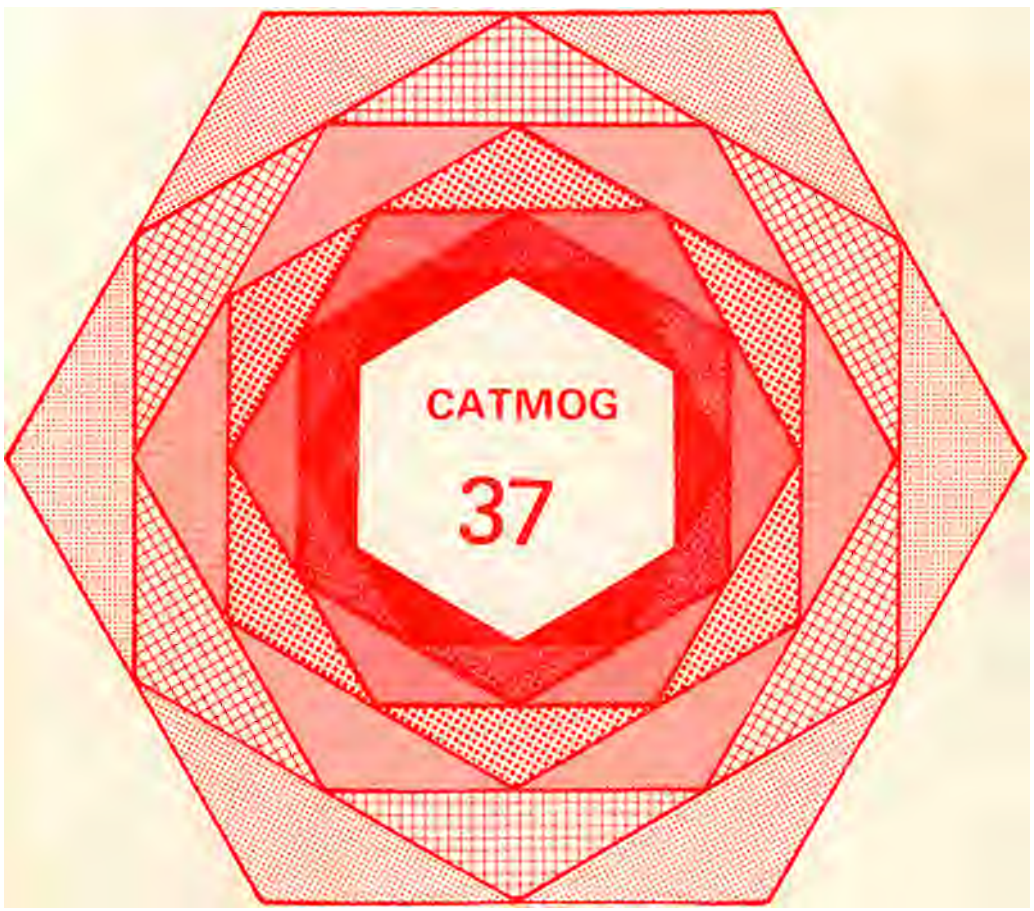
CLASSIFICATION USING INFORMATION STATISTICS.

R. J. Johnston and R. K. Semple



ISSN 0306-6142
ISBN 0 86094 133 7

Published by Geo Books. Norwich — Printed by Headley Brothers Ltd. Kent

CATMOG - Concepts and Techniques in Modern Geography

CATMOG has been created to fill in a teaching need in the field of quantitative methods in undergraduate geography courses. These texts are admirable guides for teachers, yet cheap enough for student purchase as the basis of classwork. Each book is written by an author currently working with the technique or concept he describes.

1. Introduction to Markov chain analysis - L. Collins
2. Distance decay in spatial interactions - P.J. Taylor
3. Understanding canonical correlation analysis - D. Clark
4. Some theoretical and applied aspects of spatial interaction shopping models - S. Openshaw
5. An introduction to trend surface analysis - D. Unwin
6. Classification in geography - R.J. Johnston
7. An introduction to factor analysis - J.B. Goddard & A. Kirby
8. Principal components analysis - S. Daultrey
9. Causal inferences from dichotomous variables - N. Davidson
10. Introduction to the use of logit models in geography - N. Wrigley
11. Linear programming: elementary geographical applications of the transportation problem - A. Hay
12. An introduction to quadrat analysis (2nd edition) - R.W. Thomas
13. An introduction to time-geography - N.J. Thrift
14. An introduction to graph theoretical methods in geography - K.J. Tinkler
15. Linear regression in geography - R. Ferguson
16. Probability surface mapping. An introduction with examples and FORTRAN programs - N. Wrigley
17. Sampling methods for geographical research - C.J. Dixon & B. Leach
18. Questionnaires and interviews in geographical research - C.J. Dixon & B. Leach
19. Analysis of frequency distributions - V. Gardiner & G. Gardiner
20. Analysis of covariance and comparison of regression lines - J. Silk
21. An introduction to the use of simultaneous-equation regression analysis in geography - D. Todd
22. Transfer function modelling: relationship between time series variables - Pong-wai Lai
23. Stochastic processes in one dimensional series: an introduction - K.S. Richards
24. Linear programming: the Simplex method with geographical applications - James E. Killen
25. Directional statistics - G.L. Gaile & J.E. Burt
26. Potential models in human geography - D.C. Rich
27. Causal modelling: the Simon-Blalock approach - D.G. Pringle
28. Statistical forecasting - R.J. Bennett
29. The British Census - J.C. Dewdney
30. The analysis of variance - J. Silk
31. Information statistics in geography - R.W. Thomas
32. Centrographic measures in geography - A. Kellerman
33. An introduction to dimensional analysis for geographers - R. Haynes
34. An introduction to 0-analysis - J. Beaumont & A. Gatrell
35. The agricultural census - United Kingdom and United States - G. Clark
36. Order-neighbour analysis - G. Aplin
37. Classification using information statistics - R.J. Johnston & R.K. Semple
38. The modifiable areal unit problem - S. Openshaw
39. Survey research in underdeveloped countries - C.J. Dixon & B.E. Leach

This series is practiced by the Study Croup in Quantitative methods, of the Institute of British Geographers.
For details of membership of the Study Group, write to the Institute of British Geographers, I Kensington Gore, London SW7 2AR, England.

The series is published by:
Geo Books, Regency House, 34 Duke Street, Norwich NR3 3AP, England, to whom all other enquiries should be addressed.

CONCEPTS AND TECHNIQUES IN MODERN GEOGRAPHY No. 37

CLASSIFICATION USING INFORMATION STATISTICS

by

R.J. Johnston and R.K. Semple
(University of Sheffield) (University of Saskatchewan)

CONTENTS

Acknowledgements

We are grateful to Stan Gregory and Alan Hay for their comments on an early draft of the manuscript; to Jim Forrest for allowing us to use the results of his work with RJJ on voting in London; to Tony Gatrell and two referees for useful assistance; and to Dave Unwin for much valuable help in bringing the computer program up to reasonable portability standards.

I INTRODUCTION

Classification, as described in an earlier monograph in this series (Johnston, 1976a), is a fundamental aspect of scientific activity. Prior to any analysis, the classes of phenomena under consideration must be clearly defined (Grigg, 1965). In addition, classification's widely used, as a descriptive tool, summarising large data sets in a readily appreciated format (see Openshaw, 1983).

