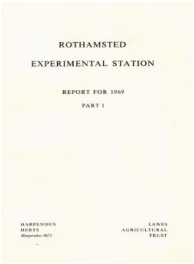
ROTHAMSTED RESEARCH

Report for 1969 - Part 1

Full Table of Content



Statistics Department

J. A. Nelder

J. A. Nelder (1970) *Statistics Department ;* Report For 1969 - Part 1, pp 264 - 271 **- DOI: https://doi.org/10.23637/ERADOC-1-124**

STATISTICS DEPARTMENT

J. A. NELDER

Statistical programming

The department has a large and continuing task in providing and maintaining computer programs for statistical analysis. The impending installation of the new computer caused us to concentrate on programming in Fortran. However, some additions were made to existing Orion programs; in particular the Maximum Likelihood Program was extended to deal with problems of genetic linkage estimation, graphical facilities for principal-coordinate analysis were improved, and the General Lattice Program extended. (Ross and Lauckner) Other existing programs were amended to remove faults discovered during routine use, and an account of the Multivariate Analysis Program was published. (12.1)

The GENSTAT system. The main effort was in developing the GENSTAT system of statistical programs outlined in the 1968 Report, and substantial progress was made. A teleprinter terminal to a remote computer was installed during the year, and extensively used. The computer provides editing facilities, so that programs can be modified and re-run immediately. The terminal proved very effective for debugging basic sub-routines and about 8000 lines of Fortran program were tested, a figure that could not have been achieved without its use. The facilities to be provided in the first release of GENSTAT in 1970, and the progress made in implementing them, are as follows:

Data description. Basic data structures have been defined for the *data matrix* (to hold the original data from surveys and experiments), three types of *matrix* for matrix arithmetic, *latent roots and vectors* (fundamental to multivariate analysis), *multiway tables* (important in, for example, survey analysis), and plain-language *text*; the user can define *sets* of items in a general way, and also *constants* (scalars) such as conversion factors. All GENSTAT programs use these standard structures as input and produce others of the same kind as output; their use thus gives the automatic compatibility between programs that is so important if the system is to be easy to use. Methods of accessing the structures are standard and the underlying programs are tested. (Alvey)

Input–output. To facilitate programming at a higher level, a basic input–output package was programmed that is more flexible than standard Fortran I/O. (Simpson)

It allows variable format of input items, and helps the input of material containing names, and symbols other than those found in numbers. Using it, general programs are being tested for reading and printing the GENSTAT standard structures. The reading program allows data to be specified in various ways, accepts missing value symbols and various conventions for blank fields. The program is coded and under test. (Simpson)

https://doi.org/10.23637/ERADOC-1-124

The printing program contains a general rectangular table-printing routine, allowing parallel tables of variates, automatic page overflow for large tables, optional suppression of labels, etc. Similar facilities exist for triangular structures. The program is tested. (Nelder)

The ability to store data between one run and the next is vital in any multi-stage analysis. Complex problems of specification arise when named structures previously stored are merged with new ones existing in core, or conversely when current structures are added to an existing file of data. Very powerful facilities for these operations were specified, and the programs are coded and under test. (Rogers)

Derived variates. In the system variates can be transformed and combined in various ways. New variates can be defined by general arithmetic expressions, including various standard mathematical functions, such as square root and logarithm, and standard matrix operations, like inversion. Multiway tables can be combined in a quite general way, extending the provisions for table manipulation in the General Survey Program to tables with arbitrary classification sets. The programs are coded and partly tested. (Alvey, Krzanowski, Nelder)

Regression analysis. The facilities in Anderson's Orion multivariate analysis program (12.1) were transferred, with an important generalisation. Arbitrary sets of constants can now be fitted by including qualitative independent variables. This allows constant-fitting to be subsumed under regression, and permits general linear models to be easily specified. The programs are coded and partly tested. (Anderson, Lowe, and Wedderburn)

Analysis of variance. A very general algorithm for the analysis of variance developed by G. N. Wilkinson was programmed for the system. It produces the analysis of variance, with associated tables of effects for the class of generally balanced designs. These include all standard orthogonal factorial designs in randomised blocks, Latin squares, split-plots (to any depth) etc., in addition to the balanced incomplete block designs, group-divisible designs, and many of the lattice designs. Unequal numbers of treatments are allowed when balance is preserved, also nested treatment structures. Balanced confounding and fractional replication are catered for. The basic algorithm is tested, the remaining facilities (output, etc.) are under construction. (Wilkinson and Rogers)

