



GENSTAT NEWSLETTER

NO. 6

JUNE 1980

Genstat is now used at so many centres that Rothamsted have asked the Numerical Algorithms Group to take over the licensing and distribution of the program. As NAG have considerable experience in promulgating the NAG Library and GLIM the service to users should be much improved. Queries about the program should now be directed to the Statistical Package Coordinator, NAG Central Office, 7 Banbury Road, Oxford OX2 6NN (Telephone: Oxford 511245). Rothamsted are still willing to answer queries about the technical content and use of Genstat.

Since the last newsletter was issued an interactive version of Genstat 4.01 has become available for Prime (400+) computers. The assistance of John Parkinson and Chris Partridge of Prime Computer International is acknowledged. A version for the Siemens computer has been produced by Dr. Jan v.d. Vooren of the Glasshouse Crops Research Station, Naaldwijk, Netherlands; this version is distributed by Siemens Data B.V.

Versions of 4.02 are available for the IBM 360/370 range (distributed by Nick McLaren of the Cambridge Computer Laboratory); CDC 7600 (converted by John Lloyd-Jones of Manchester) and CDC 6000 (converted by Peter Baghurst of Adelaide). Versions of 4.02 are being prepared by Annie Bouvier (CNRZ, France) for the Isis 80 and John Lemon (Aberdeen) for the Honeywell (GCOS).

We hope to provide versions of 4.03 for 15-20 different computers, including the MODCOMP Classic 7870, TELEFUNKEN TR 440 and VAX.

CHANGES IN RELEASE 4.03

There have been several minor changes introduced in release 4.03 and these are listed below:-

- Cluster Analysis: output can be labelled with integers or names
- 'INPUT' directive: this now has new options for skipping records and for rewinding
- 'OPTIMISE' directive: the options have been re-ordered; also operations on restricted data structures and missing values are now allowed
- 'PRINT' directive: new option to control printing of unknown value cell of tables
- 'READ' directive: new option to return to a specified input channel after an error
- 'REGRESS' directive: the RECY option has been changed
- 'TABULATE' directive: the default setting of the NAME option has been changed and the tables can be specified using nameable lists.

The most important addition for release 4.03 is that of Box-Jenkins time-series analysis; also a new histogram directive, smoother lines in graphs and additional table functions for the 'CALCULATE' directive have been included. Short items on each of these follow.

ADDITION OF TIME SERIES FACILITY

Genstat 4.03 will contain directives for fitting autoregressive integrated moving average (ARIMA) and transfer function (TF) models. Seasonalities, Box-Cox transformations and missing values are also catered for.

Various aids to time series model-building are included; calculation of autocorrelations, partial autocorrelations, cross-correlations and chi-squared statistics to test for lack of correlation; also preliminary estimation of parameters.

Parameter estimates and standard errors are obtained by maximizing the likelihood function, and they can then be used for forecasting future values of the time series.

Joe Pearlman
R. E. S.

A NEW DIRECTIVE FOR DRAWING HISTOGRAMS

It is possible to draw histograms using the GRAPH directive but this can be an unnecessarily complicated procedure, often involving the use of the GROUPS directive.

Histograms can now be produced very easily using a new directive HISTOGRAM. The simplest form of the directive is

'HIST' variate name

which will use a number of groups equal to the square root of the number of values of the variates. The basic form of the histogram is as follows:-

```
HISTOGRAM OF variate name
- 10  3 ***
10 - 20  8 *****
20 - 30  5 *****
30 -    1 *
```

The histogram can take more complex forms, allowing you to specify the number of groups and the group labelling.

The current facilities for drawing histograms using GRAPH will remain available in release 4.03 but, if we receive no strong representations to the contrary, they will be withdrawn in release 4.04. Please let the Genstat Secretary have your views. If you do not, you may lose a facility which is useful to you.

Stuart Watson
R. E. S.

SMOOTH LINES IN 'GRAPH'

In addition to point plots and line plots, smooth-line plots can now be produced by the 'GRAPH' directive. The graph produced by

```
'HEAD' LINE = 'L'
'GRAPH' X ; Y $ LINE
```

has successive pairs of points joined to one another by straight lines. It is now possible to put

```
'HEAD' SMOOTHLINE = 'S'
'GRAPH' X ; Y $ SMOOTHLINE
```

This will produce smooth curves calculated by fitting cubic splines.

Stuart Watson
R. E. S.

NEW TABLE FUNCTIONS IN CALCULATE

Up to now there have been two special table functions TTOTAL and TMEAN which calculate totals and means respectively over any factors indexing the table used as the argument of the function but not indexing the resultant table. The functions TMIN, TMAX, TVAR, TMED and TMV have been added which calculate minima, maxima, variances, medians and number of missing values respectively. TSUM is now a synonym of TTOTAL.

These functions can be used to calculate marginal values for a table. For example

```
'CALC' T = TVAR(T)
```

where T is a table with margins will calculate the between cell variances for the margins of T.

Norman Alvey
R. E. S.

