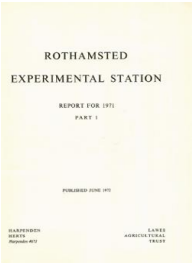
ROTHAMSTED RESEARCH

Report for 1971

Full Table of Content



Statistics Department

J. A. Nelder

J. A. Nelder (1972) *Statistics Department ;* Report For 1971, pp 229 - 237 **- DOI: https://doi.org/10.23637/ERADOC-1-126**

STATISTICS DEPARTMENT

J. A. NELDER

The central part now played by the computer in the department's work is both inevitable and desirable. In providing general programs for statistical analysis we can not only give the research worker useful tools to examine his data, but we also gain new insights into the underlying theory by the process of explaining to the computer, in the act of programming, what we are trying to do. The response by users of all kinds to the facilities provided and the information put out by the programs gives us valuable guidance on the relevance of our work to our customers.

Statistical programming

The development of the Genstat statistical system was severely hampered by faults in the Multijob operating system on the 4-70 computer. Thus it took five months before the Composer, the system program that links together various subprograms for execution, could be persuaded to work properly on Genstat. Persistent operating-system crashes, with accompanying loss of files, led to much waste of programmers' time, and to delay in making Genstat available to users of the 4-70. Three versions now exist, a large one with maximum space for data; a small one for running concurrently with terminal use; and a checking version, which detects and lists any errors in syntax but does not execute the program. A version at the Edinburgh Regional Computing Centre has been in use on their IBM 360/50 since the spring. The reliability of the system is encouraging, and useful suggestions for extensions have been received from several A.R.C. Stations. Additions and modifications include:

Input/output. A 'DESCRIBE' directive allows descriptive material, such as units of measurement, to be associated with a variate name, and output in suitable contexts. The output routines were re-written to speed operation. (Simpson, Nelder, P. K. Leech)

Two input channels are now available, so that a user may read previously stored data from a file different from that holding his program. Output may be direct to the line printer, or to a disc file for subsequent punching on paper tape. A separate output channel is provided to hold error messages which can be examined from a terminal. (Simpson)

The graphical routines were simplified and facilities for scaling improved. (Alvey)

Regression. Sums of squares and products matrices can now be formed within groups and an associated set of group means saved. This gives the basic facility on which analyses of hierarchical data may be built. (Wedderburn)

Analysis of variance. Two major extensions were implemented, covariance analysis and the splitting of sets of effects into individual degrees of freedom, both being done in the context of a very general type of analysis involving possibly many error terms. These extensions led to a complete re-structuring of the basic algorithm, involving the use of complex list structures; these proved their worth in clarifying the underlying logic and in making the program smaller. (Wilkinson and Rogers)

https://doi.org/10.23637/ERADOC-1-126

Multivariate analysis. Canonical variate analysis, as a special case of principal coordinate analysis, was provided. (Banfield)

Program control. A directive 'JUMP' for both conditional and unconditional jumps was incorporated, and work began on a general looping facility in the user's language. (Simpson)

Data movement. A statistical system should make it easy for the user not only to read in his data in many ways but also to reassemble them inside the machine in new forms as the analysis proceeds. The directive 'EQUATE' allows subsets of data from one or more structures to be transferred to parts of other structures, the transfer being controllable by format lists allowing repetition or skipping of items. The flexibility of data-handling internally is thus made comparable to that obtainable on input with the 'READ' directive. (Nelder)

Links with other programs. The program to link Genstat with Yates' General Survey Program (GSP) is now working on the Edinburgh computer (F. B. Leech and P. K. Leech). Faults in the operating system on the 4-70 prevented the mounting of the Genstat package for disc storage. A link was programmed to allow data stored by Genstat to be accessed by SYMAP, a package originated by Professor H. J. Fisher, Laboratory for Computer Graphics, Harvard University, for constructing maps on the computer. (Rogers)

Direct recording of data. Data produced directly on paper tape by the balance at Rothamsted Farm can now be checked on the 4-70 computer and stored on a disc in a form usable as input to Genstat. (Lauckner)