Classification is widely used in geography, both as a descriptive methodology and as a prelude to scientific analysis. A great variety of classification methods is available (for a review see Everitt, 1974) and a considerable number have been applied by geographers. Introductions to these are provided in several texts (e.g. Mather, 1976; Johnston, 1976a, 1978). Unfortunately, most of these methods - including those employed in the popular CLUSTAN computer program package (Wishart, 1978) - are not well-suited to many geographical data sets, because of the problems of orthogonalisation and closed number sets (see Johnston, 1977, 1978; Evans and Jones, 1981). Thus geographers have been investigating other methods that may be more appropriate to their data (see Gatrell, 1981; Beaumont and Gatrell, 1982). The present monograph outlines one such method, indicating its statistical basis, illustrating its use, and providing a computer program for its application.

II INFORMATION THEORY: AN INTRODUCTION

Information theory provides a means of analysing closed number sets. With it one can both categorise an individual distribution and compare two or more distributions. Its characteristics have been outlined for geographers in an earlier monograph in this series (Thomas, 1981), as well as in other texts (e.g. Chapman, 1977, Chapter 8). For fuller treatment relevant to geographical applications see Theil (1972), and for applications relating to classification (in ecology) see Williams, Lambert and Lance (1976) and Orloci (1968). Only those aspects of information theory relevant to the classification algorithm presented here are outlined in the present section.

To illustrate our presentation we use the small data set in Table 1. This shows the proportion of the votes cast for each of three political parties in six separate constituencies.

For constituency A, the distribution [0.55, 0.25, 0.20] can be categorized by a single, information, measure. This is the average uncertainty of the various proportions, and is calculated by the formula

$$H = \sum_{i=1}^{n} p_i \log_2 (1/p_i) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (1)$$

Table 1. The hypothetical distribution of votes in six constituencies

Constituency	Party		
	Conservative	Labour	So-Dem-All
A	0.55	0.25	0.20
B	0.20	0.65	0.15
C	0.35	0.48	0.17
D	0.25	0.45	0.30
E	0.55	0.20	0.25
F	0.30	0.30	0.40

where

p_i is the proportion in the ith component of the distribution

n is the number of components in the distribution, and

H is the information measure.

The information value is calculated using logarithms to base two* since it is a binary measure of information content based on yes-no replies to questions which when answered decrease our uncertainty. (See Chapman, 1977, pp. 228ff. for further details.)

For the distribution [0.55, 0.25, 0.20] the information content, using formula (1) is

$$0.55 \log_2 (1.8182) + 0.25 \log_2 (4) + 0.20 \log_2 (5) = 1.4389$$

In itself, this measure of the information content of a distribution is a dimensionless number whose value represents the average number of questions that must be asked with yes-no answers to find out each party's proportional vote. If one party got all the votes our uncertainty about the other parties' votes would be zero. Our uncertainty is maximised when we know that the vote is equally distributed across all parties. To discover that this was the case, given three parties, we would ask an average 1.5851 questions. Using formula (1), $-3(\frac{1}{3} \log_2 \frac{1}{3}) = 1.5851$. This maximum uncertainty, H_{max}, is defined as

$$H_{max} = - \sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n \qquad (2)$$

With the six constituencies in Table 1, the values of H are

A 1.4389 B 1.2790 C 1.4731

D 1.5397 E 1.4389 F 1.5711

Thus we are least certain about the location of any particular voter in constituency F, and most certain in B.

* The base two logarithm can be obtained as the base ten logarithm multiplied by 3.3223.

For any distribution, therefore, there is a maximum value of H which is directly related to n. The relationship between H_{max} and H_{min} can be expressed as a measure of information gain. That is when H = zero we have no uncertainty; when $H = H_{max}$ we have total uncertainty. By asking questions we gain information and thus move towards a position of less uncertainty. Information gain may be defined as

$$I = \log_2 n - H \qquad (3)$$

As indicated by formula (2), the maximum value for H over n components is achieved when $p_1 = p_2 = \ldots p_n$, and is equal to $\log_2 n$. In such a case I = 0, which indicates that the phenomenon being studied (votes in the above example) is equally distributed across the components (parties). The other extreme case occurs when there is a maximum inequality in the distribution, with all of the elements of the distribution in one component and the other components empty. With a three-component distribution, this would be [1.0, 0, 0] giving a value of $H = 1.0 \log_2 1.0 = 0$. 'I' can be calculated from equation (3) or alternatively as:

$$I = \sum_{i=1}^{n} p_i \log_2 (np_i) \qquad (4)$$

and may be interpreted as a measure of the inequality of the distribution across the n components. It follows that the larger the value of I, the greater the inequality (see Semple and Gauthier, 1972). (This is only one of the possible interpretations of I, and is the one relevant to the present task: for other interpretations in different contexts, see Semple and Golledge, 1970; Semple and Wang, 1971; Thomas, 1981; and Chapman, 1977). For the data in Table 1, the values of I are

A 0.1462 B 0.3061 C 0.1120

D 0.0455 E 0.1462 F 0.0140

Note that formula (4) is a special case of the general measure of information gain

$$I = \Sigma\, p \log_2 (p/q) \qquad (5)$$

where

q is a prior or expected proportion, and

p is an observed proportion.

In the present case, q = 1/n (see Thomas, 1981).

We have, then, an inequality statistic which tells us something about the form of a distribution but it must be noted that it is a system measure that does not discriminate according to the order within the distribution. For example, the value of I is identical for the distributions $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$, $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ and $(\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$; as it is for constituencies A and E in Table 1. However, it is possible, as demonstrated in the next section, to develop the I measure in order to classify the component parts of a distribution into groups.

III CLASSIFICATION USING INFORMATION STATISTICS

The aim of this section is to outline a classification methodology developed by Semple in 1971 utilizing the I measure. We first review a sample procedure analogous to the analysis of variance, and then expand this to the more useful multivariate case developed by Semple and Scorrar (1975).

(i) A univariate distribution

To illustrate this procedure we use the small example of five settlements (A - E: see Figure 1) with numbers of retail establishments as follows:

A				
1	3	6	10	12

There are 32 establishments in all, therefore, giving a proportional distribution over the five settlements of [0.03125, 0.09375, 0.18750, 0.31250, 0.37500], for which we derive values of H 1.9844 and I 0.3378; Hmax 2.3222.

Instead of treating all five settlements as a single group, we divide them into two groups, the first containing settlements A and B and the other settlements C, D, and E. The information measure for the whole distribution then becomes the sum of the measures for the separate groups (i.e. we are decomposing the information statistic; for fuller details see Theil, 1972). If we call the groups S_1 and S_2, so that $H(S_1)$ is the information measure for the first group, $H(S_2)$ is the measure for the second and $H(p)$ is the measure for the entire distribution - i.e. $H(p) = H(S_1) + H(S_2)$ - then

$$H(S_1) = \sum_{i=1}^{2} p_i \log_2 \left(1/p_i\right) \tag{6}$$

$$H(S_2) = \sum_{i=3}^{5} p_i \log_2 \left(1/p_i\right) \tag{7}$$

Formulae (6) and (7) can be generalized as

$$H(S_r) = \sum_{i \in S_r} p_i \log_2 \left(1/p_i\right) \tag{8}$$

where S_r is the rth group ($r = 1 \ldots R$) so that $H(S_r)$ is the measure for the rth group. The expression $i \in S_r$ indicates that summation occurs for those i that are members of group S_r. Equation (2) then becomes

$$H(P) = \sum_{r=1}^{R} \left[\sum_{i \in S_r} p_i \log_2 \left(1/p_i\right) \right] \tag{9}$$

The term within square brackets indicates a summation for each group [i.e. each value of $H(S_r)$] and the summation sign outside the square brackets indicates summation across all groups.