NEW LIBRARY MACROS

There have been eight additions to the Macro Library and six of these are concerned with multivariate analysis.

INDSCAL performs an individual scaling analysis; SVD3 gives a canonical decomposition of a three-way table, and is used by INDSCAL; BIPLTV provides a biplot for a list of variates; BROWNE provides the basis of an oblique Procrustes rotation; GPROCPLT provides various graphs to aid interpretation of a generalised Procrustes analysis; GPROCLAB generates plotting labels for GPROCPLT.

If each of m individuals assesses the similarities between n objects Carroll & Chang propose individual scaling analysis for the m similarity matrices, each of size n x n. The analysis locates the n objects in an r dimensional space, based on an average similarity; also, the m individuals are located in an r dimensional space. The user must supply a list of similarity matrices, the three values, m, n and r, and three option settings. The two sets of locations are returned as matrices, also full tabular and graphical output is (optionally) available.

The canonical decomposition of a three-way table, being a generalization of singular value decomposition, can be expressed as:

$$x_{ijk} = \sum_{t=1}^r u_{it} v_{jt} w_{kt} s_t$$

If X is the (n x p x q) table the u's, v's and w's are held in the matrices U(n x r), V(p x r) and W(q x r); the s's are held in the (r x r) diagonal matrix S. The U, V and W matrices are normalised but are not, in general, orthogonal. The user must supply the three-way table (without margins), the values n, p, q and r, and the maximum number of iterations required. If the iterative process has not converged by that stage a warning is printed, together with the current matrices U, V, W and S and these matrices from the previous iteration. There is no other printed output. The four matrices are returned to the user.

The biplot allows inter-individual and inter-variate relationships to be displayed on one graph. The macro BIPLTV provides either or both of two forms of the biplot for a set of variates, each of which may be relocated to zero mean and optionally rescaled to unit variance. The user must supply the list of variates, a names vector to label the points for the variates, and three scalars giving the graph and transformation options and the number of variates.

The oblique Procrustes rotation described by Browne can be reduced to a series of constrained regression problems; the constraint being that the regression coefficients have unit sum-of-squares. Given the usual regression matrix (X) and vector (y) and their sizes the macro BROWNE calculates, prints and returns to the user the coefficients, fitted values and residual sum-of-squares.

The interpretation of a Generalised Procrustes analysis (as provided by the macro GENPROC) can be assisted by the series of plots produced by the macro GPROCPLT. The user must provide the initial and final configurations, three scalars defining the size of the data, headings for the configurations and objects, and seven names structures containing plotting labels. These plotting symbols can be automatically generated by the macro GPROCLAB, which needs two scalars giving the numbers of configurations and objects. GPROCLAB will optionally (as defined by a third scalar) print the labels in a form suitable for the user to create a key to the graphs from GPROCPLT.

We have had to make a slight change to CLASSF; however this has no noticeable effect.

A small change has been made to D3PLOT: it now draws the z-x base before plotting the points. Also, I find it easier to visualise the points in 3-dimensions if I draw in extra lines to create the impression of looking into a cuboid, e.g. a squash court viewed from the centre of the back wall. On the Genstat example for D3PLOT the extra lines join points as follows: 03-07-08-04 mark the edges around the top of the cuboid; 07-05 and 08-06 delimit the walls.

Pete Digby
R. E. S.

In addition there is a macro PERMLEV for permuting, combining or omitting levels of one or more of the factors classifying a table.

Linda Bassill
R. E. S.

Finally, there are a number of macros contained in the subfile NPARAMCS for calculating some of the most commonly used rank tests. These include those for Kendall's coefficient of concordance, the Kruskal-Wallis one-way analysis of variance and Spearman's rank correlation coefficient.

Nick McLaren
Cambridge

REPORT ON GENSTAT CONFERENCE, CAMBRIDGE, APRIL 1979

Few statistical computer programs have yet reached the stage where a whole conference can be devoted to them: for fewer still would such a conference bring together those who originated the program, those who have converted it to various ranges of computer and those who use it for solving practical problems. This however was achieved for Rothamsted's General Statistical Program Genstat at a conference held in Cambridge from April 9th to 11th, 1979, and attended by nearly a hundred people from six countries. Those taking part came from universities and from research stations and other institutions in both the government and private sectors. This very successful conference was superbly organised by N.G. Alvey with support from other members of Rothamsted Statistics Department. The success was also due in large measure to the arrangements provided by Emmanuel College, where most of the participants ate and resided.

The conference began with an introductory survey by J.A. Nelder of Rothamsted Experimental Station who, with G.N. Wilkinson, was originally responsible for Genstat. At an early stage in the development of the program data manipulation procedures were introduced and now many different types of data structures are supported, far more than simply data matrices. This method of working was essential, since Genstat was designed to be, and has remained, a single statistical program of great generality rather than a set of routines for doing particular analyses. Dr. Nelder acknowledged that, by the nature of the program, documentation is very difficult; however, an improved manual is now available, the newsletter appears regularly, there is a macro library and a computer file of examples. An interactive version of Genstat is also being developed. Further, Genstat is being evaluated independently, in studies such as that in progress at Cornell, and it is found to compare very favourably with other statistical programs and packages.

