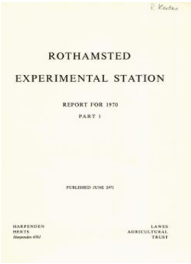
This work is licensed under a Creative Commons Attribution 4.0 International License.

Thank you for using eradoc, a platform to publish electronic copies of the Rothamsted Documents. Your requested document has been scanned from original documents. If you find this document is not readible, or you suspect there are some problems, please let us know and we will correct that.

Report for 1970 - Part1

Full Table of Content



Statistics Department

J. A. Nelder

J. A. Nelder (1971) *Statistics Department ;* Report For 1970 - Part1, pp 208 - 214 **- DOI: https://doi.org/10.23637/ERADOC-1-125**

This work is licensed under a Creative Commons Attribution 4.0 International License.

STATISTICS DEPARTMENT

J. A. NELDER

The department continued to build up a unified flexible set of computer programs for all kinds of statistical analysis, for use both inside and outside the department. Work on statistical methodology goes on alongside much practical application of statistical techniques to a great range of problems, originating both inside and outside the Station. Long-established collaboration with the agricultural advisory service of the Ministry of Agriculture continued, and the service for Commonwealth agricultural institutes was expanded.

Statistical programming

The Genstat system of statistical programs (described in the 1969 Report) was further developed and tested. The new 4-70 computer was not usable until November, so all testing had to be done on the IBM 360 machine at Edinburgh through our direct link. The putting together in one system of many component parts programmed independently exposed some anomalies requiring correction and modification, but in general the specifications laid down for programmers worked well and the system was assembled quite smoothly.

A user's manual was given a limited distribution and useful comments have been received and acted on. The standardising of conventions governing the presentation of data and accompanying instructions by the user greatly diminished the bulk of documentation of our programs. The manual will contain an appendix of worked examples with the associated output, to supplement the formal description. The facilities described in the manual cover:

The GENSTAT language. This is a flexible language oriented towards statistics, with a free format and controlled by a set of directives, such as 'PRINT' or 'REGRESS', defining the different operations. All data are described by standard structures, which can be named by the user (see 1969 Report), and instructions are translated by a compiler into a coded form that drives the complementary programs. (Simpson)

Input–output. The 'READ' directive allows data to be accepted in many different forms, in fixed or free format, serially or in parallel, on one card or many, with or without missing values, etc. (Simpson)

The 'PRINT' directive similarly provides for the printing out of results. Standard formats and layouts are defined, but the user may modify these, e.g. to permute the ordering of tables, restrict the page width, suppress labelling etc. (Nelder)

Both directives have been extensively used, tested by many people, and continuously refined.

Derived variates. Standard mathematical operations are provided, with a useful extension of the customary notation to allow various kinds of patterned operation to be written down simply. Matrix operations and a general calculus for multiway tables are provided. (Alvey and Krzanowski)

Regression. The user can specify linear models containing mixtures of qualitative and quantitative independent variables. Terms can be added, dropped, or interchanged and

https://doi.org/10.23637/ERADOC-1-125

This work is licensed under a Creative Commons Attribution 4.0 International License.

the best new one to add, or worst old one to drop, determined. Regressions can be constrained to go through the origin, or have an intercept. (Wedderburn and Anderson)

Analysis of variance. The Wilkinson algorithm required modification to cope with the accumulation of rounding-errors that occurs in machines of the IBM 360 type. Some parts of the calculation must be done in double precision, and the necessary changes were made. The algorithm can deal with complex designs, and produces full information on standard errors in tables having many different kinds of comparison. (Rogers)

Classification. The user can define various coefficients of similarity between individuals with various measured characteristics, and cluster them using single-linkage cluster analysis. Operations on the similarity matrix are provided. (Ross and Lauckner)

Multivariate analysis. Principal co-ordinate and principal component analysis are provided, including extensive facilities for keeping parts of the output as new data structures for future use. (Krzanowski)

Storage on disc and magnetic tape. Data, with all their associated labels and plain-language description can be stored easily as a computer file for future use; they can be recovered and merged with new data in several ways. In addition the user can inspect the list of contents of GENSTAT files. These facilities give powerful general techniques for handling data, and provide as a bonus a way of 'dumping' a GENSTAT program at any stage during its execution so that it can be used later, e.g. to recover from a machine failure. (Rogers)