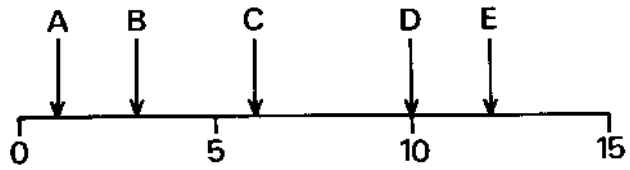


Fig. 1 The number of retail establishments in five settlements.

For the proposed division into two groups, therefore, - where the membership of S_1 is A, B and of S_2 is C, D, E -

$H(S_1) = (0.03125) \log_2 \left(1/0.03125\right) + (0.09375) \log_2 \left(1/0.09375\right) =$

$= 0.15626 + 0.32022 = 0.4765$

$H(S_2) = (0.1875) \log_2 \left(1/0.1875\right) + (0.3125) \log_2 \left(1/0.3125\right) +$
$(0.375) \log_2 \left(1/0.375\right)$

$= 0.4528 + 0.5244 + 0.5307 = 1.5079$

and $H(P) = 0.4765 + 1.5079 = 1.9844$

$H(P)_{max} = \log_2 n = 2.3223$

so that $I(P) = 2.3223 - 1.9844 = 0.3379$.

Within each of the groups it is possible to obtain a measure of the within-group information. In group S_1, for example, the distribution is [0.03125, 0.09375] which in proportional terms, relative to the group total of 0.125, is [0.25, 0.75]. (Going back to the original data, settlements A and B have four retail establishments between them, of which only one (0.25) is in A.) The information for the group can be calculated as

$$p_r \sum_{i \in S_r} \left(p_i/p_r\right) \left[\log_2 \left(p_r/p_i\right) + \log_2 \left(1/p_r\right)\right] \tag{10}$$

in which p_r is the proportion of the total distribution in group r. In formula (10), the first term in brackets is the equivalent of the first term p_i in equation (1) and the term in square brackets is equivalent to $\log_2 \left(1/p_i\right)$ in equation (1). In each case the value of p_i is being expressed as a proportion of p_r, the group total. The summation is standardized by multiplying by p_r (by weighting the information measure by the proportion of the total to which it relates). It is a measure of the information content or evenness of the distribution within group r, and is equivalent to the within-group variance in analysis of variance (see Silk, 1981). Extending it to all groups gives:

$$\sum_{r=1}^{R} p_r \left[\sum_{i \in S_r} \log_2 \left(p_r/p_i\right) + \log_2 \left(1/p_r\right) \right] \tag{11}$$

which can be simplified to

$$\sum_{r=1}^{R} p_r \left[\sum_{i \in S_r} \left({}^{p_i}/_{p_r} \right) \log_2 \left({}^{p_r}/_{p_i} \right) \right] \qquad (12)$$

According to formula (12), the within-group measure is obtained as the sum of the intra-group information measures (the portion of the formula within square brackets), with each intra-group measure weighted by the proportion of the total population in that group.

For the data on retail establishments in the five settlements, evaluation of equation (12) gives the following.

$$\begin{aligned}
&4/32\ [((1/32)/(4/32))\ \log_2\ ((4/32)/(1/32)) + \\
&\quad ((3/32)/(4/32))\ \log_2\ ((4/32)/(3/32))] \\
&+\ 28/32\ [((6/32)/(28/32))\ \log_2\ ((28/32)/(6/32)) + \\
&\quad ((10/32)/(28/32))\ \log_2\ ((28/32)/(10/32)) + \\
&\quad ((12/32)/(28/32))\ \log_2\ ((28/32)/(12/32))] \\
&=\ 0.125\ [(0.03125/0.125)\ \log_2\ (0.125/0.03125) + \\
&\quad (0.09375/0.125)\ \log_2\ (0.125/0.09375)] \\
&+\ 0.875\ [(0.1875/0.875)\ \log_2\ (0.875/0.1875) + \\
&\quad (0.3125/0.875)\ \log_2\ (0.875/0.3125) + \\
&\quad (0.375/0.875)\ \log_2\ (0.875/0.375)] \\
&=\ 0.125\ [(0.25)(2.0) + (0.75)(0.415)] + \\
&\quad 0.875\ [(0.2143)(2.2226) + (0.3571)(1.4856) + (0.4286)(1.2225)] \\
&=\ (0.125)(0.81125) + (0.875)(1.5308) \\
&=\ (0.1014) + (1.3394) \\
&=\ 1.4409
\end{aligned}$$

To complement the within-group measure of equation (12) it is necessary to compute the between-group information measure, which is given by the formula

$$\sum_{r=1}^{R} p_r \log_2 \left({}^{1}/_{p_r} \right) \qquad (13)$$

This treats each group as a single component of a distribution, and hence equation (13) is equivalent to formula (1). From the data for the five settlements, formula (13) is evaluated as

$$\begin{aligned}
&(4/32)\ \log_2\ (1/(4/32)) + (28/32)\ \log_2\ (1/(28/32)) \\
&=\ (0.125)\ \log_2\ (1/0.125) + (0.875)\ \log_2\ (1/0.875) \\
&=\ 0.5436
\end{aligned}$$

The sum of the between-group and within-group measures is

$$H(P) = \sum_{r=1}^{R} p_r \log_2 (1/p_r) + \sum_{r=1}^{R} p_r \left[\sum_{i \in S_r} (p_i/p_r) \log_2 (p_r/p_i) \right] \qquad (14)$$

which, given the data for the five settlements,is

$0.5436 + 1.4408 = 1.9844$

This is the value of H(P) previously reported for that distribution.

As indicated earlier, it is usual to employ the inequality statistic 'I' rather than the information statistic when classification of group inequalities is preferred to group concentrations. The formula, equivalent to that in (14), which divides 'I' into its between and within-group inequalities, is:

$$I(P) = \sum_{r=1}^{R} p_r \log_2 \left(\frac{p_r}{N_r/N} \right) + \sum_{r=1}^{R} p_r \left[\sum_{i \in S_r} (p_i/p_r) \log_2 \left(\frac{p_i/p_r}{1/N_r} \right) \right] \qquad (15)$$

The left-hand term is the between-group inequality, whose value would be zero when $p_r = N_r/N$ - when each group had the same proportion of the distribution as of the membership. That is, with groups of 3 and 2, p_r would be 0.6 and 0.4 respectively. The right-hand term is the within-group inequality, whose value would be zero when every member of each group had the same proportion of the distribution i.e. when $p_i/p_r = 1/N_r$

Evaluating formula (15) for the data on the five settlements gives, for the left-hand term,

$$\begin{aligned}
&(4/32)\ \log_2\ ((4/32)/(2/5)) + (28/32)\ \log_2\ ((28/32)/(3/5)) \\
&=\ (0.125)(-1.6783) + (0.875)(0.5444) \\
&=\ (-0.2098) + (0.4764) = 0.2666
\end{aligned}$$

And for the right-hand term

$$\begin{aligned}
&4/32\ [(1/32)/(4/32)\ \log_2\ ((1/32)/(4/32))/(1/2) + \\
&\quad (3/32)/(4/32)\ \log_2\ ((3/32)/(4/32))/(1/2)] \\
&+\ 28/32\ [(6/32)/(28/32)\ \log_2\ ((6/32)/(28/32))/(1/3) + \\
&\quad (10/32)/(28/32)\ \log_2\ ((10/32)/(28/32)/(1/3) + \\
&\quad (12/32)/(28/32)\ \log_2\ ((12/32)/(28/32))/(1/3)] \\
&=\ 0.125\ [(0.25\ \log_2\ (0.5) + (0.75)\ \log_2\ (1.5)] + \\
&\quad 0.875\ [(0.2143)\ \log_2\ (0.6429) + (0.3571)\ \log_2\ (1.0715) + \\
&\quad (0.4286)\ \log_2\ (1.2853)] \\
&=\ (0.125)(0.1887) + (0.875)(0.0544) = 0.0236 + 0.0476 \\
&=\ 0.0712
\end{aligned}$$

Thus I(P) = 0.2666 + 0.0712 = 0.3378, which is the same as that evaluated earlier.

