed to be standardized. Finally, the increasing importance of storing data and computed results on magnetic tape files that could be retrieved for further analyses, also required standard storage formats.

Unification of methodology directly implies unified general programs; indeed the urge to unify programs often stimulates the research needed to unify statistical methodology. Examples are the unifications needed to provide a general framework for analysis of variance algorithms (Nelder, 1965a, b and Wilkinson, 1970), the concept of generalized linear models (Nelder and Wedderburn, 1972) which contain many well-known models as special cases and the unifying concept of distance in multivariate analysis (Gower, 1966). Linking the concepts of general programs, standard input and output conventions, standard ways of describing the structure of data and standard filing structures, points to the desirability of having a unified control language.

We decided to have a single program for all non-linear model fitting and this was to be a revised and enlarged development of the 401 program which now became known as the Maximum Likelihood Program (MLP, Ross). The classification work too was now to be accommodated in a single program, the Classification Program (CLASP, Ross) which embraced everything that the separate 401 programs could do, plus an extended range of clustering algorithms and other additions. Both MLP and CLASP were, and still are, controlled by rather simple languages. Everything else was intended to be covered by a new system, the Survey and Experiments Program (SEP, Gower and Simpson). SEP, described below, proved too ambitious a project, so survey work was handled by a rewritten GSP, while multivariate analysis was covered by matrix subroutines and some small packages, the Multivariate Analysis Program (MAP, Anderson) and the Numerical Taxonomy Program (NUT, Anderson). Experiments were catered for by a few fairly general programs, the main ones being Simple Experiments (Healy), 2^n and 3^3 (Yates) and an important program for General Factorial Designs (GENFAC, Yates), which extended the ideas developed for the 2^n program on the 401 to cover a very wide range of factorial designs; all these programs incorporated a rewritten and extended version of GIED.

The Orion Computer was provided with a compiler for Extended Mercury Autocode (EMA), a language comparable, and in some ways superior, to early versions of Fortran. Mercury Autocode was developed at Manchester University in about 1954 and the extended version around 1962 but this was not available on the Rothamsted Orion until 1966. Thus MLP, CLASP, and SEP were written in machine code. Because everyone was familiar with machine-code programming this was no hardship, indeed it was regarded as an advantage because computers remained slow and the importance of having efficient programs for frequent general use could not be ignored. There was eventually a move towards EMA, in the interest of portability of programs, and most other programs used the language, but to no avail as EMA died out.

Two important programs written for the 401 have not yet been mentioned. The first was a program written in 1957 to fit constants for main effects and interactions to multiway tables. The program FITCON (Yates) was generalized in 1958 to fit quantal data (FITQUAN, Yates). Both were rewritten in EMA. FITQUAN was a forerunner of GLIM (see below) and would now be described as analysing data in tabular form, using a generalized linear model with the binomial distribution and a probit link-function.

ROTHAMSTED REPORT FOR 1985, PART 2

The Survey and Experiments Program. Although the SEP project was abandoned after two years' work, it nevertheless gave valuable experience and produced ideas of lasting significance that have greatly affected subsequent developments. As Yates wrote in the Annual Report for 1966 'It gives us useful experience on what is required in a general statistical language for future machines'.

The Algol 60 report was published when the 401 was reaching the end of its life and plans were being made for transfer to the Orion. Experience with GEP, GSP, and the matrix and multivariate programs suggested that our routine work might be effectively done by forming a subroutine package, i.e. a library of statistical subroutines that could be linked together by the user, using some appropriate language. Two major parts of any computer language are how it handles operations and how it handles operands. Both Algol and Fortran have powerful facilities for defining general operations—procedures in Algol and subroutines in Fortran. However, both are decidedly weak in defining operands. They can handle single-valued operands well, but have only very limited facilities for arrays and other multi-valued operands. This is a distinct disadvantage for statistical work, where a multitude of different structures occur (matrices of various forms, multi-way tables with or without margins, hierarchical structures which may or may not have different kinds of information at different hierarchic levels, etc.). Statisticians are familiar with these structures and often think in terms of operating on them as entities rather than operating on the single elements contained within the structures. Another thing required of a statistical subroutine package was that input, output and storage of all structures should be standardized. Accordingly SEP was specified to have the following main features:

1. To analyse surveys, experiments and multivariate data etc.
2. An algebraic control language to be compiled and interpreted (in practice influenced by Algol 60 and specified in Backus-Naur form). Conditional and unconditional jumps were included.
3. To have a block structure and procedure facility to support a subroutine library.
4. To accept data-structures as operands (scalars, variates/factors, tables, matrices, code-tables, etc.). The result of operations on elements of data-structures to be further data-structures of the same set.
5. To accept elements of operands in several modes (real, integer, name, binary, code, etc.).
6. To include multi-way table and matrix operations, with recognition of rectangular, symmetric and diagonal matrices.
7. Some special statistical and mathematical functions.
8. Excellent input/output facilities and filing of both data and programs. Thus data were acceptable in a variety of free- and fixed-formats and all output was carefully presented with proper headings and labelling.
9. The ability to operate on variates (suitable for experiments) and units (suitable for surveys).

SEP was intended for three levels of user:

1. Those who would use a standard library of SEP procedures for routine statistical analyses.
2. More expert users who would intersperse SEP-instructions between standard procedure calls.
3. The expert/research worker who would write:
   (a) the standard statistical procedures for the library,
   (b) special SEP programs for unusual analyses and research projects.

Clearly SEP was an ambitious project. It failed because it was too ambitious for the

resources available (2½ people for two years) and because when it reached a stage where it was used for some statistical production it showed itself to be unacceptedly inefficient (especially with respect to input) for the computers of the day. Furthermore, being in machine code it would have had too limited a life to be worth continuing the effort. Frederick P. Brooks, Jr, in his excellent book *The mythical man month* comments aptly on the architects of computer systems:

> 'An architect's first work is apt to be spare and clean. He knows he doesn't know what he's doing, so he does it carefully and with great restraint. . . . The second system is the most dangerous a man ever designs. . . . The tendency is to over-design the second system using all the ideas and frills that were cautiously side-tracked on the first one.'

SEP certainly had many of the characteristics of such a second system. One of its frills was to operate on sets of tables all of which could be classified by different factors and contain different margins, and different types of margin (totals, means, percentages). When required margins did not exist, or did not contain values, or contained values of the wrong kind, an impressive array of default rules determined how to proceed. These rules were very difficult to understand and worse, because they operated in default of any precise specification, they could give inappropriate settings. This experience has prejudiced me against almost all default rules in statistical programs; if the user does not know, and cannot be bothered to specify, what he wants then it is asking for trouble to get a machine to make a decision that the user may not realize could be wrong.

During this period John Nelder, then Head of Statistics at National Vegetable Research Station (NVRS), Wellesbourne, was a frequent visitor and user of the Orion at Rothamsted. Because of the variety of ways that data were collected at NVRS he developed what was termed a Three-Tier System for the Analysis of Experiments. The first tier consisted of a set of programs to read in data and convert them into a standard form. The second tier, a modification of GIED, operated on the stored data. The third tier consisted of a set of programs for analysis. Other programs in the series were concerned with the writing to, reading from and editing of magnetic tapes connected with the long-term storage of experimental data and intermediate results. Thus there was a strong concern with standardized conventions and filing formats.

In the period 1964–69 some 90 programs were written for the Orion program library. Many more, of course, were written for special research projects that had only transitory value and were not worth recording. Figure 1 shows a steady rise in the numbers of experiments and variates analysed, reaching 6124 experiments and 50 373 variates in 1967.

Genstat

Yates retired in 1968 and Nelder was appointed Head of the Statistics Department. Interest in computing was growing and Rees was appointed Head of a new Computer Department at Rothamsted. The remit of the Computer Department was to run the computer service and to deal with non-statistical computing; responsibility for statistical computing remained with the Statistics Department. Thought was already being given to the replacement of the Orion and hence what was to be done about our statistical programs, as it was clear that EMA would cease to be available.