Maximum likelihood program. This program can now be used on the 4-70. It includes automatic display of the likelihood function for any model and any set of data. It can be used for curve fitting, probit analysis, dilution series, fitting frequency distributions and for genetic linkage estimation. The user can supply his own models by writing short Fortran subroutines. More than 20 kinds of curve can be fitted and the output includes standard errors of fitted values and graphs of fitted curves and original data. There are methods of comparing different sets of data by fitting parallel curves and several weighting schemes are allowed. Probit analysis includes the fitting of parallel lines and the probit plane, and estimation of unknown control mortality. The distributions fitted include the Poisson, Negative Binomial, Neyman Type A and Polya-Aeppli for discrete data (counts) and the normal, compound normal and log-normal for continuous data, all with histograms of observed and fitted frequencies. The models for genetic linkage estimation are for F_2 and F_3 segregations with dominance in both loci, and for F_2 segregation with one semi-sterile marker. Many extensions of the program are in progress. (Ross, Jones and Kempton)

Theory. The Genstat compiler is being rewritten to make use of modern techniques of precedence grammars in the analysis of syntax. As a by-product, inconsistencies in syntax are exposed, and a more uniform notation can be developed with fewer rules for the user to master. (Simpson)

The description of models for the analysis of variance requires symbolic formulae to express a model succinctly. The notation originally developed for the analysis of variance algorithm in Genstat was extended and simplified and an algorithm developed

https://doi.org/10.23637/ERADOC-1-126

for expanding a formula into an explicit list of terms. The way is now open for a unified way of writing linear models for fitting to both balanced data (using the analysis of variance algorithm) and unbalanced data (using the regression algorithm). Further extensions are envisaged to describe non-linear models, particularly those having a partly linear component (14).

The analysis of the lattice designs is facilitated by a device introduced by Yates whereby the, say, 25 varieties in the lattice are treated as if they were represented by a 5×5 factorial structure of two pseudo-factors. After the analysis of variance has produced the basic standard errors, the pseudo-factorial structure defines the various types of comparison possible among the varietal means. The analysis of the lattice designs can be simply specified if the required pseudo-factors can be detected by the computer; a way of programming this has now been found (Wilkinson) and will be included in a later version of Genstat. Meanwhile the user can specify lattice structures himself in a straightforward way. (Alvey)

The standard method of finding principal coordinates from the similarity matrix is often limited by restrictions on the amount of core store. When the first few vectors only are required, an old method based on iterative multiplication has considerable advantages because it leaves the similarity matrix unchanged and can easily be programmed to allow the matrix to be segmented and stored mostly on backing store. Much larger problems can then be tackled. (Ross)

A recent algorithm in Algol for extracting the latent roots and vectors of a symmetric matrix was translated into Fortran and exhaustively tested. It was also compared with the algorithm currently used in Genstat. The new algorithm was more efficient for matrices of order less than about 45, but less efficient for larger ones. It was generally more accurate and more robust, e.g. less likely to cause overflow in awkward cases. A version was prepared for publication. (Todd with Mr. D. E. Sparks of Audits of Great Britain Limited)

Statistical theory

Generalised linear models underlying the standard procedures of regression and the analysis of variance are linear models in which the measured yield, say, is supposed to be built up from the sum of systematic effects attributable to treatment factors and their combinations, plus further random effects describing residual variation. For a continuous quantity, such as a weight, a linear model may be adequate and the random effects often have an approximately Normal distribution. Sometimes a transformation is applied to the data before analysis to make the linear model more appropriate; a difficulty now appears in that the best transformation for the systematic components of the model may not correspond to the best one for the random components. When the data are discrete, e.g. relate to counts or percentages, such difficulties are certain to arise, together with further complications arising from the non-normality of the errors. To cope with this situation various techniques of analysis have been devised, including probit analysis, the fitting of constants to tables of percentages using the logit scale, models for contingency tables of counts, and methods of estimating components of variance. These underlying models can all be shown to belong to a class of generalised linear models, in which the systematic effects are linear on one scale of measurement, and the random effects, on another scale, belong to a class of distributions including the normal, Poisson, binomial and gamma. This class of models can be fitted with a single algorithm, that of iterative weighted least-squares, convergence being guaranteed in many commonly-occurring situations. These generalised linear models unify the analysis of many kinds of data. (Nelder and Wedderburn)

https://doi.org/10.23637/ERADOC-1-126

Methods of classification. The logical basis of numerical methods of classification, as used increasingly in taxonomy, etc., received further study. The raw material for the methods are the measurements for a set of characters on a set of individuals; in the simplest case the characters are of the qualitative yes–no kind. The general aim is to produce a set of classes (whose number is not generally known in advance) the individuals in which resemble each other more than they do the individuals in the other classes. A common approach is to define a taxonomic distance between two individuals based on the similarity or dissimilarity of their characters, and then to use the set of distances so formed to define the classes by minimising distance variability within classes compared to that between them. However, although the criterion of classification is here well defined, it tells us nothing directly about the members of the classes so formed. It also carries with it an implicit assumption of homogeneity within the classes, an assumption often not justified in taxonomic problems.