Recent developments in the fitting of generalised linear models and optimisation procedures for fitting non-linear models were described by P.W. Lane of Rothamsted. The generalised linear model fitting facilities of the interactive program GLIM have been incorporated into Genstat, while the optimal fitting of non-linear models is related to the maximum likelihood program MLP. In this way features of two other statistical programs have been introduced into Genstat. The generalised linear model part of Genstat permits a considerable expansion of the regression directives, and is already proving valuable. For non-linear models the procedure is somewhat more difficult to operate; Mr. Lane illustrated it with a practical example of a double Langmuir isotherm with two linear and two non-linear parameters.

An entirely new development for Genstat will be the introduction of facilities for Box-Jenkins time series analysis. These were described by G. Tunnicliffe-Wilson of Lancaster University, who has worked closely with G.M. Jenkins. There will be many applications including forecasting, analysing structures, testing relationships and extracting components of series. Dr. Tunnicliffe-Wilson gave several examples of the sorts of series it will be possible to analyse; he stressed that this development is still in the future as far as Genstat is concerned, but should occur within a year.

Genstat may be used not only as a program in its own right, but also before or after other programs. C. Jameson of East Malling Research Station described the use of Genstat to validate and process input for other programs,

for which purpose it is very useful because of its power and flexibility in data reading and manipulation. Mrs. Jameson gave as an example an experiment on plantains in Nigeria where only 45 out of 63 plots survived in what was a non-orthogonal design initially; the data were read into Genstat and output in a form suitable for analysis by East Malling's general non-orthogonal analysis program NOBLE. Moving along the West African coast she described the validation of possibly unreliable data on Ghanaian cocoa. By contrast J. Beasley of Rothamsted described how Genstat could be used to process output from other programs, taking as examples data from Rothamsted's general survey program RGSP, which is more suited to the input of hierarchical data, and from GLIM.

A view from a computer centre was then given by G.R. Field of the University of Manchester Regional Computer Centre. In a university environment the criteria for selecting a statistical program are geared to teaching and so differ from those considered by research workers. Dr. Field gave figures to show that Genstat was not used nearly as much as SPSS - to use a four-letter word that almost seemed like other four-letter words at this conference. He felt that the marketing of SPSS, in particular its ease of use and simple documentation, was the chief reason for its wide use, even though Genstat may be better for many statistical purposes.

The first day's programme was rounded off by a general discussion chaired by G.H. Freeman of the National Vegetable Research Station; the subjects covered included methods for making Genstat available, difficulties encountered in using it and the education of staff in its use. Identifying Genstat users was felt to be one problem, especially by participants from the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO), where there may well be up to 200 users. A problem oriented approach often works in introducing new users to the program; nevertheless, for many non-statisticians the output is like "giving dynamite to people who are not mining engineers", as one Australian put it. The usual introduction for new users is by copying other programs, by a macro or by filling in blank spaces in a form. None of these is completely satisfactory, but a new trend towards teaching Genstat before Fortran may help.

The second day's programme began with a journey into the internal workings of the Genstat compiler with H.R. Simpson of Rothamsted. With Mr. Simpson as guide this journey was a real tour de force, because to maintain the interest of an audience in such an esoteric subject for well over an hour is probably beyond anyone else. In the more recent versions of Genstat the syntax has been modernised and many possibilities are now available, such as substitution of identifiers at the time of compilation. Mr. Simpson gave various examples of what can be done, including one with the apt reference WEIRD which had eight consecutive + and - signs. The compiler can be used in more ways than originally intended, but it can be abused; however, to eliminate all possible abuses would be costly. Mr. Simpson described the hypothetical implementation of an imaginary 'HELP' directive, which would enable users to ask for help in various languages and could be particularly useful in the interactive version under development. In the discussion following this talk comments were made on the peculiarities of different machine ranges and the consequences for the implementation of Genstat.

It was then the turn of users to describe various statistical analyses and the part Genstat had played in them. M.I. Nimmo-Smith of Cambridge University described an application to psychology. The particular problem, which was not Dr. Nimmo-Smith's own project, was concerned with how well intermittent users of an interactive computing system understand the meaning of commands. Various multivariate techniques available in Genstat were used, to look at similarities both between users and between commands.

S. Evans of Bristol University described some morphometric studies on late Pleistocene molluscs: again, this was not Dr. Evans' own project, though she had been closely concerned with the data analysis, using Genstat throughout. Fossil shells were available from four distinct strata in the Turkana region of East Africa, and the problem was to sort out the evolutionary patterns. Much data manipulation had been necessary, and finally principal component and canonical analyses were performed, which showed the differences between strata clearly.

