COMPUTER TECHNIQUES FOR THE ANALYSIS OF EPIDEMICS

by J. KRANZ, *Tropeninstitut, Justus Liebig-University, Giessen, West Germany*

The use of computer techniques for quantitative problems in epidemiology is briefly reviewed. Two computer programmes which, amongst others, have proved to be useful in our studies so far are presented : The Multivariate Regression Analysis with Test (REGT) and a Factor Analysis with Orthogonal Rotation (PAFA).

INTRODUCTION

Computers are fashionable. But they also might be useful, if properly employed. This is particularly true in epidemiology. With epidemiology we mean a comprehensive study of the various quantitative aspects of host and pathogen populations and their mutual impacts under the influence of environmental factors, and the actions of man. Modern approach to epidemiological research involves experimental and measurements under conditions as natural as possible, supplemented by laboratory studies. Field studies of this kind, however, can yield conclusive results only if as many independent variables (like climate, host control measures, etc.) and dependent variables (like sporulation, infection, disease incidence, etc.) as feasible are measured with the highest possible precision. This in turn amounts to a large number of variables and, consequently, vast and confounded data. It is against this background that the use of computers in epidemiology has to be viewed.

Computers may be used 'on the line' for the registration and conversion of data obtained by other devices (e.g. evaluation of disease incidence as measured by aerial infra-red photography, compound weather registration, remote sensing, etc.). Computers could also serve for the storage and retrieval of experimental data. Methods catering for this are at present being developed by some chemical firms and perhaps other institutions. When the considerable organizational problems involved have been overcome this technique will be of great importance, and epidemiologists certainly will benefit from it. The computer as simulator of epidemics has been looked forward to by many plant pathologists, hoping that this could facilitate their problems in disease forecasting. However, little exceeding the numerous formulae and regressions already proposed in literature has come from computers in this respect since. The few examples one may cite are Oort (1968) and Eversmeyer and Burleigh (1970) with mathematical solutions for wheat rusts, the so-called 'negative prognosis' for late blight, a mathematical solution by Schrödter and Ullrich (1966) as well as the interesting, chiefly logical solution offered for early blight of tomatoes by Waggoner and Horsfall (1969). The negative prognosis appears to be the only method of practical use so far. Wheat rusts, early and late blights, however, are pet diseases for the pathologists and therefore comparatively well known. Nevertheless, the authors quoted (except Oort) still were compelled to conduct further experiments to analyse and define suitable para-

meters for their prediction and simulation methods. Analysis of epidemics thus invariably precedes simulation (though simulation could be reverted to analysis), particularly, if knowledge of factors relevant to disease progress is scanty. Above this, analytic methods are required to solve several other problems in epidemiology (analysis of epidemics, fungicidal effects, resistance, etc.)

Results

In our epidemiological studies of the complex nature already outlined, various computer programmes have proved to be suitable tools to achieve analysis (Kranz 1968*a*, 1968*b*; Kranz and Lörinz 1970). One is the Multivariate Regression Analysis with Test (REGT) which shall be described here briefly. This method can deal with a large number of independent and dependent variables. This increases the chances to discover factors having a decisive effect on the course of an epidemic. If successful, it might enable us to define models for simulation and other purposes with very few but essential parameters only. The clue to this is r^2 (the partial coefficient of determination) which, expressed in percentages, gives us the proportion of the over-all variation of the variables measured, explained by one respective variable. Schrödter and Ullrich (1966) also used this criterion, whereas Eversmeyer and Burleigh (1970) chose R^2 (the multiple coefficient of determination) to sort out relevant parameters; REGT dose both of them (and others) in one run and at low costs.

REGT

REGT (Gebhardt 1967) is an abbreviation of 'regression analysis with test'. This programme allows for the input of data of up to 99 variables, of which independent variables x and dependent variables y may be chosen as wanted and warranted by biological reasoning. REGT, therefore, is a multivariate method. The only limitation is that the number of x must not exceed 30. The programme first computes a hypothesis which includes all variables. Thereafter, if possible, so-called 'reduced hypotheses' are being computed stepwise. This means that the independent variable with the absolutely smallest t-value will be eliminated, then a new regression analysis will be started, after which again the independent variable with the smallest t-values will be eliminated, and so on. At each step an analysis of variance is included to test if the residual variation has been significantly increased by the reduction of variables. If this is true and the F-value becomes significant, no further elimination of variables is allowed. Another indication to terminate the analysis is when all remaining t-values have become significant. The output of REGT consists of the inverses of the matrix $x'X$, the $\bar{x} \pm s$ (of the independent variables x only), $b_k \pm s$, t-values, partial correlation coefficients, the multiple correlation coefficient R of each independent variable x with all the remaining independent ones, as well as R of each dependent variable y with all the independent variables x. The programme also prints all the relevant parameters for the analysis of variance and the test of normality of residues. REGT is well suited for a great number of our multivariate problems.

PAFA

Apart from simple correlation analysis, which the computer does easily from bulky material, we have employed the multivariate correlation analysis which is known as Factor Analysis.

