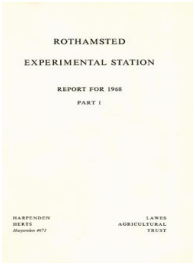
ROTHAMSTED RESEARCH

Report for 1968 - Part 1

Full Table of Content



Statistics Department

J. A. Nelder

J. A. Nelder (1969) *Statistics Department ;* Report For 1968 - Part 1, pp 233 - 239 **- DOI: https://doi.org/10.23637/ERADOC-1-123**

STATISTICS DEPARTMENT

J. A. NELDER

Statistical programming

Although no longer concerned with the operation of the computer, the department retains responsibility for maintaining and developing statistical programs.

Orion programs. Numerous extensions and modifications were made to our standard Orion programs. A new version of the General Survey Program is now complete. It is a great improvement on the previous program, with many additional features, including simpler and more compact specification of the tables required and the operations to be performed on them, excellent facilities for printing the results with adequate headings, etc., and better provision for storing tables on magnetic tape. (Anderson, with F. Yates, Computer Department) Simpson wrote a general input program for survey data, to be used in association with the main program. Anderson incorporated programs for principal components analysis and canonical correlation into his system MAP, and added plotting facilities to the regression program. Major improvements were made to the polynomial regression program. (Ross and Lauckner) A simplified version of the General Factorial Program was written to deal with the simpler, commonly occurring factorial designs. (Preece) A few errors in the program for summarising groups of experiments were corrected, and this is now working well. (Alvey) Two programs were developed to present national data graphically: both draw maps of Britain on the line printer, one on a county basis, and the other using the 10 km squares of the National Grid. In either, the user can specify the symbols to be printed in each subdivision. (Ross and Lauckner)

Statistical algorithms. Numerical analysts have long realised the importance of publishing thoroughly tested computing recipes or algorithms, and several scientific journals have sections for this purpose. However, statistical computing procedures are not yet adequately covered, and the journal *Applied Statistics* has recently begun an algorithm section. Several members of the department have already contributed. (12.5, 12.13, 12.15) (Gower, Herraman, Nelder and Ross)

Programs for the new computer. The projected arrival of the new ICL 4/70 computer in 1969 will inevitably entail a major upheaval, and programming for the change is well under way.

Differences between the Orion and the 4/70 are considerable: first, the 4/70, as befits a third-generation machine, will have a more advanced operating system than the Orion; secondly, it will offer much greater storage space (on discs) with relatively fast access time; thirdly, it will allow the use of more powerful general-purpose programming languages than EMA, namely Fortran IV and Algol; and fourthly, there will be some

https://doi.org/10.23637/ERADOC-1-123

form of direct-access to the machine. Future programs must take account of these new aspects, and exploit them to the maximum for the benefit of the user. The key word here is 'modularity'; effective use of the new computer will depend on the user being able to call on a system of well-designed program modules, each able to do an operation in a self-contained way, and to be linked together with other modules to form many different patterns. With such a system the user should be able to construct rapidly the special pattern of modules he needs for his own problem.

The GENSTAT system. The system under construction (for GENeral STATistical work) is based on the idea that the easiest way to ensure compatibility of different program modules is to define the basic data structures on which they operate. Some of these structures can be seen in raw experimental data, whereas others are created during analysis. At any stage, the analysing program will have assembled a set of these structures in the core-store of the machine. GENSTAT will offer the user, for example, the ability to store on magnetic tape, as a self-contained file, any subset of the current structures, and to fetch them again when required. Also, each standard structure will have a standard program for printing it, so that the user is largely freed from the tedious business of programming his output in detail. GENSTAT is being written in Fortran, and work to date has been chiefly concerned with the 'housekeeping' aspects of the system, which manipulate the structures in the core-store, and move them to and from magnetic tapes and the outside world. The definitions of the structures and working of the housekeeping programs are being fully documented, and will be provided to outside programmers wishing to add facilities to the system. Programs now coded and undergoing tests include some for accessing structures in core (Alvey), for manipulating multi-way tables (the SCAN facility) (Gower and Krzanowski (12·5)), for storage on magnetic tape (Nelder) and a general input package (Simpson). The formulation of flexible input/output facilities may well prove to be of wider use to Fortran programmers, who often find the standard procedures restricting. Gower drew up a specification for a package of programs for multivariate analysis, in collaboration with interested people at East Malling and elsewhere, for future inclusion in the system. Ross is considering how the facilities provided by the Orion classification program can best be incorporated into the new system.