(ii) Evaluating alternative groupings

The previous section has reviewed a method of decomposing the formulae for measuring information content of distributions [formula (1)] and inequality [formula (4)], to assess the between and within-group variations in the derived index. Clearly, the larger the between-group inequality relative to the within-group inequality, the more effective the grouping, using the criteria of effectiveness of a classification generally employed in such analyses. This effectiveness can be assessed by a test statistic

$$R_s = [I_B(P)/I(P)] \times 100 \qquad (16)$$

where:

I(P) is the total inequality for the given distribution, formula (4);

$I_B(P)$ is the between-group inequality for the given grouping of distribution; the left hand term of formula (15); and

R_s is the test statistic, expressing the between-group inequality for the grouping as a percentage of the total inequality.

For the data analysed here and its groupings [(A,B); (C,D,E)], the value of R_s is [(0.2666)/(0.3378)] x 100 = 78.92. This calculation of R_s evaluates the particular grouping. It can be interpreted as a percentage, with limiting values of 0 and 100.

The purpose of inductive classification is to find the best grouping, so all possibilities must be evaluated. With two groups and five components of the distribution, there are 15 possibilities, as indicated in Table 2. Ten combinations have two members in one group and three in the other; five have a 1:4 division. The values of R_s show that only one other grouping [(A,B,C), (D,E)] comes close to that for the grouping [(A,B), (C,D,E)]. The latter is the best two-group solution for the data set.

Having determined the best classification into two groups, a researcher may wish to proceed to a three-group classification. For the data set employed here, with five components to the distribution, a large number of grout is possible. However, given that the settlements are arranged along a univariate continuum (Figure 1) only those groups which combine adjacent individuals are worth considering. There are six such possibilities, and Table 3 shows that the classification [(A), (B,C), (D,E)] is the best on the analysis of variance criterion used here. (Note that the analysis using this procedure, unlike many others (Johnston, 1976a), is not hierarchical, so that the classification at one level does not constrain that at another.) And, of course, at this finer level of aggregation, three groups rather than two, the test statistic for the best classification is larger; a greater percentage of the variation is between-groups. The best four-group classification is [(A), (B), (C), (D,E)], with a test statistic of 98.76.

(iii) A multivariate extension

The discussion so far has shown how the inequality statistic, I, can be used in an analysis of variance framework to evaluate the best grouping of a population, given that the number of groups is predetermined. As such, the technique is interesting but of little practical value; a conventional analysis of variance would achieve the objective equally well, although its use is limited by certain statistical assumptions. The benefit of the 'I' statistic comes when the grouping is of a multivariate nature.

Table 2. All possible two-group groupings of five components*

Groups		P_r		$I_B(P)$	R_s
(A,B)	(C,D,E)	(4/32)	(28/32)	0.2666	78.92
(A,C)	(B,D,E)	(7/32)	(25/32)	0.1071	31.71
(A,D)	(B,C,E)	(11/32)	(21/32)	0.0097	28.72
(A,E)	(B,C,D)	(13/32)	(19/32)	0.0001	0.03
(B,C)	(A,D,E)	(9/32)	(23/32)	0.0443	13.11
(B,D)	(A,C,E)	(13/32)	(19/32)	0.0001	0.03
(B,E)	(A,C,D)	(15/32)	(17/32)	0.0140	4.14
(C,D)	(A,B,E)	(16/32)	(16/32)	0.0295	8.73
(C,E)	(A,B,D)	(18/32)	(14/32)	0.0773	22.88
(D,E)	(A,B,C)	(22/32)	(10/32)	0.2431	71.96
(A)	(B,C,D,E)	(1/32)	(31/32)	0.1838	54.41
(B)	(A,C,D,E)	(3/32)	(29/32)	0.0606	17.94
(C)	(A,B,D,E)	(6/32)	(26/32)	0.0007	0.21
(D)	(A,B,C,E)	(10/32)	(22/32)	0.0509	15.07
(E)	(A,B,C,D)	(12/32)	(20/32)	0.1175	34.78

Table 3. All possible three-group groupings*

Groups			P_r			$I_B(P)$	R_s
(A)	(B)	(C,D,E)	(1/32)	(3/32)	(28/32)	0.2901	85.88
(A)	(B,C,D)	(E)	(1/32)	(19/32)	(12/32)	0.2475	73.27
(A,B,C)	(D)	(E)	(10/32)	(10/32)	(12/32)	0.2472	73.18
(A,B)	(C)	(D,E)	(4/32)	(6/32)	(22/32)	0.3100	91.77
(A,B)	(C,D)	(E)	(4/32)	(16/32)	(12/32)	0.2913	86.23
(A)	(B,C)	(D,E)	(1/32)	(9/32)	(22/32)	0.3106	91.95

* The data refer to the five settlements whose numbers of retail establishments are shown in Figure 1.

Table 4. A simple multivariate example

A. Original Data

	Establishment Type (j)				
Centre (i)	Food	Clothing	Other		
A	40	30	30		
B	45	35	20		
C	30	35	35		
D	25	35	40		
				Grand Total	400

B. Proportional Data

	Food	Clothing	Other	Total	
A	0.10	0.075	0.075	0.25	
B	0.1125	0.0875	0.05	0.25	
C	0.075	0.0875	0.0875	0.25	
D	0.0625	0.0875	0.10	0.25	
Total	0.35	0.3375	0.3125	Grand Total	1.00

C. Summary Statistics

Centre Mean	0.0875	0.0844	0.0781	
SD	0.0198	0.0054	0.0185	

The data set in Table 4A shows the functional structure of four shopping centres, each containing one hundred establishments; the establishments are subdivided into three categories. The aim is to classify the centres into groups with similar functional structures. To achieve this, the data are expressed as proportions of the grand total: there are 400 establishments, so that the food stores in centre A comprise 0.1 of the total, etc. (Table 4B). The inequality statistic can then be calculated for this matrix as

$$I(P) = \sum_{j=1}^{J} p_j \sum_{i=1}^{N} p_i \log_2 (N\, p_i) \qquad (17)$$

where:

p_j is the proportion of the establishments in type (column) j;

J is the number of columns;

p_i is the proportion of establishments in type (column) j that are in centre (row) i, such that

$$p_i = p_{ij}/p_j \qquad (18)$$

p_{ij} is the proportion of the total in row i, column j; and

N is the number of rows.

