Note on the classification of multi-level data

By G. N. Lance and W. T. Williams*

This note considers the problem of classifying data in which the specification of a single element by a single attribute is represented, not by a single value, but by a set of values.

Let a set of elements be specified by s attributes, of which the jth attribute takes m_j values ("levels") for a single element. This situation arises in classifying soil profiles, where attributes are recorded at different depths on the same profile; it also arises in classifying elements for which time-dependent attributes are available. There are four obvious solutions: (*a*) to regard the levels as independent, so that each element is specified by a total of $\sum_{j=1}^{s} m_j$ independent attributes, (*b*) to average over levels for each element and compare the averages, (*c*) to compare over the m_j pairs of levels and average the comparisons, and (*d*) to fit a level-dependent function for every element/attribute, and compare the resulting parameters. Of these, (*a*) is theoretically objectionable; it can be shown that, whether or not this is the overt intention, existing numerical methods of classification tend to find "final" groups within which attributes are independent; and the concept of seeking elements with *internally*-independent levels is highly artificial and unlikely to be profitable. Solution (*b*) is unacceptable since it involves discarding what is inherently likely to be important information. Solution (*d*)—commonly used in comparing growth-curves of living organisms—is theoretically the most attractive; but, even if there is *a priori* knowledge of the function to be fitted, the computation is often formidable. Solution (*c*) has obvious advantages; but it rests on the assumptions (i) that m_j is constant for all individuals, and (ii) that the kth level of any one element in some sense corresponds to the kth level of every other.

Being unwilling to concede assumption (ii), Rayner (1966) has suggested a further solution for the case where m_j is constant (say, $=m$) for all attributes. Briefly, it involves, for a pair of elements, first calculating the m^2 similarity coefficients between all pairs of levels; the first level of one individual is then compared with all levels of the other, and that level with the highest similarity is taken as being in correspondence. The process is repeated with the remaining levels of the first element, and then again with the second element as primary referent. If level-correspondence is perfect, the m values in the principal diagonal of the $m \times m$ matrix will be selected, each twice; at worst, $2m$ separate values will be obtained. In any case, the set of $2m$ values is averaged. Even apart from its limitation to constant m_j, we are not attracted to this solution since it, too, rests on an assumption: that if, as is commonly the case, the levels are ordered, the ordering is in the same sense for all elements. This assumption, as Rayner is careful to point out, is substantially valid for his soil-profiles; but it may fail completely if the levels are time-dependent. For example, if the elements are industrial firms, and an attribute is net profit, this might be increasing over levels (here, time intervals) in one firm, decreasing in the other. The similarity coefficients picked out would then be those, not on the principal diagonal, but on the opposite diagonal, and two firms which show the greatest possible difference in this attribute would appear to be identical.

We therefore suggest that renewed attention be paid to solution (*c*). Let the values at the kth level of two elements for an attribute with m_j levels be x_{1k}, x_{2k}; and let the contribution of this attribute to an additive overall inter-element similarity measure be given by

$$\frac{1}{m_j} \cdot \sum_{k=1}^{m_j} \frac{|x_{1k} - x_{2k}|}{x_{1k} + x_{2k}}.$$

Providing all x_{ik} are positive, this measure is constrained between 0 and 1 and is dimensionless; other measures with similar properties could easily be devised, and we have selected this particular one only because it is compatible with that used in MULTIST, the mixed-data classificatory program at our disposal (Lance and Williams, 1966). Missing or inapplicable values for particular levels of either or both elements can be accommodated by not entering a contribution from the level or levels concerned, and reducing m_j accordingly.

The system has been used, with the authors' permission, to classify the set of soils described in Loveday and McIntyre (1966); there were 21 elements with 11 attributes, each at three levels, information for the uppermost level being sometimes absent. A classification was also carried out using the "independence" method (solution (*a*) above). The hierarchies were taken in each case down to the 5-group level, and the results tested by a between/within analysis of variance of two estimates of yield, which had not been used in the classification. With the measure defined above, both variance ratios were significant at the $P = 0{\cdot}05$ level; with solution (*a*) the prediction, as expected, failed, the between-variance being actually less than the within-variance for both estimates. Method (*c*) clearly has the predictive properties that are to be expected from a satisfactory intrinsic classification, despite its theoretical drawbacks. We have therefore incorporated this "linked-level" system, for qualitative or numerical data, in a version of the MULTIST program entitled LINKED.

* *C.S.I.R.O. Computing Research Section, Canberra, A.C.T., Australia.*

References

LANCE, G. N., and WILLIAMS, W. T. (1966). "Computer programs for classification", *Proc. 3rd Aust. Comp. Conf.*, Canberra, paper 12/3.