Multivariate analysis. Facilities to be provided include principal component and principal coordinate analysis, canonical variate and canonical correlation analysis, and the matching of one set of coordinates to another by rotation of axes. The basic algorithm for the extraction of latent roots and vectors is tested, and all directives are fully specified (Gower and Krzanowski)

Cluster analysis. The system will have most of the facilities in the Orion CLASP program, e.g. the evaluation of similarity matrices using

https://doi.org/10.23637/ERADOC-1-124

various similarity coefficients, the construction of minimum spanning trees and single-linkage cluster analysis, and various techniques for subsequent grouping and sorting of the data. The programs are coded and partially tested. (Ross and Lauckner)

User's language. A user's language for specifying standard operations was developed, with the syntax made as simple as possible; each statement starts with a directive word, such as 'READ'. The user is offered various abbreviations that allow him to express patterned sets of both data and instructions compactly. A compiler translates the directives into an internal code, which is then referenced by the interpretive programs that carry out the directives. The forms of all directives (71 in number) are fully specified, the underlying routines for the compiler are tested, and the compiler itself is under construction. (Simpson)

Documentation. Three levels of documentation are being produced. At the bottom level, all Fortran subroutines are described according to the conventions set out for the algorithm section of *Applied Statistics*. Next, the programs underlying individual directives are described together with the internal representations of the standard structures. Finally a user's guide is in preparation, describing the facilities available with each directive, together with a general introduction to the user's language. The three levels of documentation would be most relevant, respectively, to a programmer contributing to the system, the sophisticated user wanting detailed information on the workings of the system, and the general user.

Other programming. Kruskal's program for his method of multidimensional scaling was adapted to our dialect of Fortran and provided with the necessary input facilities; algorithms for hierarchical agglomerative cluster analysis with printing of the resulting dendograms were programmed in Fortran. (Anderson and Lowe)

Gower collaborated with Professor Reyment of the University of Uppsala on a general program for principal coordinate analysis. Ross translated substantial parts of his Orion program for maximum likelihood estimation into Fortran.

Theory. Gower and Nelder described some ways in which data structures relevant to statistical computing can be defined, and the extensions required to existing general-purpose computing languages if such structure definition is to be made simple for the programmer (12.11, 12.17). Work is continuing on the definition to the computer of the external structure of data (Nelder) and their internal representation (Gower), and the extent to which the special needs of statisticians are met by existing computer languages. Gower and Preece investigated a combinatorial problem suggested by an algorithm of Nelder (12.18) and extended his results (12.13).

Statistical algorithms. Algorithms accepted for publication include the evaluation of marginal means (12.8), the analysis of variance for a factorial table (12.9), the calculation of effects (12.10) (Gower), and the re-ordering

https://doi.org/10.23637/ERADOC-1-124

of multiway structures (12.14) (Krzanowski). All these are in Algol. One in Fortran deals with the efficient formation of sums of squares and products when only a few of the variates involved can be held in the core store at any one time. (12.18) (Nelder)

Statistical theory

Ross extended his work on numerical methods of maximum likelihood estimation to include the fitting of the double exponential curve, the estimation of genetic linkage, and reaction rates in the equations of enzyme kinetics. He also investigated sampling problems where the populations concerned are aggregated or clumped.

Preece discovered near-cyclic representations of the sets of interactions defining some recently-discovered fractional factorial designs. (12.20)

Krzanowski investigated the possibilities of using Kruskal's measure of monotone stress to choose a time scale in growth analysis. The problem arises when crop growth is measured at intervals and various possible relevant meteorological data are collected simultaneously. It is required to assess which of several possible alternatives to time (such as day degrees) is the best, without assuming that the form of the growth curve is known, except in so far as size increases with time. Kruskal's technique allows the goodness of a time-scale to be measured when there are measurements of growth from several sowing dates, and the results are being compared with methods which assume a particular form of equation for growth.

Anderson continued work on the approximate representation of a sample of n-variate points in n dimensions by points in a few, say two, dimensions. In particular the properties of a representation defined by minimising the sum of squared deviations of reduced configuration distances and the true distances are being examined. He also continued work on the theory of divisive hierarchical clustering. Where apparent clusters are formed in what is really a sample from a single distribution, the average value of the maximum sum of squares between clusters can easily be calculated for univariate distributions, but the multivariate case is more complex.

Wedderburn attempted to define the class of estimation problems reducible computationally to iterative weighted linear regression. Probit analysis is such a problem; here the weights depend on the fitted values, and so change in each cycle of iteration. The effective independent variate also changes. It seems that many other problems, though non-linear, can be similarly treated. The solution of such problems can be easily programmed, given a good basic set of regression subroutines.