An alternative approach is being developed, based on the idea of maximal predictive classification. For any partition into classes, the values of each character can be listed for each class and used to predict the characters of an individual on being told that it belongs to that class. These class predictors can be chosen so that the maximum number of characters is predicted correctly. The maximal predictive criterion selects the partition into classes so that the total number of correct predictions is as large as possible. It can be used to define an optimal hierarchical classification in a way that does not depend on the idea of taxonomic distance. The relationship between this and other methods of classification is discussed in (12·5). (Gower)

Bacteriologists would like to have a set of standard tests guaranteed to identify with certainty any unknown specimen. A method using a criterion based on information theory was developed and involves calculating the decrease in uncertainty caused by adding each test to the batch and selecting the best one. Though this method cannot be guaranteed to select the best set, it should lead to one very close to the best. (Ross) A computationally simpler, but nearly equivalent, criterion was derived and used to determine a minimum set of sugars to identify yeasts (12·6). (Gower with Dr. J. A. Barnett, Food Research Institute)

Many computer programs for classification methods use transfer algorithms, in which a provisional allocation is modified by making suitable transfers of individual units from one group to another, to improve the allocation according to the criterion adopted. A problem arises in defining an initial allocation not too far away from the final (optimal) one; Ross investigated the dissection of the minimum spanning tree to provide such an initial allocation. Such dissection techniques also give a powerful method for getting the most out of a single run of the classification program. A study of different measures of distance for multinomial data was completed (12·8), and a useful application found in work on metabolic pathways in yeasts when the data takes the form of contingency tables relating to assimilation of different sugars. (Krzanowski)

Mathematical basis of analysis of variance. In simple least-squares theory each parameter in the model can be estimated by a distinct combination of data values, and a standard error assigned to the estimate based on a single error term. Complex experimental designs require extensions to the simple theory first to cope with multiple error terms and secondly to allow for the possibility of aliasing, which occurs when distinct parameters are estimated by the same combination of data values. With partial aliasing there is a partial overlap between the sets of estimators for two sets of parameters, and the detection of these situations during analysis by a computer program requires a characterisation of analysis of variance in terms of various kinds of symmetry that can arise. Such a characterisation was discovered and, when fully developed, will provide

https://doi.org/10.23637/ERADOC-1-126

a unified method of least-squares analysis applicable to the full range of possibilities, from the highly organised type of data obtained from balanced experimental designs to the unorganised type to which regression analysis is commonly applied.

Multiple error terms can lead to a situation where more than one estimate of a parameter can be made, with different accuracies, and a single estimate is required combining the information. Balanced incomplete block designs provide a well-known example. The general characterisation of analysis of variance mentioned above also leads to a simple way of combining information applicable to a great range of experimental designs. (Wilkinson)

Likelihood optimisation and transformations. The fitting of non-linear models by maximum likelihood usually needs iterative trial-and-error methods, whereby provisional estimates of the parameters are refined and the likelihood brought nearer to the true maximum. The iterative process may be stable from some starting points and converge to the right answer, or diverge and run wild from others. Techniques have been developed whereby the zones of convergence and divergence can be readily identified and rapidity of convergence shown by a set of nested contours. The reasons why transformations of the data can improve the possibility and speed of convergence, and why some standard methods diverge can now be better understood. (Ross)

When some types of complex model are being fitted to data the predicted values against which the data are being matched cannot be calculated explicitly, but must themselves be estimated by simulating instances of the model and averaged to make the unwanted random error smaller. Clearly the amount of simulation should be as little as possible, consistent with achieving sufficient accuracy, and for this complex computer algorithms are required to control the fitting process automatically. Gause's famous data on the growth of *Paramecium* in a limited culture medium were used with a density-dependent birth-and-death model to explore the problem further. Changing the form of the parameters in the model affected the speed at which the best fit was obtained. (Kempton)

Practical applications

These are again divided, with some inevitable overlap, into: (i) work involving analyses of an exploratory kind; (ii) surveys, for which the organisation of data collection and processing are major components; (iii) experiments, mostly groups of experiments done cooperatively where the statistician often has an important role in interpretation.