The final user was J.A. John of Southampton University, who was concerned with the inspection of residuals from the analysis of a designed experiment. Dr. John described two macros originating from CSIRO which he used in his analysis. One set of data gave many strange residuals, so a macro was used to fit the appropriate Box-Cox power transformation; he then used the other macro to display residuals and normal probability plots for the transformed data.

For a year in which spring was apparently always round the next corner the afternoon of April 10th appeared designed to show Cambridge at its best. Nevertheless, the originators of Genstat had arranged to be available to discuss problems with users, many of whom took their opportunity. Some surprising results were shown up, including analyses apparently correct on one machine range and incorrect on another, while some problems required all the combined expertise available for their solution.

The first session of the final day, on macros, was introduced by J.C. Gower of Rothamsted, who described how the use of macros allows the extension of the Genstat system, a feature not present in every language. The possible inefficiencies of macros are outweighed by the advantages of not cluttering up the system; however, documentation of library macros is difficult to keep up to date. P.W. Lane introduced library macros by showing the principles of their construction, and also gave some brief examples. Macros can be retrieved from the system and compiled either with the rest of the program or during its running.

More extensive examples of macros were given by two CSIRO speakers, R.I. Baxter and D. Ratcliff. Information about macros is available to CSIRO users in a computer file. Mr. Baxter described some macros developed at CSIRO, including those used by Dr. John the previous day. He also told us about the definitive text published by the Queensland University Press on "The Two Faces of Deviance", though deviance was not apparently used in its statistical sense. Mr. Ratcliff also related to a previous speaker by prefacing his contribution with the remark that he would be using Mr. Simpson's imaginary directive 'HELP' with the option 'LANG=STRINE'. He went on to describe both some of CSIRO's existing macros and some proposed new ones; these covered analysis of complex experimental designs and multivariate methods, among other features.

C.F. Banfield of Rothamsted concluded the macro session by considering present and future library macros for multivariate analysis. Again, Mr. Banfield's contribution related to that of a previous speaker, Mr. Ratcliff. Since multi-variate methods take a lot of space the only way to accommodate them in a system that is already large is by the use of macros. Fourteen of the twenty current library macros are multivariate, and other multivariate macros are in preparation.

The final contribution, by P. Phillips of the Agricultural Research Council Unit of Statistics (ARCUS), Edinburgh, described the difficulties of making Genstat accessible to inexperienced users, so was related to the discussion session on the first afternoon. Miss Phillips was mostly concerned with training for junior statistical staff, both formal teaching and practical sessions, starting with very simple programs and progressing to more complex ones. A Genstat clinic is also maintained by ARCUS to help anyone with Genstat problems, and this is used by staff from Edinburgh University as well as other research workers.

What then were the impressions of the conference as a whole? Possibly the most enlightening feature was that, at least among research workers in various fields, Genstat is no longer a novelty but a standard tool for use as and when it is needed. The conference certainly provided an opportunity for many concerned with this wide-ranging program to meet and find common ground. As can be seen from the cross-references between talks, the same features often recur in more than one context. Genstat has always been thought to be a program requiring a fair degree of statistical sophistication on the part of its users, so perhaps more than most it needs contact between the various sides. The mix of computer professionals, statisticians and other scientific research workers is not always achieved, but at Cambridge the blend was just right to ensure the success of the Conference.

Geoff Freeman
N.V.R.S.

COUNTING OVER TRIANGULAR ARRAYS

Given a list of variates $X(1...4)$, say, then to plot the six scatter diagrams requires the parallel lists

$X(1,1,1,2,2,3); X(2,3,4,3,4,4)$ to be generated.

As the number of variates increases this explicit listing of pairs becomes tedious and is totally unacceptable if one is writing a program with any pretensions to generality. Although the condensed notation for generating suffix lists in Genstat does not seem to lend itself to counting over triangular arrays, the following pseudo-Genstat code shows how it can be done:-

$X((2...n)\frac{n}{2}); X((1...n)\frac{n-2}{2}, (1)\frac{n}{2})$ n even

$X((1...n)\frac{n-1}{2}); X((2...n)\frac{n-1}{2}, (1)\frac{n-1}{2})$ n odd

When n is known in advance the appropriate form (n odd, or even) can be selected and values substituted for n, n/2, (n-1)/2 and (n-2)/2. The following diagrams show the order in which the pairs are generated for n=5 and n=6:

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|-------------|---|----|----|----|---|--|---|---|---|---|--|--|---|---|---|--|--|--|---|---|--|--|--|--|---|---|---|---|----|----|----|---|--|---|---|----|---|---|--|--|---|---|----|---|--|--|--|---|----|---|--|--|--|--|---|---|--|--|--|--|--|
| 1 2 3 4 5 | 1 2 3 4 5 6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table style="width: 100%; border-collapse: collapse;"> <tr><td style="width: 10%;">1</td><td style="width: 10%;">1</td><td style="width: 10%;">6</td><td style="width: 10%;">9</td><td style="width: 10%;">10</td></tr> <tr><td>2</td><td></td><td>2</td><td>7</td><td>5</td></tr> <tr><td>3</td><td></td><td></td><td>3</td><td>8</td></tr> <tr><td>4</td><td></td><td></td><td></td><td>4</td></tr> <tr><td>5</td><td></td><td></td><td></td><td></td></tr> </table> | 1 | 1 | 6 | 9 | 10 | 2 | | 2 | 7 | 5 | 3 | | | 3 | 8 | 4 | | | | 4 | 5 | | | | | <table style="width: 100%; border-collapse: collapse;"> <tr><td style="width: 10%;">1</td><td style="width: 10%;">1</td><td style="width: 10%;">7</td><td style="width: 10%;">13</td><td style="width: 10%;">14</td><td style="width: 10%;">15</td></tr> <tr><td>2</td><td></td><td>2</td><td>8</td><td>11</td><td>6</td></tr> <tr><td>3</td><td></td><td></td><td>3</td><td>9</td><td>12</td></tr> <tr><td>4</td><td></td><td></td><td></td><td>4</td><td>10</td></tr> <tr><td>5</td><td></td><td></td><td></td><td></td><td>5</td></tr> <tr><td>6</td><td></td><td></td><td></td><td></td><td></td></tr> </table> | 1 | 1 | 7 | 13 | 14 | 15 | 2 | | 2 | 8 | 11 | 6 | 3 | | | 3 | 9 | 12 | 4 | | | | 4 | 10 | 5 | | | | | 5 | 6 | | | | | |
| 1 | 1 | 6 | 9 | 10 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | | 2 | 7 | 5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | | | 3 | 8 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | | | | 4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 1 | 7 | 13 | 14 | 15 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | | 2 | 8 | 11 | 6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | | | 3 | 9 | 12 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | | | | 4 | 10 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | | | | | 5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

For a general program it is inconvenient to have to select between different codes for n odd or even. Both eventualities may be absorbed into the single sequence:

X((p...n)m); X((q...n)r,(1)m)

or equivalently: X(p...n); X((q...n)r,(1)m)

where p, q, r and m have been pre-set by the following code:-

| | values | |
|----------------|---------|---------|
| | n even | n odd |
| m = intpt(n/2) | n/2 | (n-1)/2 |
| r = n-m-1 | (n-2)/2 | (n-1)/2 |
| p = n-2r | 2 | 1 |
| q = 3-p | 1 | 2 |

Example:

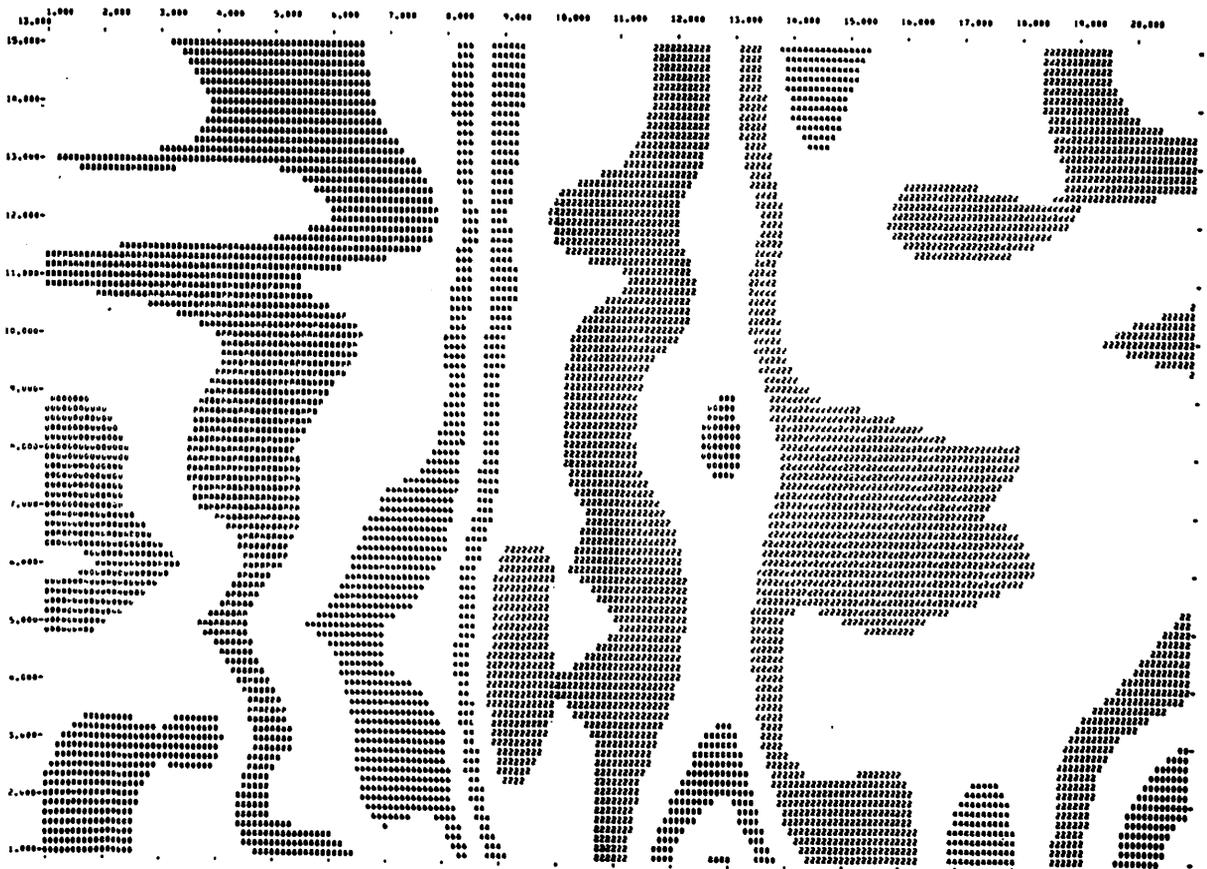
```
'REFE' SCATPLOT
'' To plot scatter diagrams for all comparisons of variates V(1...5)
'SCAL' N = 5 : M,P,Q,R
'CALC' N = INTPT(N/2) : R = N-M-1
      : P = N-2*R : Q = 3-P
'RUN'
'VARIATES' V(1...5) $ 40
'READ' V(1...5)
'FOR' Y = V(P...N) ; X = V((Q...N)R,(1)M)
'GRAPH' Y ; X
'REPE'
'RUN'
data
'CLOSE'
'STOP'
```