Extensions to GENSTAT. Various extensions to the system are being developed. The language is to contain a 'skip' directive for conditional and unconditional jumps (Simpson), logical operations in derived variates (Alvey), facilities for plotting graphs (Alvey), ways of building up the values of one structure from parts of others (Nelder), and a special directive 'OWN' to allow a programmer to add his own private program to the system (Simpson). The regression section is being expanded to allow for hierarchical structure of data, and to incorporate iterative weighted regression. Covariance analysis has been specified to augment the analysis of variance (Rogers). Canonical variate analysis will be provided more simply after the regression extensions. A program is being developed to act as an interface between GENSTAT and the General Survey Program. It will allow tables to be moved between the two programs, so that the special facilities of each system can augment the other. (Leech)

Maximum likelihood program. This program does the same thing as the Orion MLP, but offers a syntax compatible with GENSTAT. It contains a suite of minimisation algorithms, with standard routines for curve and distribution fitting, probit analysis and genetic problems, The user can insert his own Fortran subroutines. Some techniques used in this program have been published (12.19).

A program for producing contour maps on a line printer using interpolated values from equally spaced data values was written primarily to study multi-dimensional likelihood functions, but was also used to represent pictorially the concentration patterns of chemicals injected into the soil. (Ross)

Theory. The extensions to general-purpose computing languages required to meet the needs of statistical computing were considered by Gower for internal data structures

https://doi.org/10.23637/ERADOC-1-125

This work is licensed under a Creative Commons Attribution 4.0 International License.

(12.11), and by Nelder for the input/output of data (12.16). Structure-defining (extensible) languages, such as Algol 68, though providing powerful new programming tools, may not diminish overlapping programming unless statisticians (or any other class of user) define carefully their own standards for data structures (12.15).

An algorithm was developed for defining groups by removing links from a minimum spanning tree. This has useful practical applications in the analysis of large sets of data, and helps to make feasible the classification method proposed by A. W. F. Edwards, by restricting the number of splits of the data set to be examined to splits of the minimum spanning tree. (Ross)

Statistical theory

When measurements exist of a set of characters for a set of individuals (or species etc.), each individual can be represented by a point in many-dimensional space, and the similarity between any two individuals can be described in terms of a distance between their corresponding points. This distance depends not only on the set of characters chosen but also on their relative scales and the particular definition of distance used. Two analyses, using different constellations of characters, give two sets of distances, and it may be asked how similar these descriptions are. An analytical technique was devised whereby one configuration of points is rotated and scaled to minimise the sum of squared distances between corresponding points (R^2), the value of R^2 being a measure of the discrepancy. The technique is relevant to a wide range of problems in multivariate analysis (12.8). (Gower and Krzanowski)

It adds greatly to the power of an analytical method when an initial analysis, say of principal components, using a given set of measurements on given sampling units, can be modified in a stepwise way to take account of adding or dropping a sampling unit or measurement, thus avoiding complete recalculation from the beginning. Stepwise procedures for principal components, principal co-ordinates, canonical variate and canonical correlation analysis were worked out and found to depend on a single basic algorithm. (Gower)

Similarities are often calculated from multivariate ecological data recording the presence or absence of species in a set of quadrats. Such similarities do not allow for the rarity or commonness of a species correctly, and it is better for very rare or very common species to be first excluded, as these species can provide no information on possible associations. (Ross)

Gower considered further problems of weighting characters in similarity coefficients, and of treating correlations between characters in classification techniques (12.9). Krzanowski investigated measures of distance when the material concerns relative proportions of different attributes, such as blood groups in a set of populations.

In regression models of the second kind, the parameters in a model are assumed to have a probability distribution, rather than the observations. Thus, in a fertiliser experiment, variation in the yields obtained might be treated as arising from variation in the parameters connecting yield to amount of fertiliser applied. The estimation of parameters from models of this kind is complicated by singularities in the likelihood function, and the nature of these is being investigated. Iterative weighted linear regression techniques show promise. (Wedderburn)

Practical applications

Members of the department applied a range of statistical techniques to diverse data and the following examples indicate the scope of their work. Multivariate techniques were

https://doi.org/10.23637/ERADOC-1-125

This work is licensed under a Creative Commons Attribution 4.0 International License.

again in demand, and canonical analysis was applied to aphid populations (Entomology Department), nematode populations (Nematology Department) and soil profiles (Weed Research Institute). (Krzanowski) Classification methods and principal coordinate analysis were helpful in analyses of amino-acid in protein (Biochemistry Department) and with botanic data collected as part of the International Biological Programme (Aberdeen University). (Lauckner and Anderson) The rotational fit technique described above was used to compare different multivariate analyses of measurement on skulls from six hominid populations with eight recognisable constellations of characters. (Gower and Krzanowski) Cluster analysis was used in an ecological study of organisms in polluted waters, in the search for associated groups of organisms (Water Pollution Laboratory). (Ross)