A frequent aim in exploratory analysis of data is the search for relationships between varying quantities. Data from the Entomology Department were used to look for relationships between night illumination and insect catches from light traps. (Church) Weather data during the first four months of the year were examined to see whether any useful prediction could be made of the ensuing incidence of sugar beet yellows. (Lauckner with Marion Watson, Plant Pathology Department) For some situations well-established types of relation exist; Ross is compiling a systematic description of all the commoner types of curves depending on a few parameters met in biology, including their origins and methods for fitting them. Many are included in the Maximum Likelihood Program (see above).

Canonical variate analysis attempts to replace many individual measurements, on a set of individual units previously grouped into classes, by a few combinations of measurements (canonical variates) that are as similar as possible within classes and as divergent as possible between them. When, in particular, two canonical variates suffice, graphical display of their values gives a simple, readily appreciated representation of the original data. This form of analysis was applied in the continuing investigation on the relation-

https://doi.org/10.23637/ERADOC-1-126

ships between human skeletons taken from different localities (Krzanowski with Mr. D. R. Brothwell, Natural History Museum), in the differentiation of carrot varieties (Kempton), and to soil measurements in a long-term experiment with hops at Rosemaund Experimental Husbandry Farm. (Ryan)

Cluster analysis seeks to use multiple measurements on units having no prior structures to define a set of classes. The CLASP program was used to classify water types throughout the world in an attempt to produce a small set of basic types that would simplify the instructions for mixing agricultural chemicals. (Lauckner with Mr. B. Crozier of the Plant Pathology Laboratory) Assistance was given in using the program to the Local Government Operations Research Unit, Reading, who aimed to clarify discussions of the needs of elderly people by identifying typical contrasting cases. Individuals could then be assigned to a group by matching with the nearest typical case, and the whole population thus summarised in terms of these groups. Over 900 cases were used and a very satisfactory classification emerged. (Ross)

For many years the form of yield response to nitrogen fertilisers was thought likely to differ too unpredictably from place to place and from season to season for its detailed investigation to be worthwhile. However, re-examination of past experiments, together with results of many recent multi-level nitrogen experiments with cereals done by the Agricultural Development and Advisory Service, has shed much new light on the nitrogen response curve, and hence on nitrogen requirements and on how these are affected by factors such as disease, crop rotation and kind of soil. Although no one type of response curve best fitted all the results, response on most sites was characterised by a linear or only slightly curved rising portion with a rather abrupt transition to a second portion where further N was ineffective or yield slowly decreased; as expected, functions without a falling asymptote, such as the unmodified exponential and inverse linear fitted poorly; the quadratic, because of its symmetry, consistently exaggerated fertiliser requirements for maximum yields in these experiments. (Boyd, Tong Kwong Yuen and Sparrow)

Surveys

Rothamsted Insect Survey. The data from this survey of insect catches from both light and suction traps made over the past five years would occupy about 75 000 punch cards. Extraction of information from the original records was extremely laborious and processing by computer had become essential. So far about 30 000 cards with data from 110 sites have been punched, each card containing the daily catches of species at one site for a 28-day period. The total yearly catches of 64 species of special interest have been extracted and the information used to make distribution maps. Species frequency distributions from the yearly catch for 20 sites are almost all J-shaped and the Fisher logarithmic series fits them moderately well.