Thus:

$$\sum_{j=1}^{J} p_j = 1.0 \qquad (5)$$

and

$$\sum_{i=1}^{N} p_i = 1.0 \qquad (6)$$

but note that the values of p_i refer to the separate columns of the table. Using equation (17), the value of I(P) is the inequality in the distribution of establishments across the centres, weighted by the proportion of the establishments in each type. Its minimum value is zero, when $p_i = {}^1/_N$; the proportion of establishments in each centre is the same in each type. Its maximum value is $\log_2 N$, when some values of p_i are 1.0 and others are zero (i.e. in each type, all of the establishments are in one centre).

Evaluating formula (17) for the data in Table 4B gives

$$\begin{aligned}
&0.35\,[(0.10/0.35)\log_2(4)(0.10/0.35) + (0.1125/0/35)\log_2(4)(0.1125/0.3) \\
&\quad + (0.075/0.35)\log_2(4)(0.075/0.35) + (0.0625/0.35)\log_2(4)(0.0625/0.35)] \\
&+ 0.3375\,[(0.075/0.3375)\log_2(4)(0.075/0.3375) + (0.0875/0.3375)\log_2(4) \\
&\quad (0.0875/0.3375) + (0.0875/0.3375)\log_2(4)(0.0875/0.3375) \\
&\quad + (0.0875/0.3375)\log_2(4)(0.0875/0.3375)] \\
&+ 0.3125\,[(0.075/0.3125)\log_2(4)(0.075/0.3125) + (0.05/0.3125)\log_2(4)(0.05/ \\
&\quad 0.3125) + (0.0875/0.3125)\log_2(4)(0.0875/0.3125) + \\
&\quad (0.10/0.3125)\log_2(4)(0.10/0.3125)] \\
&= 0.35\,[0.0551 + 0.1166 + (-0.0477) + (-0.0867)] + \\
&\quad 0.3375\,[(-0.0378) + 0.0136 + 0.0136 + 0.0136] + \\
&\quad 0.3125\,[(-0.0141) + (-0.1030) + 0.0458 + 0.1140] \\
&= (0.35)(0.0373) + (0.3375)(0.0034) + (0.3125)(0.0426) \\
&= 0.0274
\end{aligned}$$

Having calculated the I(P) statistic - the weighted, inter-row inequality - it is possible to decompose formula (17) to give a between- and within-group inequality. For purposes of classification, as illustrated in the previous section, only the between-group inequality is needed. This is calculated as:

$$I_B(P) = \sum_{j=1}^{J} p_j \sum_{r=1}^{R} p_{jr} \log_2 \left(\frac{p_{jr}}{N_r/N}\right) \qquad (21)$$

which is the multivariate equivalent of the left-hand term in formula (15). In that equation p_{jr} is the proportion of the establishments in type (column) j in group r, or

$$p_{jr} = \left(\sum_{i \in S_r} p_{ij}\right) / p_j \qquad (22)$$

Using the data in Table 4B, evaluation of formula (21) for a two-group classification, with the groups [(A,D), (B,C)], gives:

$$0.35\ [(0.1625/0.35)\ \log_2\ ((0.1625/0.35)/(2/4) +$$
$$(0.1875/0.35)\ \log_2\ ((0.1875/0.35)/(2/4))]$$
$$+\ 0.3375\ [(0.1625/0.3375)\ \log_2\ ((0.1625/0.3375)/(2/4)) +$$
$$(0.1750/0.3375)\ \log_2\ ((0.1750/0.3375)/(2/4))]$$
$$+\ 0.3125\ [(0.1750/0.3125)\ \log_2\ ((0.1750/0.3125)/(2/4)) +$$
$$(0.1375/0.3125)\ \log_2\ ((0.1375/0.3125)/(2/4))]$$
$$= (0.35)(0.0037) + (0.3375)(0.0010) + (0.3125)(0.0104)$$
$$= 0.0049$$

with an R_s value from equation (16) of 17.88. Of the other potential groupings, the best is [(A,B), (C,D)], with an R_s value of 77.06.

(iv) Characterizing groups

Having established the best set of groups in a population, the analyst will probably wish to characterize the groups by identifying their particular characteristics.

Part C of Table 4 gives summary statistics - means and standard deviations - for the proportion of all establishments that are in a particular type, across all four centres. Thus on average, 8.75 per cent of all establishments are food shops, with a standard deviation of 1.98 per cent. Once the groups have been identified, the mean for each can be determined. For the grouping of the four centres into [(A,B), (C,D)] these means are

Group	Establishment Type		
	Food	Clothing	Other
(A,B)	0.10625	0.08125	0.06250
(C,D)	0.06875	0.08750	0.09375

Thus the centres in the first group (A,B) have, on average, a much larger proportion of the total establishments that are food shops than do those in the second group (C,D), whereas the converse applies to the 'other shop' type. There is little difference between the two groups in their proportions in the clothing category.

The mean proportions can be used to characterize each group, therefore. An alternative characterization contrasts the average in each group with the average for all centres. The group means can be compared against the overall means, and presented in standardized form, by using the Z statistic (Blalock, 1960, p. 144) which has the formula:

$$Z = (\bar{x} - \mu)/(\sigma/\sqrt{N_g}) \qquad (23)$$

where

$\bar{x}$ is the mean of the group

μ is the mean for the total population

σ is the standard deviation for the total population; and

N_g is the size of the group

Thus, for the data in Table 4C and the group means given above, the group Z statistics are

Group	Establishment Type		
	Food	Clothing	Other
(A,B)	1.34	-0.82	-1.19
(C,D)	-1.34	0.82	1.19

These show that the centres in the first group (A,B) have above average proportions of food establishments and below average proportions in the other two types; the reverse is the case for the other group (C,D).

The Z statistics are not used in any hypothesis-testing framework to establish the statistical significance of a difference. They simply provide a standardized statement of the degree to which the characteristics of a group differ from those of the population from which it has been separated. As illustrated in more detail below, the larger the Z statistic the greater the difference between the group and the population means on that column characteristic. In this way, the particular features of each group can be emphasised.

IV MORE DETAILED EXAMPLES

The preceding two sections have outlined the relevant portions of information theory for present purposes and have demonstrated, using small examples, how this can be used in multivariate classifications. The basic features of the procedure as outlined are:

1) it groups individuals on the basis of a distribution of values for each individual across J components, so that all members of each group have similar distributions;

2) the basis of this classification is the analysis of variance criterion - the between-group differences are maximised and hence the within-group differences are minimised; and

3) the statistic used is the inequality statistic - I.

For any population, there is a finite number of ways into which the N individuals can be classified into K groups, where K can be any number between 1 and N. The classification procedure finds that grouping (where K is fixed) which produces the maximum value of R_s. It is necessary, therefore, to calculate the value of I(P) for the total population, and that of $I_B(P)$ for each grouping. In the previous section, the small examples allowed the calculations to be done by hand; for larger problems, the tedious procedure for finding

the largest value of R_s can be done using the computer program developed by Semple et al (1972) and modified later in a minor fashion by Johnston (see Section VI).

To illustrate the use of this classification procedure with larger data sets than used in the previous section, several examples are given for only the best groupings at each iteration. As is common with most classification procedures (the notable exceptions being those described in Semple et al, 1969, and Lankford and Semple, 1973), the decision on the optimal number of groups is very largely a subjective one, although a plot of the values of R_s at each stage may assist in the decision-making.