The programme PAFA (Schnell 1965) doing a Principal Axis Factor Analysis with subsequent Orthogonal Rotation according to the Varimax criteria has served our purposes.

The number of variables which the programme can deal with goes up to 150, the data of which may either be read in as raw data, factor matrix or correlation matrix respectively. Various communalities may be used in the diagonal of the matrix, e.g. *l*, the highest coefficient of a line, or R^2. After extraction of coefficients from the matrix, normally 4–6 iterations will be done to obtain factors. These, however, cannot make sense without rotation. When this is done, factors eventually contain variables with factor loadings either very high or very low (near 0). One should accept factors only when they have appeared (more or less) in all steps during iteration, although their sequence might change. According to our experience, a general factor and 1 or 2 secondary factors, as usually expected in psychological research, should not be designated in biological research. The reason lies in the fact that the sequence of factors does not imply a rank, though the proportion of common variance explained by each factor decreases. PAFA prints, amongst others, mean the correlation matrix, the factor matrix unrotated and rotated, both the latter for each iteration.

PAFA is a fast and consequently cheap programme. A certain disadvantage is the lack of objective criteria for significance. We have adopted the 0.35 level of factor loadings according to experience in the Institute of Psychology, Giessen University. Nevertheless, results can be quite clear-cut. The factor-analysis could *inter alia* supplement REGT, by testing the highest correlation amongst the variables themselves. This would be of interest, if a relevant variable is difficult to measure or even inaccessible under normal field conditions. In this case a variable with a low *t*-value in REGT which, however, in PAFA appears together in the same factor with the relevant variable having a high loading, both then could serve as a substitute variable if it is easily accessible. Substitutes for real parameters are common in epidemiology (e.g. rainfall).

We have been talking of digital computers so far. Brief mention should be made of the usage of analogue and electronic table computers. The analogue computer can cope with differential equations only, and the results are plotted in curves. Electronic table computers are most appropriate for all common statistical methods unless the data are too confounded. They may also digest and convert data from field lists for the digital computer. This could reduce the high costs arising from the preparation of punch cards, and subsequent filing by the computer. It also would help to save internal storage space, one of the limiting factors in computer usage. We have, for instance, a programme which computes from width and length and number and diameter of leaf spots leaf, the leaf surface area, the percentage of diseased surface area for each individual leaf, and for the whole plant in one run.

Discussion

Computers are extremely quick, precise, and they are inexpensive provided a programme is available and the data are already on punch cards or other units. The computer, on the other hand, being a machine can answer questions only as good as they have been asked. Keeping this in mind, the computer opens venues of research one could not imagine before. In epidemiology it can help to undo much of the com-

plicated interactions one is faced when studying a biological process called epidemic. However, owing to the variation intrinsic to organism, biomathematics only can work out probabilities. Their accuracy and validity entirely depend on the relevance of the variables chosen and the precision of their measurements.

The definition of relevant variables very often has to be achieved stepwise, by retesting hypotheses and thus refining them. But there might be a price for the possible progress. The computer, though it is just a tool, like the microscope, is perhaps more rigid and exacting than any other equipment known to plant pathologists before.

It is therefore conceivable that, as the application of computer techniques increases, all working in epidemiology will have to agree to more commonly accepted terms as well as to standard techniques and approaches.

References

Eversmeyer, M. G., and Burleigh, J. R. (1970). A method of predicting epidemic development of wheat leaf rust. *Phytopathology*, **60**, 805-811.

Gebhardt, F. (1967). Regressionalyse mit Tests (REGT). Programm des Deutschen Rechentzentrums, Darmstadt.

Kranz, J. (1968*a*). Eine Analyse von annuellen Epidemien pilzlicher Parasiten. I. Die Befallskurven und ihre Abhangikeit von einigen Umweltfaktoren. *Phytopath. Z.*, **61**, 59-86.

——— (1968*b*). Eine Analyse von annuellen Epidemien pilzlicher Parasiten. III. Uber Korrelationen zwischen quantitativen Merkmalen von Befallskurven und Ahnlichkeiten von Epidemien. *Phytopath. Z.*, **61**, 205-217.

Kranz, J., and Lörincz, D. (1970). Methoden zum automatischen Vergieich epidemischer Abläufe bai Pflanzenkrankheiten. *Phytopath. Z.*, **67**, 225-233.

Oort, A. J. P. (1968). A model of the early stages of epidemics. *Neth. J. Pl. Pathol.*, **74**, 177-180.

Schnell, P. (1965). Faktorenanalyse (Principed Axes Factor Analysis—PAFA), Programm des Deutschen Rechentzentrums, Darmstadt.

Schrödter, H., and Ullrich, J. (1966). Weitere Untersychungen zur Biometeorologie und Epidemiologie von *Phytophthora infestans* (Mont) de By. Ein neues Konzept zur Lösung der epidemiologischen Prognose. *Phytopath. Z.*, **56** , 265-278.

Waggoner, P. E., and Horsfall, J. G. 1(969). EPIDEM, a simulator of plant disease written for a computer. *Bull. Conn. agric. Expt. Stn.*, **698**, 80.