Practical applications

The department continued to be consulted on a wide range of biometrical problems by other Rothamsted departments, other Institutes financed by the A.R.C., the National Agricultural Advisory Service, the Plant Pathology Laboratory of the Ministry of Agriculture, Fisheries and Food, and

https://doi.org/10.23637/ERADOC-1-124

other organisations concerned with biological problems. Again there was a considerable demand for assistance on classification problems using numerical techniques. Examples include the effect of long-lying snow patches on high-altitude plant communities (Anderson), the classification of earthworms and fishes (Gower), the taxonomic status of British water voles (12.6) (Krzanowski), and classification problems associated with influenza virus, dyslexic children, dental organisms and West African cowpeas. (Ross)

Principal coordinate analysis was applied to data from 612 soil profiles obtained by the Soil Survey of England and Wales (Gower and Krzanowski), and to data on the New World plant genus *Oplonia* and various groups of bacteria. (Ross and Lauckner)

Leech investigated the value of classification techniques in the study of necropsy records, using data from the calf survey, but found them mainly unhelpful.

Work on the construction of a key for about 350 species of yeasts is in progress; problems arise when some test values are unknown for certain species. (Gower and Lowe)

Wedderburn analysed results from the Soil Microbiology Department on the selection of strains of Red Clover to improve the fixation of nitrogen by nodule-forming bacteria, also the effect of radiation on another clover species.

The process of establishing the innocuity of a batch of vaccine can be tedious and expensive, and no certainty is possible, only statements of probability. The statistical aspects of innocuity testing of foot-and-mouth disease vaccines using two techniques were evaluated. (12.2) (Leech)

Ross investigated the distribution of plants in a row-sown crop when an intended regular spacing is upset by the combined effects of germination failure and occasional multiple seeds at a site. Simulation of a simple model on the computer gave results compatible with field data.

Surveys. This is the first year of the new scheme for the Survey of Fertiliser Practice (see *Rothamsted Report for 1968*). The field work for the survey was almost complete by the end of the year. Using analyses already developed with the General Survey Program (GSP2) the main tables giving results for England and Wales can be produced within a week or two of receiving the final returns. (Church and Hills)

Surveys of maincrop (997 farms) and early potatoes (374 farms) done by the Potato Marketing Board in collaboration with this department and the National Institute of Agricultural Engineering were analysed, and a report prepared on the early potato survey. This survey, the first of its kind, covered seven districts specialising in early potatoes, and provided information on seed preparation, cultivation and planting, on the use of fertilisers, pesticides and herbicides, and on harvesting methods. An estimated 40% of the acreage was planted with certified seed and most of the rest with home-produced once-grown seed. In the south-west and Wales the average amount of seed planted was markedly greater than for the main crop. Eighty per cent of the crop was machine-planted and there were striking effects of herbicide use on cultivation methods. Two-thirds

https://doi.org/10.23637/ERADOC-1-124

of the crop had some herbicide treatment, and cultivation after planting was needed on 70% of the acreage without herbicide but only on 30% with. (Church and Hills)

Summary tabulations were again prepared for the Survey of Foliar Diseases in barley organised by the Plant Pathology Laboratory of the Ministry of Agriculture, Fisheries and Food. Samples of 50 tillers from each of 300 crops were assessed for *Rhyncosporium*, Mildew, Brown Rust and other diseases. It is planned to extend the survey to wheat in 1970. Collaboration on pesticide surveys continued, and data from orchards and hop gardens were analysed. (Church and Hills)

Analysis was completed on a survey of diseases of pedigree beef cattle jointly sponsored by the National Cattle Breeding Association and the Animal Health Trust, and reports sent to the participating Breed Societies. The survey results suggest that, in beef cattle, milk fever and ketosis were less frequent and grass tetany more frequent than in dairy cattle. Mastitis was infrequently recorded; the milder forms of mastitis in many beef cows probably pass unnoticed because the husbandry method does not bring individuals under close observation.