(i) Rainfall regions

For this first example, the small hypothetical data set of Table 5 is used. It provides data for ten stations, giving the average rainfall at each for the four seasons. The aim is to classify the stations according to both their proportion of the total rainfall and its distribution across the four seasons. Inspection of the table suggests that there are five main groupings: 1) high rainfall with winter maxima (A,B); 2) medium rainfall with summer maxima (F,G); 3) medium rainfall with no seasonal variation (C,E); 4) low rainfall with summer maxima (I,J); and 5) low rainfall with winter maxima (D,H).

Table 5. Hypothetical rainfall data (in mm.) for ten stations

Station	Season				Total
	Winter	Spring	Summer	Autumn	
A	500	300	100	300	1200
B	550	280	70	310	1210
C	100	95	105	90	390
D	100	50)	10	50	210
E	180	200	190	210	780
F	50	100	400	100	650
G	60	90	450	110	710
H	80	30	10	30	150
I	10	20	80	20	130
J	20	20	70	20	130
Total	1650	1185	1485	1240	5560

The classification procedure operates on the proportional data in Table 6. With only two groups, and an R_s value of 57.63, the ten stations are divided according to amount of rainfall, with mean proportions of:

Group	Mean Group Proportions			
	Winter	Spring	Summer	Autumn
(A,B,E)	0.0737	0.0468	0.0216	0.0492
(C,D,F,G,H,I,J)	0.0108	0.0104	0.0289	0.0108

The members of the first group have, on average, much larger proportions of the total rainfall in each of the seasons except summer. (Reference back to Table 5 will show that stations A, B and E have the largest annual totals.) The Z statistics for these two groups emphasize this.

Group	Group Z Statistics			
	Winter	Spring	Summer	Autumn
(A,B,E)	2.28	2.46	-0.34	2.46
(C,D,F,G,H,I,J)	-1.49	-1.61	0.22	-1.61

The Z statistics are not particularly large, however, reflecting both the relatively high within-group variation (recall that the R_s value was 57.63), and the small size of the groups (note the role of N_g in formula (23)).

Table 6. The data in Table 5 in proportional form

Station	Season				Total
	Winter	Spring	Summer	Autumn	
A	.0899	.0540	.0180	.0540	.2158
B	.0989	.0504	.0126	.0558	.2176
C	.0180	.0171	.0189	.0162	.0701
D	.0180	.0090	.0118	.0090	.0378
E	.0324	.0360	.0342	.0378	.1403
F	.0090	.0180	.0719	.0180	.1169
G	.0108	.0162	.0809	.0198	.1277
H	.0144	.0054	.0018	.0054	.0270
I	.0018	.0036	.0144	.0036	.0234
J	.0036	.0036	.0126	.0036	.0234
Total	.2968	.2131	.2671	.2230	
Mean	0.0297	0.0213	0.0267	0.0223	
Standard Deviation	0.0334	0.0179	0.0264	0.0189	

With five groups the R_s value increases to 96.53. The members of the groups are those identified above from inspection of Table 5. The group mean values (with Z statistics) are:

Group	Group Mean Proportion (and Z Statistics)							
	Winter		Spring		Summer		Autumn	
(A,B)	0.0944	(2.74)	0.0522	(2.43)	0.0153	(-0.61)	0.0549	(2.43)
(D,H)	0.0162	(-0.57)	0.0072	(-1.11)	0.0018	(-1.33)	0.0072	(-1.13)
(C,E)	0.0252	(-0.19)	0.0265	(0.41)	0.0265	(-0.01)	0.0270	(0.35)
(F,G)	0.0099	(-0.84)	0.0171	(-0.33)	0.0764	(2.66)	0.0189	(-0.26)
(I,J)	0.0027	(-1.14)	0.0036	(-1.40)	0.0135	(-0.71)	0.0036	(-1.40)

Thus the first group (A,B) is characterized by above-average rainfall (proportion of the total amount) in three of the four seasons, whereas the second and fifth groups are characterized by below average amounts (the second group has an even drier summer than the fifth, but a slightly wetter winter). The two stations in the third group have rainfall amounts that are close to the average for all ten stations: those in group four (F,G) have the wet summers.

(ii) Prior standardization of data

The analysis of the data in Tables 5 and 6 produced groups of stations according to total rainfall amount and its seasonal distribution. It may be, however, that the researcher is not interested in the former: the classification required refers to seasonal distribution only. To exclude variability in terms of rainfall amount and to avoid variations in total amount dominating the classification, the seasonal quantity for each station is expressed as a percentage of the station total (as illustrated in Table 7A). This gives each station equal weight in the analysis (i.e. each station has 0.1, or 1/N, of the total rainfall, as shown in Table 7B). In information theoretic terms, Table 7B displays less information than Table 6. Such a prior standardisation may be valuable in many empirical investigations, where the absolute totals are meaningless, as illustrated below.

The plot of R, values for the data of Table 7B is given in Figure 2. With only two groups, over 80 per cent of the variation is between-group - separating the summer minimum stations (A,B,C,D,E,H) with Z statistics of (1.70, 1.88, -1.90, 1.86) from those with summer maxima (F,G,I,J), with Z statistics of (-2.09, -2.30, 2.33, -2.27). With three groups (R, 97.12), the two stations(C,E) with relatively low winter and high summer rainfall are separated from the rest of the first group (A,B,D,H); the second group in the two-group solution remains unchanged. At a finer level still, the five-group solution (R_s = 99.23) identifies the following differences:

Group	Z Statistic for			
	Winter	Spring	Summer	Autumn
(A,B,D)	1.76	1.45	-1.81	1.59
(H)	1.51	-0.02	-1.03	-0.13
(F,G,I)	-1.99	-2.08	2.17	-1.97
(C,E)	-0.27	1.49	-0.35	1.36
(J)	-0.72	-1.01	0.91	-1.14

The first group contains three stations with a summer minimum and a slight winter maximum; the second comprises the one station with a marked winter peak, to be contrasted with the third, comprising three stations with pronounced summer maxima. The fourth group contains stations with an even distribution over the four seasons, which relative to all the means is above average in spring and autumn. Finally, station J has a summer peak, but apparently not as pronounced as in (F,G,I).

These two analyses of the data of Table 5 illustrate the output of the procedure and the relative merits of using raw and standardised data. (Standardised data, which contain less information than non-standardised data, are likely to give higher R_s values for a similar number of groupings.) The data set is a very small one, however; the next example uses a much larger one.

Table 7. The data of Table 5 as percentages of row totals

A. Percentaged data

Station	Seasons				
	Winter	Spring	Summer	Autumn	Total
A	41.67	25.00	8.33	25.00	100.00
B	45.45	23.14	5.79	25.62	100.00
J	19.38	15.38	53.85	15.38	100.00

B. As proportion of total

A	.0417	.0250	.0083	.0250	.1000
B	.0455	.0231	.0058	.0256	.1000
J	.0154	.0154	.0539	.0154	.1001

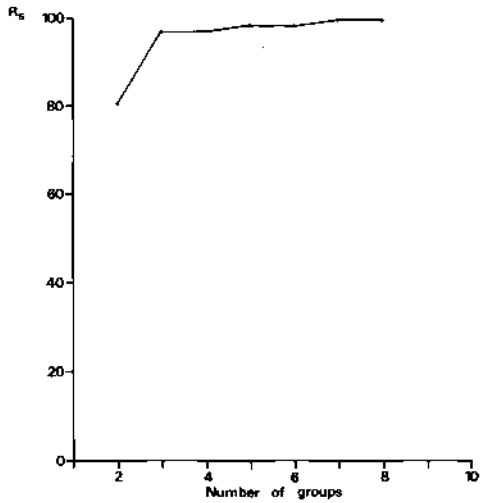


Fig. 2 The plot of R_s against K for the classification of the data in Table 7.