In 1966 Nelder spent a period in Adelaide, Australia working with G. N. Wilkinson at the Waite Institute of the University of Adelaide and at the CSIRO Division of Mathematical Statistics. While there he developed ideas for the concise specification of experimental designs in terms of their separate block and treatment structures. These ideas stemmed from two papers (Nelder, 1965a, b) that unified understanding of design and had repercussions on analysis. Wilkinson developed a very general algorithm that operated on the design

specification to give an analysis of variance which in a recursive form has great versatility. In practice the algorithm was programmed in a non-recursive form, which copes only with first-order balance but which nevertheless handles a wide class of commonly occurring block designs with confounding, fractional replication and error terms associated with multiple strata. To this program was added a derived variate section, equivalent to GIED, but now expressed in a Fortran-like language. This was named Genstat (General Statistical Program) and first appeared in May 1966. The introduction to the 1970 version of the user manual indicates the scope of this Genstat:

> 'Genstat 4 is a computer program system for statistical analysis of observational (*sic*) data, developed initially at the WARI and CSIRO Division of Mathematical Statistics. The system provides general facilities for analysis of variance, multiple regression and covariance analysis, and for generating, operating on, storing and retrieving, listing and tabulating data files'.

The relation of this program to the program Genstat developed at Rothamsted after Nelder's arrival, has caused much misunderstanding. It was certainly one of the contributory strands but there were others, not least the work that had been done at Rothamsted over the preceding ten years. Probably it is now impossible to disentangle the many threads but the following is my understanding of the various influences on the design of Genstat.

1. In common with SEP, Genstat:
   - (i) Has an algebraic control language (now Fortran influenced) which is compiled into an interpretative form.
   - (ii) Has a recognized set of similar data structures. The main addition is of Factors (variates that can take only a few qualitative or quantitative values, usually coded as integers). In SEP a factor was indicated as a special attribute of an ordinary variate.
   - (iii) Has a matrix and table algebra (now much simplified).
   - (iv) Has similar input/output facilities giving great flexibility and good annotation of printed output.
   - (v) Has a macro facility. This is an alternative to the procedure facility planned for SEP and was easily provided as a simple extension of the Genstat text-handling operations.

2. Analysis of experiments came via the Adelaide Genstat:
   - (i) Using Wilkinson's analysis of variance algorithm, several times recoded and revised (Wilkinson was at Rothamsted from 1970–75).
   - (ii) Using Nelder's structure-formulae for specifying treatment and block aspects of the experimental design.

3. Multivariate analysis:
   - (i) For multiple regression, much influenced by Anderson's MAP program and its predecessors.
   - (ii) Standard multivariate directives for components analysis and canonical variates analysis were also in the style of MAP.
   - (iii) The standard matrix-algebra facilities of the Genstat language, supplemented by the multivariate directives, give a powerful language for writing multivariate macros, of which some 30 are now in the Genstat macro library. This is essentially the sub-routine package approach envisaged for SEP, which allows new statistical programs to be written without extending the language itself.
   - (iv) The classification directives are a direct transfer from CLASP.

4. File handling:

   (i) Some ideas have come from the three-tier system, perhaps via Adelaide.
   (ii) Others to do with the storage both of data and library programs were possibly influenced by SEP.
   (iii) Technical details have been heavily influenced by the methods used in Fortran.

Later major additions to Genstat were, in 1977, Generalized Linear Models, through GLIM, a project formally of the Royal Statistical Society with which Rothamsted was much involved (see below) and in 1979, Time Series Analysis (Dr G. Tunnicliffe-Wilson, University of Lancaster). In 1981 the optimization sections of MLP were incorporated into Genstat.

Initial development of Genstat at Rothamsted was much hampered by problems with the computers. The programming language was to be Fortran, so the Orion Computer could not be used and it was not until November 1970 that the new ICL 4-70 was commissioned in the Computer Department and this gave trouble for several months in 1971. Thus development had to be done remotely, initially through a bureau in London and then using a card-reader/line-printer link with an IBM machine in the Edinburgh Regional Computing Centre.

Early versions of Genstat were available in 1971 but it was not until March 1972 that the system became generally available, after which it was quickly to become the standard statistical computing language of the A(F)RS Institutes. The outstanding success of the Genstat project owes much to John Nelder's leadership and to his many contributions. The whole project was (and continues to be) overseen by a committee; Nelder was chairman until his retirement in 1984.

With the development of Genstat most of the Statistics Department's computational needs for routine analysis were accommodated and the Genstat language provides a powerful tool for programming research problems and for assembling a macrolibrary for statistical computations of a less routine nature. Currently there are 43 programs in the macro library. Thus Genstat provides for the three levels of user mentioned above.