Theory. Succinct and logical ways of describing the structure of experimental data, as distinct from their content, are needed, both outside and inside the computer. Formulae for structure description are being developed (Nelder) that can be used either externally or internally. The internal description of structures relies greatly on the use of pointers as connecting links between the structural elements, and Gower is studying the implications of this in relation to existing and proposed programming languages. The current programming languages, Fortran and Algol, have acknowledged defects for statistical work, and for the future the department is interested in BCL, a language being developed by a team under Mr. David Hendry of Radics Ltd., an independent software house closely connected

https://doi.org/10.23637/ERADOC-1-123

with the Institute of Computer Science. Programs written in BCL look in many ways quite different from, say, the Fortran programs we write today. BCL gives the programmer the facility to describe complex structures in his data and to map them into structures inside the computer, and *vice versa*. The hitherto rigid separation in programs between *definitions*, describing what entities the program operates on, and *commands*, describing what the program does with those entities, now disappears. Because so much of statistical programming is concerned with the problem of steering large amounts of data in and out of the computer, languages such as BCL have great potential importance.

Statistical theory

Gower and Ross developed the relationship between the method of cluster analysis known as single-linkage and the concept of a minimum spanning tree. (12.7) The existence of compact algorithms for determining the minimum spanning tree (12.13) allows single-linkage cluster analysis to be applied to large sets of data, where other forms of cluster analysis would be impracticable. Gower discussed the interrelationship of several commonly-used methods of cluster and ordination analysis (12.6). He (12.4) also showed how to add a new point to any ordination analysis, given the distance between the new point and all the old ones. From work on the analogous problem for cluster analysis, Ross developed some empirical techniques that work well in practice (12.14). In addition to their theoretical interest, the results are important in large-scale analysis when exact methods are impracticable. The exact analysis can be done on a selected sub-set of the data and the new methods used to add in the remaining data.

Ross continued his work on the numerical estimation of parameters by maximum likelihood. The problem of devising a suitable program for a computer is analogous to that of providing a blind man with a strategy for finding his way to the highest ground of the landscape surrounding him. For the estimation problem the landscape may be a surface not of two dimensions, but of five or more. Some statistical problems generate surfaces with long attenuated ridges, whose highest point is difficult to find rapidly; special strategies, adapted to the problem in hand, can greatly improve searching efficiency.

Preece investigated modifications of the standard iterative method of fitting missing values in experimental data, using the General Factorial Program. The rate at which the final values converge can often be improved considerably by subtracting at each stage, not the current residual, but a multiple of it. Results were obtained on the best choice of the multiplier in certain situations.

Anderson continued his work on methods of ordination and clustering, considering in particular the use of least-squares in one- and two-dimensional ordinations. Krzanowski looked at the constancy of variance produced by some angular transformations used in taxonomic work.

https://doi.org/10.23637/ERADOC-1-123

Practical applications

Established statistical techniques were applied to a great range of data from Rothamsted departments and other organisations. There was a continuing demand for the various multivariate techniques for which there are programs on the Orion. Classification methods were applied to organisms as diverse as field beans, marine animals (*Euaugaptilus*), wild plants (*Crotalaria*), bacteria (*Klelosiciella Achromobacter* and *Staphylococci*) and soil fauna. (Gower and Ross) Canonical analysis was used on data from the wild plant *Erophila* and other ecological data (Anderson and Lowe), from plant genetical studies, and from ancient British skulls. (Gower) Principal co-ordinate analysis was found useful in a study of tobacco species (Gower), and discriminant analysis in determining the sex of fallow deer from some of their bone measurements. (Anderson and Lowe)

Regression analysis (including probit analysis) was applied to problems relating the shrinkage of clay soils to various other properties, the burial of litter by earthworms to weather (Anderson), and Borstal admissions to social factors. (Ross) Anderson made a study of time trends in the treatment responses on the Broadbalk wheat experiment. An attempt to use a combination of regression analysis and clustering techniques to investigate relationships between plant growth and soil characteristics did not produce biologically meaningful results, and ended.