An unusual problem in simulation concerned the ordering of octahedral cations in micas, when random linking is constrained by Pauling's rules for the structures of ionic compounds (Pedology Department). The simulated structures are being investigated to try to explain some of the properties of micas. (Krzanowski)

The patterns of diseased plants in a crop may be complex and methods are required to summarise and describe them (Plant Pathology Department). Methods for studying non-randomness in spatial patterns are being surveyed, with special attention to the derivation of the various contagious distributions. (Kempton)

Various curves have been proposed to describe crop response to fertilisers, but little has been done to compare the different curves on the same data. Curves based on the exponential (Mitscherlich) relation are being compared with Nelder's inverse polynomials and with the conjunction of two straight lines proposed by Boyd, which are representable as a limiting case of a simple extension of the inverse polynomial curves. Results from experiments by the Ministry of Agriculture's advisory service on the nitrogen manuring of spring barley are being used initially for this work. (Tong Kwong Yuen)

Data from the National Survey of Health and Development were used in two papers on measures of educational attainment and the effect of various factors on attainment of children between 8 and 18 (12.20, 12.21). Discriminant analysis was the principal statistical technique used. (Simpson)

Surveys

Crops. For the *Survey of Fertiliser Practice* 545 farms in England and Wales were surveyed. Provisional results from 438 farms were obtained by the end of November and the final tabulations by the end of the year. This speed of processing resulted from an unusual stability in both computer hardware and software (programs). A paper (12·7) describes the aims of the current series of surveys and compares practice in England and Wales during 1969 with advisory recommendations. Reports were also prepared on the use of different types of fertiliser and the influence of region and previous cropping on fertiliser use.

Results of chemical analyses from a survey of lime requirements done by the Ministry of Agriculture's Regional Soil Scientists were tabulated and a report prepared. Reports on the Potato Marketing Board's surveys of Main Crop and Early Potatoes were completed (12.17, 12.18) and an examination was begun of the extent to which yield differences can be explained by differences in farm characteristics and practices. (Church and Hills)

Results of a survey of nutrients in glasshouse soils were also reported (12.12). The Survey of Foliar Diseases in Barley, organised by the Ministry of Agriculture's Plant Pathology Laboratory, was extended to wheat, and the basic tabulations made using the

https://doi.org/10.23637/ERADOC-1-125

This work is licensed under a Creative Commons Attribution 4.0 International License.

General Survey Program. (Hills) Consultative work on that Laboratory's Pesticide Surveys continued. The place of sample surveys in methods of assessing crop loss was discussed in a contribution to the F.A.O. manual (12.6). (Church)

Livestock. Data from the Royal Veterinary College on the reproductive system of ewes at different times of year when subjected to different treatments were analysed by the technique of fitting constants. The General Survey Program was used on data from a survey of biting flies in Uganda. A study on teat necrosis in piglets was planned.

Results of a survey of pig parasites in slaughtered animals, collected by the Animal Health Trust, were analysed.

Five national surveys on cattle disorders provided data to study the relation between the proportion of affected herds in a population and the proportion of affected animals, and a formula was derived to predict the former from the latter. The formula gives useful prediction over the disorders investigated—calf mortality, brucella infection in cows, abortion, stillbirth, and Johne's disease in cows. The predictions could be useful in calculating the size of the administrative problem created by proposals to control a given disorder. (Leech)

Experiments

Routine analysis. The volume of work again increased, the number of jobs by 11% and the total of data by 28% to more than 1·5 million. Average job size increased from about 590 to 700 data; about 41% of jobs, involving 34% of the data, came from the agricultural advisory service of the Ministry of Agriculture, a similar proportion to last year. Records from the Computer Department on analyses of variance show that 4487 experiments (an increase of 16% over 1969) were analysed with an average of 8·2 variates per experiment. However, increasing use of other programs is not reflected in these figures. (Dunwoody, Ryan and Williams)

Crop experiments. Several papers were prepared jointly with other workers. The subjects include the relationship between soil phosphorus and response by sugar beet to phosphate fertiliser on mineral soils, potato manuring on fen silt soils, and long-term residues of phosphorus fertilisers. A new series of experiments testing many amounts of nitrogen for continuous cereals is being done on Experimental Husbandry Farms and commercial farms; the experiments are expected to show how far the form of the N-response curve differs between sites and years, and to indicate the reasons for these differences (12.5). (Boyd)