The diversity of a sample is a concept of interest to ecologists and various measures of it have been proposed. There are, however, surprisingly few examples of such measures being used to compare populations, or of attempts to investigate their statistical properties. Data from this survey are being used to study measures of diversity derived from the logarithmic series and the log-normal distribution. (Kempton)

Fertiliser practice. The series of country-wide surveys was continued in 1971, with a sample of 585 farms. Final tabulations for England and Wales were completed by December, using the Mk. II General Survey Program on the Orion computer. Final results for 1970 were ready early in the year, and more precise regional estimates of fertiliser use were obtained by combining information for 1969 and 1970. Data for these two years allowed for the first time a straightforward comparison between the survey estimates of total consumption and those based on subsidy claims. The subsidy figures were about

https://doi.org/10.23637/ERADOC-1-126

10% less. Comparisons are affected by changes in the stocks held, but any consistent bias should soon be known. (Church and Hills)

Livestock surveys. Data from a survey of bovine dentition sponsored by the Smithfield Club were analysed, to determine whether or not the rules for determining age from the pattern of tooth eruptions should be changed. The current rules are based on observations made more than a century ago. The analysis showed that the error in predicting age from dentition is considerable, and that measurements with greater predictive value are needed.

Published reports of national surveys of disease in farm livestock were digested and some general conclusions drawn on methods of improving the planning and analysis of this type of survey (12·11).

Final analyses were completed on a study of teat necrosis in piglets, and of a survey of nematodes in pigs at slaughter. (Leech)

Other work. Consultative work continued on the Survey of Seed Tuber Diseases (Plant Pathology Department and Potato Marketing Board); the Survey of Foliar Diseases in Wheat (M.A.F.F. Plant Pathology Laboratory and Computer Department), and on the M.A.F.F. Plant Pathology Laboratory's pesticide surveys.

Experiments

Routine analysis. The volume of work increased slightly, the total amount of data reached 1·66 million items, an increase of about 4%, though the number of jobs barely changed. The fraction of data from the Agricultural Development and Advisory Service of the M.A.F.F. remained at about one third. Troubles with the 4-70 operating system delayed the intended transfer from the Orion of the bulk of the routine work; however, by the end of the year these had been overcome and the transfer was proceeding smoothly. The first five periods of a large experiment at Gleadthorpe E.H.F. on temperature and ventilation rates for laying fowls were among the first data to be successfully analysed using Genstat. As well as directly supervising the running of jobs on the computer, we advised other Departments within Rothamsted, other A.R.C. institutes and A.D.A.S. on the design, analysis and interpretation of experiments. (Dunwoody, Dyer, Ryan and Williams)

Crop experiments. The Department's contribution to problems of crop nutrition goes back at least to the 1939–45 war and we again collaborated with other disciplines in investigations to improve fertiliser recommendations, for example with workers at Broom's Barn on the phosphorus requirement of sugar beet and its prediction by soil analysis (12·4); similar work on sodium and potassium requirements is nearing completion. In the light of Experimental Husbandry Farm results, experimental techniques for evaluating phosphorus residues were developed with members of A.D.A.S. (12·3); plans for further tests on E.H.F.s were prepared. (Boyd and Sparrow)

Using enough but not too much nitrogen is important in profitable cereal growing; A.D.A.S. recommendations, based on previous cropping, were checked against the results of the many cereal experiments done at Rothamsted and Woburn since 1964. Compared with the small nitrogen residues from previous cereal crops, the nitrogen residues of grazed leys, legumes, potatoes and root crops were much as predicted, each unit of the A.D.A.S. Nitrogen Index being equivalent to about 25 kg/ha N. However, the additional N requirements of wheat and barley in a cereal rotation were much less than this, because their yields were limited by soil-borne diseases, a conclusion broadly confirmed by a study with E. J. Mundy (Norfolk Agricultural Station) of the results of E.H.F. trials on continuous cereal cropping. (Boyd, Ryan, Starkie and Tong Kwong Yuen)

https://doi.org/10.23637/ERADOC-1-126

Analysis of the 1970 results of the national grassland manuring experiments formed one of the largest single items in our work for A.D.A.S.; response curves were fitted to the dry matter yields. (Ryan)

Livestock. Considerable statistical work was done for the final report by the Ruminant Energy Requirements Working Party to the National Conference of Nutrition Chemists on systems for predicting the energy requirements of ruminants. Results were summarised of tests on the A.R.C. system applied to data from sheep, and the relative accuracy assessed of different methods of predicting milk yield of dairy cows. The United States National Academy of Sciences has published formulae for estimating the calorific value of live-weight gain and the accuracy of these was investigated in the light of existing experimental data. The effects of revising various equations in the A.R.C. system were reported and suggestions made for simplifying the general presentation of the system.