Major versions of Genstat are termed marks and minor changes are termed releases. This gives a reference system in which Genstat 4.03 means release 3 of the mark-4 version. The 1972 version was Genstat 2 (with about nine internal releases); July 1973 saw Genstat 3(.01) with about two releases a year until version 3.09 was written in 1976 and released as 4.01 in the following year. New releases of mark 4 were less frequent and the current official version is 4.04, released in March 1983. Over the intervening years the accumulated effects of revisions, additions and corrections had produced a language with inconsistencies, so that simple completely general rules for its syntax could not be specified, with corresponding problems in learning that lead to a tendency for users to make errors. Accordingly in 1983 work began on a major revision to define Genstat 5.01, with a completely consistent syntax. The facilities for graphics are being made considerably more flexible by building on GKS, the international standard for interfacing with the great variety of graphics terminals now on the market. Also the program is being made fully interactive, which will greatly improve the ease of exploratory data analysis. Work is already well advanced on Genstat 5.01, which it is planned to release in 1986.

Genstat soon acquired users outside the AFRS and now has 388 installations in 35 countries. This success has brought associated problems. To run on a wide range of computers (some 30 models) great attention must be paid to providing portable program code. This means that programming must be in standard versions of Fortran with any sections that might be machine-dependent being carefully flagged. Then there are the problems of actually providing the different versions from the master-code; this has been handled through liaising with special implementors, often at university sites, and by providing a special conversion program to select appropriate variants of the code. Thus at external sites,

the releases mentioned above may occur months, or even years, after they have occurred at Rothamsted, so that, inevitably, several different versions of Genstat currently coexist and must be maintained. A scale of charges and suitable legal contracts have had to be worked out. A back-up service has to be provided to handle queries and reports of real or imagined errors. Sometimes users request special facilities, which have to be considered in the light of their more general utility. A great deal of time has been taken up with providing good documentation both for users and for implementors but much remains to be done (Alvey, Galwey and Lane, 1982; Alvey *et al.*, 1983). There is a continuing demand for courses on the use of Genstat both from within the AFRS and from external users. All this makes considerable demands on the Department, which has been unable to use the revenues from Genstat for its support and further development.

Some alleviation of this problem was made in 1979 when the Numerical Algorithms Group (NAG), a non-profit making organization based in Oxford, agreed to take over distribution both of Genstat and its documentation and also to act as a first line of defence for queries; since 1985 NAG has also taken on the administration and coordination of most of the implementations of the different releases for different machine ranges.

Other programs

Genstat caters for most of our needs but the language is not particularly well-suited for surveys because it has only primitive facilities for unit-by-unit operation. Thus RGSP remains our main survey program; it has been much improved and is now fully portable (see Yates (1949), fourth edition, 1981). Although many of the facilities of MLP, GLIM and CLASP are now integrated with Genstat, the freestanding programs are rather more efficient and offer a few extra facilities, and so remain in use. A further program GENKEY, whose genesis lies around 1964, was developed in a fully operational form by R. W. Payne. GENKEY calculates identification keys for groups of organisms and prints them out in a directly usable form. Thus this is a specialized program but it has many potential uses; the principal use so far has been to construct keys for 469 species of yeast and to produce, via the laser-printing service at Oxford University, a finished book published by Cambridge University Press (see Barnett, Payne and Yarrow, 1983). A few programs produced elsewhere are occasionally used and some research projects are most conveniently programmed in Fortran or another language at a similar level.

The Generalized Linear Model project, which produced the first version of the program GLIM in 1973, occupied several members of the Department. It was officially a project of the Royal Statistical Society's Working Party on Statistical Computing, with John Nelder as Chairman. The program FITQUAN mentioned as being developed on the 401 is a special case of a generalized linear model (GLM). The special feature of a GLM is that the observations are assumed to belong to a statistical distribution whose expected value can be expressed as a *function* of a linear function of the parameters concerned. It has long been known that the maximum likelihood estimation of the parameters of such models can be computed iteratively as if the problem were one of weighted least squares but with the weights changing on each cycle of iteration. In an important paper, Nelder and Wedderburn (1972) unified the theory, especially for the exponential family of distributions which covers most of the commonly occurring cases. The Working Party set about programming this approach with Nelder, Wedderburn and Rogers, and later, Baker, Payne and White all making major contributions. The associated theory is described by McCullagh and Nelder (1983). Since 1974 NAG has handled the marketing of GLIM.