LOVEDAY, J., and MCINTYRE, D. S. (1966). "Soil properties influencing the growth of subterranean clover in the Coleambally irrigation area, N.S.W.", *Aust. J. Exp. Ag. & Animal Husb.* (in the press).

RAYNER, J. H. (1966). "Classification of soils by numerical methods", *J. Soil. Sci.*, Vol. 17, p. 79.

Book Review

Systems and Simulation, by Dimitris N. Chorafas, 1965; 503 pages. (New York: *Academic Press*, 103s. 6d.)

The general nature of the title of this book indicates the wide field that it sets out to cover. The author says in the introduction "The purpose of this book is to present, explain and discuss in a fundamental manner, some of the mathematical systems which have become popular in professional practice in recent years". The extent of the range covered is immediately apparent from the chapter headings: Mathematical Abstraction and Systems Work, Solving Equations through Statistical Methods, Putting Managerial Data in Mathematical Form, Using PERT in Schedule Control, Markoff Chains in Simulation Studies, Studies in Cargo Handling, Simulation in Hydrological Works, to mention just a few.

The introductory section is followed by six chapters grouped together under the collective title "The Mathematics of the Simulator". This section is confusing in the way that various pieces of mathematics are mixed together. To start with, it is not clear what level of mathematical expertise is expected of the reader; in parts it is extremely elementary. For example, in one place the basic equation for a straight line is explained, with the meaning of "slope" spelt out, which is fair enough for the non-mathematician being introduced to basic concepts. But in previous chapters a number of examples are given involving quite complicated differential equations.

It is not clear what is the main purpose of the section on mathematics. The confusion arises from the ambitious range of the book. It attempts to cover applied mathematics from operational research and econometrics to physics and engineering. The examples in the chapters on mathematics are drawn mainly from physics and engineering; loading of beams, electric circuits, wave motions in longitudinal rods, etc. However, the subsequent sections of the book place the main emphasis on the use of mathematical models for management planning purposes. It would have been of more help to have explained in greater detail the mathematics more applicable to these types of models; in particular mathematical statistics.

The treatment of multiple regression and least squares fitting is indicative of the way that statistics is approached. The subject is curiously avoided in various places. In Chapter 4 when the fitting of an equation to a set of observations is described we are told "the most accurate and possibly the best known means for this test is the method of least squares, but, because of the involved computational procedures that it requires in many cases, we prefer to use approximate methods". It goes on to describe two such methods, referred to as the "method of the selected points" and the "straight line graphical method". In the next chapter the normal equations of simple linear regression are touched upon but not derived. Chapter 9 gives a formula to derive the slope of a fitted trend line, but this formula is not connected in any way to the previously mentioned normal equations. The chapter then goes on to explain multiple regression on a computer. Two very specific and limited routines are referred to and examples given of timings on a "small" machine. A comparison is given to show how much cheaper it is to use a computer rather than a desk calculator.

A reader unfamiliar with multiple regression and its application using a computer is likely to gain a confused picture. This is a pity. It may not be considered necessary to describe the mathematics of multiple regression, but it should be stressed that, given reasonable access to computers, it is nowadays very easy to tackle problems with a large number of independent variables and with effectively no limit to the number of observations.

The section on mathematics is followed by a number of chapters dealing with examples of mathematical models in practice. The examples chosen are challenging and of great interest. One is on a European Common Market Simulator. The problem is posed and a large number of variables defined and functional relationships are suggested. It is indicated that the gross sales potential should be expressed as a function of about a dozen specific variables such as level of public savings, level of employment, etc., and in addition a set of innovation factors. But at this point the example is dropped. The reader is left to wonder what will be the specific form of the equations, and how the coefficients will be derived and tested. The example immediately following describes a mathematical model of civilian air carrier traffic in the USA. A linear regression model is fitted to data in the period 1938–1957 and is then used to forecast the situation in 1965. A range of forecasts of passenger miles flown are plotted against available capacity measured in seat-miles. The example draws the conclusion that the greater the capacity provided the greater will be its utilization. Throughout the wide range of capacity shown this seems a somewhat improbable result.

The book concludes with two rather specialist sections. The first of these deals with hydrological applications. It gives as an example a detailed description of a model developed for the *Tennessee Valley Authority*. The last section consists of two chapters on analogue computing. An outline is given of the characteristics of analogue computers together with some examples of their use.

Very many of the examples given are of the kind of problems that are of great importance to management in practice. They are difficult problems, and any specific approach is bound to draw criticism from various quarters. They do however ensure interest. It would be unrealistic to hope for comprehensive solutions to the examples chosen. Whilst this book is of interest the claim made in the preface (by F. Gordon Smith) that "for the technical planner, and the engineer, this book must surely become a ready reference of great worth" seems a little extravagant.

P. A. B. HUGHES