With very large sets of data it might be necessary to modify the above instructions, defining MULT1 and MULT2 as symmetrical matrices and working with their transposes. If weighted means are required this can be done by preserving the original data, and calculating an appropriate linear combination of the original data plus the smoothed data.

Richard Hardwick
N. V. R. S.

Editor's Note

A simple algorithm may be used to generate the values of MULT1 and MULT2. In the case of a 5 x 5 grid we have 3, 2, 1, (N-3)!(0), 4(1), NM4!(0), (5(1), NM4!(0)) NM4, 4(1), (N-3)!0, 1, 2, 3 where NM4 has previously been calculated as N-4.



PLANT GROWTH IN THE N.V.R.S. 'CREEMER' ROOM.
CONTOUR MAP OF PLANT WEIGHT.
(WEIGHTS SCALED TO MEAN : 5 S.D. : 3)

An example of smoothed data plotted by 'CONTOUR'

SIMPLE TRANSFER OF DATA STRUCTURES TO DISC AND MAGNETIC TAPE

We think that one reason why the directives for structure storage and retrieval are not used more frequently is that it is not realised how simple the basic statements are.

To store a set of data structures only two statements are required:-

```
'FILE' filename = logical unit number
'PUT'  filename; subfilename $ structure names
```

The first names the disc or magnetic tape file and associates it with a logical unit number; available numbers will vary from site to site. The second specifies a subfile of this file and states the structures to be included.

To retrieve these structures, two almost identical statements are required with the 'PUT' statement replaced by a 'GET' statement, i.e.

```
'FILE' filename = logical unit number
'GET'  filename; subfile name $ structure names
```

The process is made particularly simple because the user does not have to bother about giving any names of structures on which a structure to be saved may depend. Thus, if you want to save a table, you give the table name only. Genstat will automatically save the factors classifying this table, any structures holding level names and the unknown value cell; similarly it will produce them again when the table is later retrieved. Note that storage and retrieval is available not just for original data, but for any derived structures, such as SSP matrices. It can be used to store macros, or mixtures of data and macros to operate upon them.

Alan Todd
R. E. S.

INCONSISTENCY OR THINGS AIN'T WHAT THEY SEEM TO BE

The flexibility of the Genstat language is a pleasing feature but there are hidden dangers for the unwary.

Suppose that you declare

```
'NAME' LNF = A,B,C
'FACT' E $ 2 : F $ LNF
'TABL' T $ E,F
```

and create values for T. In a subsequent block of statements you might redeclare the name structure LNF, either accidentally or by design, by

```
'NAME' LNF = P,Q,R,S
```

and then

```
'PRINT' T
```

In the past this would have produced curious effects, and if T had been used in calculations the results were unpredictable.

The same problem arises when F is redeclared - as a factor of course - with a different number of levels. This may happen explicitly or implicitly (via GROUPS or HIERARCHY for example). Other types of structures - matrices and regression workspaces - can be affected in the same way.

The difficulties could have been avoided by careful programming but the danger of incorrect results being accepted as correct was thought to be too great and in Genstat 4.03 we try to trap this fault. The above redeclaration of LNF now provokes the comment

```
***** LNF
***** HAVING BEEN REDEFINED THE FOLLOWING
***** STRUCTURES WERE FOUND TO BE INCONSISTENT:
***** F      T
***** AND THEY HAVE BEEN DESTROYED
```

A comment VA 18 appears, but the job is not abandoned. If an inconsistent structure is referenced, the job will fail at that point.