Advice and assistance were given to the A.R.C. Institute for Research in Animal Diseases at Compton in the analysis of extensive data of the constituents of cattle blood. This involved much use of the new version of the General Survey Program. (Leech)

Experiments

Routine analysis. The number of jobs handled increased by 6·6% over 1968, involving in all about 1 150 000 data. The average job size was about 590 data, similar to 1968; exact comparisons are impossible because records for 1968 do not cover the entire year. About 40% of jobs came from the National Agricultural Advisory Service. Data from the Computer Department gives the number of replicated experiments handled by the Statistics Department as 3865 with an average of 6·6 variates analysed per experiment. The number of experiments is more than in 1967, but fewer than in 1968. The number of variates per experiment has risen again after a fall last year. (Dunwoody, Ryan, Watson and Williams)

Crop experiments. Work continued on the interpretation of groups of experiments, done chiefly by the National Agricultural Advisory Service. The idea of representing crop response to nitrogen fertiliser as linear up to a certain dressing, after which there is no further increase (12.4), was applied to results of cereal experiments from several Regions, and a fresh series of multi-level experiments is contemplated. The contribution of soil groups, husbandry methods, and soil P and K contents, to between-site variation in the fertiliser responses of maincrop potatoes was studied. Comparison of seven methods of measuring P, confirmed the superiority of Olsen's sodium bicarbonate method over ammonium acetate/acetic acid, the present N.A.A.S. standard method. (Boyd and Ryan)

Problems in interpreting results of N.A.A.S. spacing and seed-rate trials were examined. (Starkie)

https://doi.org/10.23637/ERADOC-1-124

Livestock experiments. Livestock experiments usually cost more than those with crops, so it is even more important for the statistician to be in close touch with the experimenters. The department continued to co-operate in the design and analysis of livestock experiments on N.A.A.S. Experimental Husbandry Farms and gave much advice on the interpretation of results. Most are simple trials on husbandry problems but analysis began of results from a large-scale, more fundamental, investigation with dairy cows of the effect on milk and liveweight of changes in feed throughout a lactation, extending work started by the National Institute for Research in Dairying.

Much work resulted from the proposed adoption of the metabolisable energy (M.E.) system for expressing ruminant energy requirements. Mathematical models, which simplify calculations of rations, were constructed and other simplifications and possible improvements are being studied. In collaboration with N.A.A.S. nutrition chemists, detailed tables of energy requirements of beef cattle were produced. We used the results of many different experiments to test how well liveweight gains of beef cattle can be predicted from the amount and kind of foods given, and the M.E. and S.E. (starch equivalent) systems were compared. A similar study with dairy cattle is in progress. (Lessells and Watson)

Commonwealth and overseas

Assistance continued with the analysis and interpretation of experiments on cotton (varieties and fertilisers) in Tanzania, cocoa and cotton in Nigeria, cocoa in Ghana, rice in Malaysia and tobacco in Pakistan. Advice was sought on the statistical aspects of work on cotton and maize in Zambia, cotton variety trials in Uganda, oil-palm and cocoa in Malaysia, and pasture and maize varieties in Nigeria. (Walker)

Before a special post at East Malling for corresponding work on tree crops was filled, Walker also dealt with several enquiries on the design of tree-crop experiments, including coconuts (Gilbert and Ellice Islands), and citrus (Zambia).

Other consulting work concerned biometrical problems as varied as: application of classification methods to cowpeas (Ghana); sugar–weather relationships in sugar cane (Barbados); temperature dependence of egg and larval development in *citrus psylla* (Swaziland); the estimation of leaf area in cotton (Tanzania). (Lauckner, Ross and Walker)

Staff and visiting workers

D. A. Preece left to take up a lectureship in Statistics in the Mathematical Institute of the University of Kent. C. E. Rogers, J. D. Starkie, R. W. M. Wedderburn and Jean M. Williams were appointed.

Nelder and Gower attended the Conference on Statistical Computation held at the University of Wisconsin, Madison. Leech visited the New York State Veterinary College, Cornell University, to study and discuss methods proposed for obtaining statistics on the morbidity and mortality of farm animals. Walker paid a short visit to Zambia to advise on the planning

https://doi.org/10.23637/ERADOC-1-124 *pp 8*

and interpretation of agricultural experiments. Members of the department attended meetings of the 37th Session of the International Statistical Institute which was held in London. A paper was contributed by Gower (12.12).

Preece continued as the United Kingdom and Ireland's Regional Editor of *Statistical Theory and Method Abstracts*, published by the International Statistical Institute. During the three years 1 September 1966 to 31 August 1969 he dealt with 1103 abstracts from 53 journals. He also listed 256 published algorithms.

G. N. Wilkinson of the Division of Mathematical Statistics (C.S.I.R.O.), Adelaide, spent 10 weeks in the department implementing an improved version of his general algorithm for the analysis of variance, for inclusion in GENSTAT. Other temporary workers included two from the United Kingdom and one each from the United States, France and the Sudan.

https://doi.org/10.23637/ERADOC-1-124