Analysis of whole-lactation experiments with flat rates of feeding showed that this form of experiment could usefully be extended to include tests both of amounts and kinds of feed.

Other work included assistance with the design of a standard system, suitable for punched cards, of recording beef carcases, together with a form of primary data reduction. Current data are from experiments but the system is intended also to be useful for future surveys. (Lessells and Margaret A. Currie)

Commonwealth and overseas

The Overseas Development Administration continued to support statistical work for institutes abroad and during the year assistance was given with experiments in Bolivia, Fiji, Ghana, Kenya, Malawi, Nigeria, Sabah, Sarawak, Swaziland, Tanzania, Uganda and Zambia. Variety and fertiliser trials predominated, and the crops concerned included cassava, carrots, cocoa, cotton, cowpea, groundnuts, oil palm, maize, sugarcane, tea and tobacco.

Analyses of all possible crosses between a set of lines (diallel analysis) were done on cotton breeding material from Uganda and Nigeria. A program was written in the Genstat language, showing that it was flexible enough to cope with this rather specialised type of analysis for which no special directive had been included. The associated general data-input facilities proved particularly valuable here.

Two jobs involved insects: genetical data on *Drosophila* from Uganda required the estimation of means and components of variance and covariance arising from parental, F_1, F_2 and backcross generations. Further work is required on the effect of scales of measurement on the conclusions reached. Cluster analysis, using the CLASP program, proved useful in dealing with data on the species represented in 226 sample catches from Ghana.

Our collaboration was sought in a program in Kenya to examine the losses of chemical nutrients by forest trees through leaching from leaves. Information is needed on how various measures of leaching vary between both species and sites, and the relation of between-site variation to rainfall patterns and changes in leaf shape. (Wimble and Robinson)

Staff and visiting workers

W. J. Krzanowski, Lowsing Tong Kwong Yuen and Jean M. Williams left. C. F. Banfield, C. J. Dyer, P. E. Sparrow and G. N. Wilkinson were appointed. P. Walker was seconded to the Federal Department of Agricultural Research in Ibadan. R. H. Wimble took over

https://doi.org/10.23637/ERADOC-1-126

his work on statistical advice to overseas countries. J. C. Gower returned from a year's visit to the Biostatistics Department, University of North Carolina, Chapel Hill.

Gower, Nelder and Wilkinson attended the 38th Session of the International Statistical Institute in Washington and Gower and Nelder gave a joint paper (12·12). Nelder and Ross attended the British Ecological Society Symposium 'Mathematical Models in Ecology' and gave papers (12·2, 12·13). Nelder and Krzanowski attended the Sixth Eucarpia Conference organised by the Plant Breeding Institute at Cambridge and Krzanowski gave a paper (12·9). Nelder took part in a seminar 'The Role of Systematic Design in Forest Experimentation' which was part of the Fourth Forest Research Course at Oxford. He also gave a paper at the Conference on Teaching Mathematics to Non-Specialists organised by the Institute of Mathematics and Its Applications (12·1). He gave seminars on Genstat at the Edinburgh Regional Computing Centre, the National Vegetable Research Station, and the Department of Mathematical Statistics and Operational Research, University College, Cardiff. W. J. Lessells attended the 52nd Meeting of the British Society for Animal Production and presented a paper with Mr. M. J. Strickland of Boxworth Experimental Husbandry Farm.

Leech spent a month in Turkey acting as consultant and lecturing on the application of statistics to research in veterinary laboratories near Istanbul and near Ankara. This was part of an F.A.O. program for development of field veterinary services and regional diagnostic centres in Turkey. He also gave a paper at a symposium on 'The Future of Presymptomatic Diagnosis' at the Royal Society of Medicine, and attended a conference of the Royal Statistical Society on 'Data Validation and Estimation' and gave a paper. At the request of O.D.A., Wimble visited Uganda, Zambia, Malawi and Lesotho for statistical consultancy and also gave a lecture course to research staff at the Department of Agriculture in Lesotho.

Professor J. W. Tukey, Princeton University and Dr. J. M. Chambers, Bell Telephone Laboratories each spent a week in the department and gave seminars. Professor G. E. P. Box of the University of Wisconsin gave a seminar at Rothamsted during his visit to the University of Essex. Two other overseas workers spent several months in the department.

https://doi.org/10.23637/ERADOC-1-126