Howard Simpson
R. E. S.

THE SAVING OF PUNCHING AND PROGRAMMING TIME IN THE ANALYSIS OF EXPERIMENTS

At Rothamsted much of our time from the late summer until the new year is spent in analysing data from field experiments carried out on our various farms. We try to spread the punching load throughout the year and to reduce the amount of punching and writing of Genstat programs by the following methods:-

1. Use of Backing Store and Macros

The plans and Genstat treatment and block structures of the field experiments are punched and tested during the slack period in early summer. The analysis of variance control structure (see ACON option of the ANOVA directive) is saved on backing store, and retrieved later when the data is submitted for analysis. When the number of terms in the analysis of variance model is large, the retrieving of ACON structures can result in considerable savings in computer time. The ACON for each experiment is kept in a subfile and the ACON subfiles for all experiments of a particular crop are kept in the same userfile. To save an ACON structure, the following Genstat statements are required:-

```
'SET'  EXNAM = a unique name given to the experiment
'ANOVA/ACON = EXNAM, ...
'FILE' DSET = logical unit number
'FETCH/LIST = EXCL' DSET $ EXNAM
'PUT/Y' EXNAM $ EXNAM
'SAVE/Y, LIST = ALL' DSET
```

When the first subfile is being saved, the FETCH statement must be removed.

The following statements are required to retrieve an ACON structure:-

```
'FILE' DSET = logical unit number
'FETCH' DSET $ EXNAM
'GET/LIST = ALL' EXNAM
```

We also save certain yield data in the same subfile as the ACON. The Genstat statements are the same as for saving the ACON structure except that the FETCH statement becomes:-

```
'FETCH/LIST = ALL' DSET
```

and adding at the end of the program:-

```
'CLEAR' EXNAM
'PUT/Y' EXNAM $ EXNAM, list of variates being stored
'SAVE/Y, LIST = ALL' DSET
```

Six macros are being used to analyse about 90% of our field experiments. These macros are saved in separate subfiles within the same userfile as the ACONs and the data. Thus if the fetch statement with the list = all option is used, the following statement will retrieve the macro:-

```
'GET' macro name $ macro name
```

In addition to the ACON structure, extra structures are saved on backing store, the main one being a pointer structure called PLIST which contains a list of factors used for printing purposes. Inside the macros the following statements appear:-

```
'SCAL' NL
'START'
'CALC' NL = NVAL (PLIST)
'SET/POIN = S' TLIST = PLIST
'RUN'
```

and later on a print statement of the form

```
'PRIN/P, VAR = 1' PLOT, TLIST, ... $ 10, (10) NL, ...
```

Within the macro, default values for the PRINT and LIMA options of the ANOVA directive are set by statements of the form

```
'SET' SETPR = 13
'ANOV/PR = SETPR ...
```

Since all the macros are used at compile time these defaults can be overwritten by adding to the block of statements containing the USE statement, either before it

```
'CALC' SETPR = new value
```

or after it

```
'SET' SETPR = new value
```

The details of the macros are omitted as they are unlikely to be of general use. The main point is that if you are to carry out large numbers of similar analyses, you should consider writing your own macros.

After macros have been written, skeleton Genstat programs can be printed leaving spaces for the variable parts of the program. This reduces programming time considerably, an example of a skeleton program follows:-

```
'REFE/NUNN=150,NID=150,LINE=6000' DP1

'UNIT' PLOT$[      ]=[      ]
'SET' EXNAM=[      ]([      ])

'FILE' DSET=48
'FETCH/LIST=ALL' DSET
'GET/LIST=ALL' EXNAM:GRASS(2)

'SCAL' NCUT=[      ]
      :TARE=[      ]
      :AH,CF
      :TODW=[      ]

'HEAD' HDATE=''([      ]/[      ]/[      ])''

'CALC/M' AH=[      ]/10000
      :CF=[      ]/AH
'RUN'

'READ/P' FW(NCUT)$F,[      ]
      :ODW(NCUT)$F,[      ]
'RUN'

DATA HERE

'USE' GRASS(2)$
'RUN'
'CLOSE'
```

2. Experiments with subplots or complicated unit numbers

Unit numbers of experiments at Rothamsted with subplots, indicate with the first few digits the whole plot to which the unit belongs and with the last digit the subplot. The punching of sequences of this form can be tedious. In the following example of sixteen whole plots split into five subplots, this can be avoided by replacing:-

```
'UNIT' PLOT $ 80 = 11...15, 21...25, 31...35 etc. to 161...165
```

by

```
'VARI' A = 5(1...16)
      : B = (1...5)16
'CALC' A = A * 10 + B
'RUN'
'UNIT' PLOT = A
```

Suppose the whole plot factor in the above example is TRT with four levels A, B, C and D. It is not necessary to punch all 80 values of TRT, instead one needs only to punch sixteen values, one for each whole plot. The following statements illustrate how this can be done:-

```
'NAMES' NT = A, B, C, D
'FACT' TRT $ NT, 16
'READ/S' TRT
'RUN'
B C D A etc.
'EOD'
'FACT' TRT $ NT, 80 = 5(TRT)
```

Alan Todd
R. E. S.