An important part of the department's resources is devoted to providing a Research Statistical Service for other institutes and for the National Agricultural Advisory Service. The work embraces advice on statistical problems, analysis of data and interpretation of results. The main areas covered are farm surveys on crop husbandry and factors affecting disease and mortality in livestock, and investigational work on farms, especially experiments with crops and livestock.

Surveys. Despite the considerable amount of survey work done in past years, there is no current information on farmers' use of fertilisers in many parts of England and Wales; this not only handicaps the giving of advice on fertiliser use, but also means that estimates of trends in fertiliser use for the country as a whole can be queried in detail, because districts surveyed in different years are not perfectly comparable, so there is doubt about the relative weights to be given to estimates from individual districts.

A new scheme was therefore prepared for the Survey of Fertiliser practice, whereby a representative sample of farms in England and Wales will be visited every year, beginning in 1969. This scheme, which has been agreed with National Agricultural Advisory Service Regional Soil Scientists and with the Fertiliser Manufacturers' Association, should provide estimates of average practice on individual crops annually for major regional or farm-type groupings, and on a three-yearly basis for counties or other small groupings of interest. The continuous sampling should also make trends more evident, and will have the practical advantage of avoiding peaks both in the field work and in analysing the data. (Church)

https://doi.org/10.23637/ERADOC-1-123

Results of the 1966 surveys, on the use of fertilisers for cereals and horticultural crops, and on the types of fertilisers used for different crops, were published. (12.2, 12.3, 12.8.) (Church and Hills)

Analysis of the Potato Marketing Board survey on the quality of stored maincrop potatoes was completed. In both 1965–66 and 1966–67 about 40% of stored potatoes in Great Britain conformed to a premium 'Grade I'. A quarter of the stored crop failed to reach this grade because of surface blemishes from mechanical damage. Deeper damage, surface blemishes from diseases (mostly scab, silver scurf and blight), and greening each caused rejection of 13–18% of the crop; pest damage and rots each accounted for a further 5%. (Church)

Other survey work done during the year included collaboration with N.A.A.S. in the planning and analysis of a survey of work records (N.A.A.S. Eastern Region), analysis of a survey of plant nutrients in glasshouse soils and preliminary examination of results from a survey of lime requirements. Collaboration continued with the Ministry of Agriculture, Fisheries and Food Plant Pathology Laboratory on a series of pesticide surveys, and on surveys of foliar diseases of barley in 1967 and 1968. Surveys of the 1968 main and early potato crop were planned with the Potato Marketing Board and the National Institute of Agricultural Engineering. (Church and Hills)

A farming-type map of England and Wales based on agricultural census data, described in the 1967 report, was issued as a wall-map; copies can be obtained from the Kynoch Press, Birmingham.

A final report of the two surveys on the disposal of sheep-dip residues was prepared and work on the Guernsey Progeny Testing Scheme continues. Analysis began of the data from a survey of the diseases of pedigree beef cattle, sponsored by the Animal Health Trust and the National Cattle Breeders' Association. (Leech) An investigation of the wide variation in the proportion of cracked eggs from different farms showed that this could not be attributed to handling factors or to variation in shell strength, the reasons usually given (12.9). (Leech)

Experiments

Routine analysis. The number of replicated experiments analysed increased sharply, from 3594 in 1967 to 5200 in 1968, but the average number of variates per experiment analysed decreased slightly from 6·4 to 5·9. Analysis of replicated experiments accounted for about 70% of the work.

The number of jobs increased from 1582 in 1967 to 1834 in 1968, and of these, a little over a third came from the N.A.A.S. The smaller percentage increase in jobs (one of which may comprise several experiments) compared with that for replicated experiments shows that customers are submitting their results in a more organised way. Beginning in September, the size of a job has been estimated from the number of data to be punched. Although this criterion ignores inevitable differences in the amount of instructions to be punched between one job and another, it gives a good general idea of the distribution of job sizes. Figures for the last four months give an

https://doi.org/10.23637/ERADOC-1-123

average job size of about 540 data, and an estimate for the whole year of slightly fewer than a million data handled. (Dunwoody, Hill, Ryan and Watson)