Further analyses of results from the agricultural advisory service of the Ministry of Agriculture included a comparison of methods of pH determination; lime requirement in relation to organic-matter content and pH; a comparison of soil analytical methods for potassium and magnesium. (Starkie)

Results of the manurial reference plots of nine Experimental Husbandry Farms, and from straw disposal experiments at six, were summarised, and the analysis of the new series of national grassland manuring experiments begun. (Ryan)

Livestock. Between 1965 and 1968 four Experimental Husbandry Farms of the agricultural advisory service of the Ministry of Agriculture cooperated in an experiment involving one lactation of 189 cows and heifers given different amounts of concentrate

https://doi.org/10.23637/ERADOC-1-125

This work is licensed under a Creative Commons Attribution 4.0 International License.

during two feeding periods followed by a common régime to measure residual effects. The analysis was complicated by imbalance caused by dying and by year-to-year variation. Results suggest that effects of feeding are usually small when a cow's initial yield is poor, but increase considerably as it increases. Proportional increases in yield with increased concentrate were similar during both periods (3–10, and 12–18 weeks), and the residual effects, during the second period, of treatments applied in the first, were similar to the direct effects.

Six systems for predicting the energy requirements of dairy cattle were investigated for the Ruminant Energy Requirements Working Party, and a report prepared on three of them, the A.R.C. system and two versions of the Starch Equivalent method. This followed similar work on beef cattle (12.3) reported previously. The standard A.R.C. system was examined further using new calorimetric data from the United States Department of Agriculture. (Lessells and Watson)

Commonwealth and overseas

Work for overseas institutes, supported by the Overseas Development Administration, involved experiments in Barbados, Bolivia, Gambia, Ghana, Malawi, Nigeria, Sabah, Tanzania, Uganda and Zambia. Crops included cotton, cocoa, maize, groundnuts, oilpalm and sugarcane, and treatments comprised fertilisers, varieties, and spacing. Assistance was mainly with analysis and interpretation, including the combination of results from different sites and years. Of special interest were more than 100 fertiliser trials with cotton in Tanzania, from sites classified initially by 11 zones, 5 years, and 4 soil types. Analysis showed that fertiliser responses could be classified by combinations of six groupings of zones and two of soil types, and optimum dressings were calculated for each. Results from 25 spacing trials with cotton in Uganda agreed well with inverse polynomial models; the effect on yield of drought frequency at different stages of growth was also investigated. Other work concerned association between ant species and capsid beetles on cocoa farms, the distribution of bollworm eggs on cotton plants, and the influence of soil-moisture status on sugarcane growth. (Walker and Wimble)

Staff and visiting workers

A. J. B. Anderson and Bridget I. Lowe left. R. Jones, R. A. Kempton, Glenda Robinson, A. Todd, Lowsing Tong Kwong Yuen and R. H. Wimble were appointed, Wimble to a post financed by the Overseas Development Administration.

Gower is working for a year (from October 1970) in the Biostatistics Department, University of North Carolina, Chapel Hill. Walker gave a lecture course on field experimental technique and design at the University of Ibadan.

G. N. Wilkinson, who is joining the department in January 1971, paid a brief visit, and A. J. Vernon spent three months in the department before leaving for a second tour with the Fiji Department of Agriculture. P. Brain was appointed as A.R.C. Cadet Statistician to study at Imperial College, and will spend parts of each vacation in the department.

Nelder attended a Civil Service Department Senior Management Seminar, 'Techniques of Economic Appraisal'. He was also Chairman of the Working Party on Statistical Computing which organised a conference 'Statistical Computation, 1970' in London in July. Nelder and Ross attended the 7th International Biometric Conference held in Hanover, and Nelder and Rogers took part in a Workshop Meeting on Analysis of Variance organised by the International Institute of Statistical Computing held at the

https://doi.org/10.23637/ERADOC-1-125

This work is licensed under a Creative Commons Attribution 4.0 International License.

University of Wisconsin. Boyd attended the 9th Congress of the International Potash Institute and gave a paper (12.5). Gower attended the Royal Society/Romanian Academy of Sciences Congress on Mathematical Methods in the Archaeological and Historical Sciences (12.10). Ross visited the Mauritius Sugar Industry Research Institute to advise on their biometric problems, in particular the cane breeding and selection policy. This visit was organised by the Overseas Development Administration.

https://doi.org/10.23637/ERADOC-1-125