These other programs, especially MLP and GENKEY, have developed over the same period as Genstat and have generated similar administrative problems. In 1984 NAG took on the distribution and conversion of MLP. In 1985, because of lack of resources, work at

Rothamsted associated with the GLIM project was reduced to a residual level. Most generalized linear model analyses are done using Genstat but it is anticipated that GLIM will continue to be used.

From 1968 onwards the method for collecting annual statistics on routine jobs done on the computer was changed from numbers of experiments and variates analysed, to numbers of jobs successfully run and units of data submitted. The first method was based mainly on experimental analyses, while the latter method includes many other types of statistical jobs of a routine nature. Thus the two sets of figures are not comparable. Nevertheless an attempt has been made to include both in Fig. 1. Whether or not the peak shown for jobs run in 1970 and 1971 is real, is hard to say. The new Computer Department took on some data preparation and other A(F)RS institutes began to prepare their own data and submit their jobs directly, which could account for a significant drop in the early 1970s. The number of units analysed has fluctuated fairly wildly, but there is an overall upward trend reaching well over two million units in the past three years; we have no figures for the AFRS as a whole. The number of jobs also fluctuates but about a steady state of about 1800 jobs. Hitherto all data has been entered on keyboard-controlled equipment—firstly on a variety of paper-tape machines, then punched cards and now on floppy discs. In all these methods, careful procedures are required for checking that data has been entered correctly. Nowadays, data may arrive on floppy discs or cassettes and there is a move (first noted in 1959) towards direct entry from data-loggers or laboratory instruments. At first sight this would appear to remove the need for checking, but it is becoming clear that automated data too can be erroneous and that methods must be devised to control their quality. This is particularly important when such data is 'unseen by human eyes' so that even major discrepancies can pass unnoticed. The checking of data quality is an unglamourous, but important, aspect of statistical computing because the most penetrating statistical analysis using the latest computing equipment is futile, or worse, if the data are of poor quality.

It is not only in numerical aspects of computing that the computer has been useful to statisticians. Since 1947 the Department has been concerned with producing an annual 300 page publication of the Yields of the Field Experiments. Originally this was typed in its entirety but we gradually mechanized the task, and since 1981 the text has been edited on a word-processor (using the previous year's files) inserting the numerical part into the text by transferring files produced by Genstat analyses. This has not only saved several weeks' work but also gives a more accurate and better printed production.

Conclusion

The work begun in 1954 has progressed logically to provide a few general portable programs, the most important of which is Genstat, that handle both the research and routine statistical calculations required by the AFRS. Because statistical principles are of general applicability, the use of these programs is not confined to agricultural research but is also of value to applied statisticians working in medicine, environmental science, in the social sciences, psychology, industry, local government, etc. The guiding principle has been to unify (i) the analysis-specification and data-description level and (ii) at the statistical level. The first level of unification has implied a unified control language with good operand (data-description) definitions and methods for operating on these operands. The operations may be controlled by many parameters, some of an optional nature, that should be specified in a concise and simple manner. Unification at the statistical level has depended on statistical research within the Statistics Department into generalized linear models, multivariate data-analysis, analysis and design of experiments, estimation and optimization and into the theory of diagnostic keys.

It has been amply demonstrated that the design of effective statistical computing systems

requires extensive statistical knowledge and research. Certainly Genstat could never have been written without the considerable statistical experience and expertise available in the Statistics Department. One has only to examine the appalling statistical packages produced commercially for microcomputers, which can only have been written by non-statisticians, or perhaps by very inexperienced statisticians. Fortunately microcomputers are increasing in capacity and Genstat is already available on machines in the personal computer range.

In case readers get the wrong impression, I must emphasize that the staff resources available for work on statistical software are meagre. The number of posts concerned fluctuates enormously from year to year, depending on whether or not special effort is being put into developments or whether it is maintenance that is mainly required. At peaks, up to 12 persons have been involved but all have substantial other commitments in statistical consulting and research; an average number of posts is nearer four and a half.