SOME NOTES ON RESTRICT

Some of these features are only available in version 4.03 (et seq). These are marked (4.03) on the left.

Throughout these notes we will assume that the following declarations have been made:-

```
'UNIT' U $ 12 = 1...12
'VARI' VA $ 1.1...12.1 : VB $ 1.2...12.2
'FACT' FA $ 3 = 4 (1,2,3) : FB $ 4 = (1...4)3
      : FC $ 6 = 1, 3, 2, 4, *, 6, 5, *, 1, 4, 2, *
'INTE' IA = 1, 3 : IB = 2, 4 : IC = 2, 4, 6, *
      : ID = 1, 3, 5, 6 : I = 1, 3, 7, 11 : RA, RB
```

The simplest use of RESTRICT would be

(4.03) 'RESTRICT' VA ; I
which restricts VA to units 1, 3, 7 and 11.

(4.03) 'RESTRICT/C' VA ; I
restricts VA to the complement of I; that is to units 2, 4, 5, 6, 8, 10 and 12. If this complementary set is needed elsewhere it can be named thus:

(4.03) 'RESTRICT/C' VA ; I, RA

If the same restriction is required for VB it can be included in the same instruction, or restricted later to be the same as VA (this is a lot quicker than doing the full RESTRICT again). This is done by writing

```
'RESTRICT' VB $ VA
```

If the restrictions are no longer required they can be removed by

```
'RESTRICT' VA, VB
```

The foregoing are all "special" uses of RESTRICT. We now come to the general form.

'RESTRICT' VA \$ FA = IA

will restrict VA to those units corresponding to levels 1 and 3 of FA. If only one level of the factor (in this case FA) is required it need not be put into an integer vector, but can be used directly:

'RESTRICT' VB \$ FB = 3

We are not limited to restricting on one factor at a time.

'RESTRICT' VA \$ FA, FB = IA, IB

restricts VA to those units common to the subsets of FA and FB whose levels are defined in IA and IB respectively. We can extend a previous restriction.

'RESTRICT' VA, VB \$ VA, FC = 0, 4

will take the restriction already applied to VA, and units corresponding to level 4 of FC and make the units common to both into a new restriction applied to VA and VB. The serial numbers of these units can be stored in a named integer vector by rewriting the above instruction.

'RESTRICT' VA, VB \$ VA, FC = 0, 4 ; RB

RB could then be kept to restore this restriction later by using the first form of RESTRICT mentioned above.

The complementary subset can also be obtained in the general form.

'RESTRICT/C' VA \$ FA = IA

restricts VA to exclude units corresponding to levels 1 and 3 of FA, i.e. to include level 2.

'RESTRICT/C' VA \$ FA, FB = IA, IB

excludes units referred to by FA = IA or FB = IB.

Wherever a value of a factor used in a RESTRICT is missing that unit is normally excluded. However, it can be treated as an extra level of the factor by setting another option MV = IN (default MV = OUT).

Thus

'RESTRICT/MV = IN' VA \$ FC = IC

will include units 3, 4, 5, 6, 8, 10, 11, 12. FC = IC without MV = IN would be faulted. In the complement case

'RESTRICT/C, MV = IN' VB \$ FC = ID

will include units 3, 4, 5, 8, 10, 11, 12.

The vector being restricted need not be a variate.

'RESTRICT' FA \$ FA = IA : FB \$ FB = IB

will restrict FA and FB as you would expect.

(4.03) 'RESTRICT' VA \$ FA, FB = 0, 0

will restrict VA to units common to both restrictions.

There are many more ways in which RESTRICT can be used, probably including a lot that I haven't thought of. However, these notes should provide a guide to the most common uses, and perhaps a basis for further exploration.

Penny Leech
R. E. S.

SUGGESTED CHANGES TO BACKING STORE

When the userfile is specified in the GET or PUT directive all subfiles are transferred including the subfile specified, e.g.

'GET' UF, SF \$ list

is equivalent to

'FETCH/LIST = ALL' UF

'GET' SF \$ list

which will result in all subfiles in userfile UF being transferred to the workfile and the structures specified in the list are copied from subfile SF into core.

There are two alternative proposals for changing the program.

(1) Only the subfile specified in the GET or PUT statement is transferred to the workfile or to the userfile.

Example:

'GET' UF, SF \$ list

will become equivalent to

'FETCH' UF \$ SF

'GET' SF \$ list

(2) No subfiles are transferred. The GET statement will read the contents of subfile SF direct from the userfile UF without copying to the workfile. The PUT statement will result in a userfile UF being created containing only one subfile SF. Note: subfile SF will never appear in the workfile.

Either of the above alterations will make it more efficient to access a single macro in the macro library.

Could anyone who has strong feelings about these proposals for release 4.04 please contact:

Alan Todd
R. E. S.