(iii) Agricultural regions of England and Wales in 1801

In 1801 a census of agricultural land use was taken in England and Wales, providing a report on the arable acreage in each parish. The associated crop returns have been used by a number of researchers to identify agricultural regions in various counties (e.g. Hoskins, 1949) and in Wales as a whole (Thomas, 1963). All the available data have recently been collated at the county scale (Turner, 1981) and they are used here to illustrate the classification procedure.

Data are available for the acreage under eight separate crops (or crop-combinations). In addition the total acreage of the parishes in the 55 counties is known, which allows an estimate of the non-arable farmland. Thus one can produce a classification of: 1) counties according to arable farming practiced; and 2) counties according to all agricultural land uses (employing the non-arable area as an indication of the extent of pastoral farming). The data set is a large one, and is not reproduced here. It can be obtained from the original source (Turner, 1981), or from one of the present authors (RJJ).

The purpose of the classification is to group counties which differ very substantially in size. (The total acreage in the parishes whose returns are amalgamated in the data set ranges from 1,056,830 in West Yorkshire to 1,900 in Caernarvon.) Since county size is irrelevant to the analysis, the data are first standardized into percentages of the county totals.

Two analyses were conducted, the first using the data for arable acreage only and the second the data for all agricultural land uses. Figure 3 shows the plot of R_s values for the first of these, and Table 8 indicates the Z-statistics for the two- and ten-group solutions. The latter is mapped in Figure 4A, and provides a clear picture of the regional pattern in the

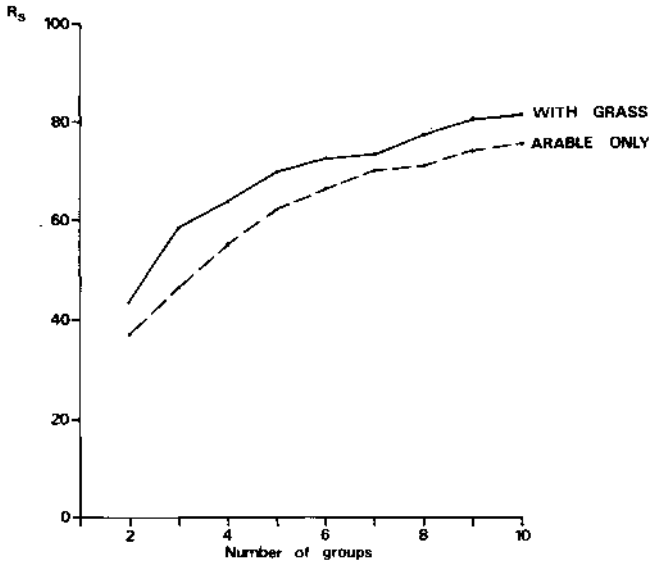


Fig. 3 Plots of R_s against K for the classifications of counties in England and Wales according to agricultural land use in 1801.

agricultural geography of the two countries at that date. Figure 4B shows the three-group solution for the data set including the estimate of grassland. (As Figure 3 shows, the R_s values were consistently higher for this analysis, no doubt reflecting the importance of pasture - an average over the 55 counties of 0.79 of the acreage.) The Z-statistics relating to this second application are given in Table 9.

This example using agricultural land use data illustrates one of the major types of application for this method of classification. Many analyses are concerned, in part if not in total, with classifying places according to a data set that comprises a closed number set. Work in industrial geography, for example, is concerned with the employment structures of different places across sets of industrial and occupational characteristics, whereas the study of urban social areas involves classifying districts according to the social and economic characteristics of their populations. For such data sets, classification based on correlations is inappropriate. Using the method outlined here, however, a ready appreciation of inter-place differences and similarities can be obtained. Since the size of the districts being analysed is usually irrelevant, the prior standardisation approach is sensible, and, as in many cases this introduces a metric with no ready interpretation, the use of Z-statistics to characterize groups is of considerable value, as illustrated in the next section.

Table 8. Classification of English and Welsh Counties by Arable Land Use in 1801

	Z Statistics for							
Group	Wheat	Barley	Oats	Rye	Potatoes	Turnip/ Rape	Peas/ Beans	Other Arable
Two-Group Solution								
1	2.54	0.91	-3.93	-1.05	-2.70	1.76	3.37	0.71
2	-2.94	-1.06	4.54	1.22	3.12	-2.04	-3.90	-0.81
Ten-Group Solution								
1	-0.40	-2.65	2.71	-0.95	3.47	-1.86	-1.59	-0.39
2	0.79	2.44	-0.25	-1.87	1.10	-1.36	-0.35	-0.85
3	2.99	1.08	-2.26	-1.54	0.10	-0.64	0.07	-0.40
4	2.13	-0.41	-2.59	0.09	-0.39	-0.73	3.64	0.32
5	-1.68	1.68	-1.50	0.51	-1.43	4.78	0.39	-0.21
6	1.29	-0.15	-2.27	-0.68	-2.60	-0.15	4.24	-0.34
7	0.13	1.11	-0.70	-0.91	-1.26	1.22	-0.61	4.02
8	-0.38	-1.65	1.40	2.42	-0.33	-0.30	-0.62	-0.33
9	-1.18	-3.43	2.66	1.28	0.09	1.84	-1.70	0.14
10	-4.79	1.72	3.74	0.75	2.75	-2.30	-2.38	-0.75

Table 9. Classification of English and Welsh Counties by All Land Uses, 1801

	Z statistics for								
Group	Wheat	Barley	Oats	Rye	Potatoes	Turnip/ Rape	Beans	Other Arable	Grass
Three-Group Solution									
1	3.74	5.29	1.06	0.85	-1.49	5.13	3.24	3.25	-4.89
2	2.50	1.34	-0.41	1.03	-0.74	0.66	2.54	-0.16	-2.00
3	-3.97	-3.42	-0.01	-1.36	1.32	-2.67	-3.81	-1.11	3.92

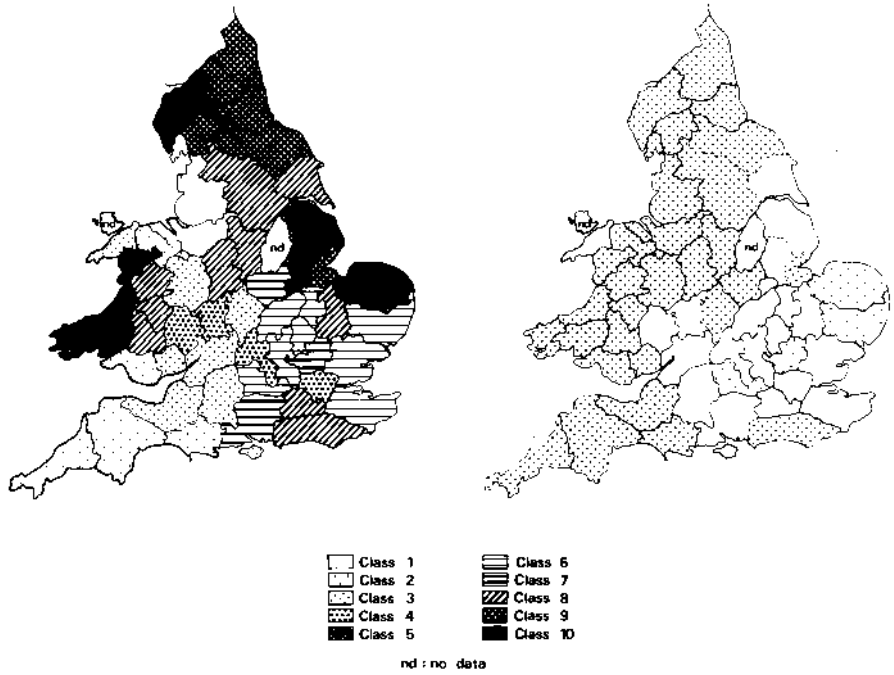


Fig. 4 Classifications of the counties of England and Wales according to agricultural land use in 1801. A: the ten-group solution for arable uses only; B: the three-group solution for all land uses.