Other groups of statisticians might have done things differently and indeed they have. In the USA the Biomedical Programs BMD(P) package was developed at the University of California, Los Angeles, the Statistical Programs for the Social Sciences (SPSS) at the University of Chicago and the Statistical Analysis System (SAS) at the Biostatistics Department of the University of North Carolina in Raleigh. All these are now profitable commercial enterprises. The system S developed at the Bell Telephone Laboratories, New Jersey, (now AT&T Bell Laboratories) has profited from its relation to the UNIX system and is oriented more exclusively to the exploratory data analyst and research statistician than also to those who do more routine analyses. I believe that Genstat compares well with the best of these, by providing better statistics, more flexibility and by using computing resources more efficiently. We look forward to the next 30 years of statistical computing at Rothamsted.

References

Alvey, N. G., Banfield, C. F., Baxter, R. I., Gower, J. C., Krzanowski, W. J., Lane, P. W., Leech, Penelope, K., Nelder, J. A., Payne, R. W., Phelps, Kathleen, W., Rogers, C. E., Ross, G. J. S., Simpson, H. R., Todd, A. D., Tunnicliffe-Wilson, G., Wedderburn, R. W. M., White, R. P. & Wilkinson, G. N. (1983) *Genstat: a general statistical program.* Oxford: Numerical Algorithms Group.

Alvey, N. G., Galwey, N. & Lane, P. W. (1982) *An introduction to Genstat.* London: Academic Press.

Barnett, J. A., Payne, R. W. & Yarrow, D. (1983) *Yeasts: characteristics and identification.* Cambridge: Cambridge University Press.

Brooks, F. P. (1978) *The mythical man-month: Essays on software engineering.* London: Addison-Wesley.

Fisher, R. A. & Yates, F. (1938) *Statistical tables for biological, agricultural and medical research.* 2nd edition 1942, 3rd edition 1948, 4th edition 1953, 5th edition 1957, 6th edition 1963. Edinburgh: Oliver and Boyd.

Gower, J. C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53,** 325–338.

Gower, J. C. (1971) A general coefficient of similarity and some of its properties. *Biometrics* **27,** 857–871.

Gower, J. C., Simpson, H. R. & Martin, A. H. (1967) A statistical programming language. *Applied Statistics* **16,** 89–99.

Healy, M. J. R. (1957) The Electronic Computer at Rothamsted. *Rothamsted Report for 1956,* pp. 229–235.

Healy, M. J. R. & Westmacott, M. H. (1957) The problem of missing values in experiments analysed on automatic computers. *Applied Statistics* **5,** 203–206.

McCullagh, P. & Nelder, J. A. (1983) *Generalized linear models.* London: Chapman & Hall.

Nelder, J. A. (1965a) The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. *Proceedings of the Royal Society of London* **A283,** 147–162.

Nelder, J. A. (1965b) The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. *Proceedings of the Royal Society of London* **A283,** 163–178.

Nelder, J. A. & Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society* **A135,** 370–384.

Simpson, H. R. (1961) The analysis of survey data on an electronic computer. *Journal of the Royal Statistical Society* **A124,** 219–226.

Wilkinson, G. N. (1970) A general recursive procedure for analysis of variance. *Biometrika* **57,** 19–46.

Yates, F. (1949) *Sampling methods for censuses and surveys.* 2nd edition 1953, 3rd edition 1960. London: Griffin.

Yates, F. (1966) Computers, the second revolution in statistics. (First Fisher Memorial Lecture). *Biometrics* **22,** 233–251.

YATES, F., HEALY, M. J. R. & LIPTON, S. (1957) Routine analysis of replicated experiments on an electronic computer. *Journal of the Royal Statistical Society* **B19,** 234–254.
YATES, F. & REES, D. H. (1958) The use of an electronic computer in research statistics: four years' experience. *Computer Journal* **1,** 44–58.
YATES, F. & SIMPSON, H. R. (1960) A general program for the analysis of surveys. *Computer Journal* **3,** 136–140.
YATES, F. & SIMPSON, H. R. (1961) The analysis of surveys: processing and printing of the basic tables. *Computer Journal* **4,** 20–24.
YATES, F., GOWER, J. C. & SIMPSON, H. R. (1963) A specialized autocode for the analysis of replicated experiments. *Computer Journal* **5,** 313–319.