Crop experiments. For many years the department has had close ties with soil scientists in other research stations and in the N.A.A.S. This collaboration has borne fruit in many joint publications in soil science and plant nutrition, aimed at improving recommendations on the fertiliser requirements of farm crops. The contributions of the department come at the planning stage, in processing results of the individual trials, and especially in summarising and helping to interpret the results of completed investigations. As much experimental work is being done to assess the nitrogen requirements of crops, and to find how and why these requirements differ from place to place and year to year, the department has investigated the form of the nitrogen response curve for sugar beet and cereals. The results suggest that, within the usual range of dressings, the relationship of sugar yield to nitrogen given can often be adequately represented by two straight lines; yield usually increases sharply to an optimum, beyond which it changes little or decreases slightly. It has so far proved impossible to predict the nitrogen requirements of either crop from a knowledge of site and soil characteristics.

Other important aspects of our work included further investigations for the N.A.A.S. soil scientists of the fertiliser responses of the potato crop, and how these differ according to soil, season and husbandry factors; also examination of the results of residual phosphate experiments on N.A.A.S. Experimental Husbandry Farms, and investigations comparing the value of methods of estimating 'available' soil P. (Boyd and Hill)

Livestock experiments. Work continued on the design and analysis of animal experiments. (Lessells and Watson) Lessells co-operated with members of the N.A.A.S. in a review of experimental techniques for beef and dairy cattle. Together with N.A.A.S. nutrition chemists, he was concerned with testing the relative efficiencies of competing systems for estimating the energy requirements of ruminants, and with developing indirect methods of assessing the digestibilities of hay and silage, on which an interim report was produced.

Commonwealth and overseas

Analyses were again provided for a wide range of crops. An examination of the results of four seasons' fertiliser trials and demonstrations on paddy rice in the Dry Zone of Ceylon showed the importance of fertilisers in diminishing the need for imported rice. (Boyd and Hill, in collaboration with Mr. D. H. Constable and Mr. F. K. A. Weerawickrema F.A.O. Ceylon Fertiliser Project) Differences in fertiliser effects between districts and from site to site within a district were evaluated; plot errors were estimated from interactions of nutrients in individual trials. Experiments were also analysed for Nigeria (cocoa and cotton), Ghana (cocoa), Tanzania (cotton) and the West Indies (maize and yams). (Walker, Ryan and Dr. B. Springer)

https://doi.org/10.23637/ERADOC-1-123

Advice and assistance were given on relating yields of Kenaf and Roselle to latitude (Nigeria), on growth studies in Sorghum (Nigeria) and on problems of sugar-cane breeding (Mauritius). Classification methods were applied to soils and trees in the Solomon Islands, and to vegetation in Northern Nigeria.

Staff and visiting workers

F. Yates retired in March after thirty-five years as Head of the Department. When he succeeded R. A. Fisher in 1933 the Department consisted only of himself and four assistants. Another statistician, W. G. Cochran, was appointed in 1934. Under Yates' leadership the department grew to become one of the world's leading statistical research groups. He greatly extended the work initiated by Fisher on the design and analysis of experiments, and the growth in the theory and practice of surveys is manifested in Yates' own book on the subject. The department pioneered the use of electronic computers in statistics, and has now accumulated an unrivalled knowledge of the problems of statistical programming. We are fortunate in that his work for the Computer Department allows us the continuing benefit of his experience.

W. Krzanowski was appointed. Dr. T. H. Anstey, Director of Lethbridge Research Station, Alberta, spent four months with us. At the request of the Ministry of Overseas Development, P. Walker spent five weeks in West Africa, visiting the Institute of Agricultural Research, Samaru, the Ministry of Agriculture, Ibadan, Nigeria, the Cocoa Research Institute, Tafo, and the Department of Agriculture, Gambia.

Preece continued as the United Kingdom and Ireland's Regional Editor of *Statistical Theory and Method Abstracts*, published by the International Statistical Institute. He dealt with 386 abstracts from 26 British journals (95 more than in 1967). Also for *Statistical Theory and Method Abstracts*, he compiled lists of almost all published algorithms of statistical interest.

Seven temporary workers, five of them from overseas, spent various periods in the Department.

https://doi.org/10.23637/ERADOC-1-123 pp 8