V APPLICATIONS

Classification is widely employed in many areas of geographical research. The typology of units produced may be used as an end in itself, as a description of the structure in the data set; it may be used as a test of hypotheses regarding the existence of classes or regions; or it may be employed as a framework for further analysis, as, for example, in the delineation of strata for sampling exercises. The classification method outlined here has a particular strength in providing a procedure that can be used with closed data sets when the research requires that all the components of that set be investigated. Such research is common among geographical applications.

(i) Social areas in cities

The study of social areas has been popular among urban geographers for more than two decades. The main purpose of such work is to illustrate the diversity of social and housing characteristics among the residential districts

of an urban complex. As such, the definition, mapping and description is to some extent an end in itself, illustrating the outcome of processes of residential congregation and segregation whose nature can only be inferred from the results (Johnston, 1980). In addition, however, the results of such classifications are frequently used as sampling frameworks and as inputs to the analysis of the morphology or urban areas.

Social area analysis almost invariably relies upon the data provided by censuses and similar surveys, relating to small districts of cities with populations of, at most, a few thousand. The data available refer to such characteristics of the population as age structure, household sizes, the occupations of the employed, educational qualifications, and, in some countries, birthplace, ethnic origin, income and religion. In addition, characteristics of the housing stock are frequently portrayed, referring to tenure, density of occupation, condition, value, and possession of certain facilities. The general expectation of social area analysis, derived from ecological theory (Knox, 1982), is that different parts of a city will have different population, household and housing characteristics as a result of the sorting processes that have been typical of the operation of housing markets and which reflect economic and social pressures (Johnston, 1980).

During the last two decades or so the methodology of social area analysis has been that of factorial ecology (Berry, 1971; Taylor, 1978). This takes a matrix of data comprising J variables and N observations. A factor (or principal components) analysis of such a matrix isolates the underlying common elements in the distributions of the J variables across the N observations, and a classification of the scores of the observations on the factors or components defines the social areas. This procedure has two apparent strong points. It removes redundancies in the data set by replacing the J variables by K common factors, and it provides orthogonal data (the factors) which can be used in standard classification procedures. Unfortunately, there are also disadvantages. Replacing the variables by factors introduces an inductive weighting: each factor is equally important in the classification. More importantly, there are technical problems (Johnston, 1976b) of which the most critical is that relating to closed number sets.

Examples of the use of the methodology outlined here for the delineation of urban social areas are provided in several recent papers. For the borough of Thamesdown in Wiltshire, for example, the closed number set referring to the occupational structure of the population was analysed to suggest six types of social area (Johnston, 1979a). A similar exercise was undertaken to categorise districts of Rockhampton according to their population age structures (Forrest and Johnston, 1981; see also Johnston, 1979b). In each of these cases, the analysis dealt with one closed number set only. Two or more such sets can be combined, however. For example, Table 10 gives data for the age, occupational, and dwelling tenure characteristics of the populations of the 20 districts of the City of Dunedin in 1966. (The data are taken from the New Zealand Census, and have been simplified for presentation and analysis here.) The three distributions refer to different segments of the population. For age, the entire population is enumerated; for occupation, only the employed male population is considered; and for dwelling tenure it is the number of separate households: for the first district, the respective totals are 53, 26 and 19. Each individual value is presented in Table 10 as a percentage of the relevant total: for people aged 0-14 as a percentage of the total population etc. For each district, then, the sum of the values is 300 - three percentage distributions.

Table 10. Age, occupation and housing tenure in Dunedin, 1966

	Age				Occupation*				Tenure	
District	0-14	15-20	21-64	65+	I	II	III	IV	Rent	Mortgage	Own Outright
1	7.5	11.3	69.8	11.3	3.8	0	76.9	19.3	47.4	15.8	36.8
2	14.9	12.4	57.8	14.9	19.2	21.8	44.8	14.3	62.0	18.9	19.1
3	15.7	15.4	54.5	14.4	25.4	19.1	38.2	17.3	53.7	23.9	22.4
4	12.8	18.8	55.1	13.3	31.0	18.4	32.8	17.8	61.8	15.8	22.4
5	27.6	13.5	50.7	8.3	22.2	22.2	42.2	13.4	16.1	52.7	31.2
6	23.0	14.4	47.2	15.4	14.8	19.3	52.0	13.9	18.1	46.0	35.9
7	35.0	12.2	47.7	5.2	12.3	19.1	52.9	15.7	36.2	48.8	15.0
8	28.5	9.1	50.5	11.9	17.0	13.0	54.0	16.0	16.6	51.6	31.8
9	12.8	30.0	49.3	15.4	24.0	17.2	43.2	15.5	51.8	20.6	27.6
10	27.8	15.4	45.8	10.9	37.3	23.0	31.1	8.6	16.8	46.2	37.0
11	39.6	7.6	48.9	3.9	19.9	19.3	51.0	9.8	10.0	68.9	21.1
12	36.7	10.4	47.8	5.1	12.2	17.8	53.5	17.5	33.9	52.8	13.3
13	27.3	14.9	45.4	12.4	34.4	22.5	32.2	10.9	10.0	47.6	33.4
14	28.7	10.1	50.5	10.7	20.5	20.2	47.8	11.5	14.5	53.6	31.9
15	27.1	12.1	49.4	11.4	12.0	20.1	53.7	14.2	24.6	46.9	28.5
16	23.9	7.7	48.7	19.7	3.7	11.0	60.0	25.3	63.6	11.7	24.7
17	21.3	11.7	49.1	17.9	6.6	15.4	62.6	15.4	25.3	38.9	35.8
18	28.2	13.9	48.6	9.3	18.4	22.7	46.8	12.1	41.4	39.5	19.1
19	31.1	10.4	48.9	9.6	24.8	24.7	40.7	9.8	9.7	59.2	31.1
20	33.5	12.7	47.3	6.5	10.1	17.4	58.3	14.2	10.0	57.5	32.5

* Key to occupations: I Professional and Managerial; II Other White Collar; III Factory Workers; IV Other.

Table 11. Z statistics for classification of Dunedin social areas

	Variable (for description see Table 10)										
	1	2	3	4	5	6	7	8	9	10	11
Two-Class Solution											
1	-3.06	1.43	2.41	2.04	-0.17	-1.67	0.13	2.37	3.30	-3.42	-0.69
2	2.00	-0.93	-1.58	-1.33	0.11	1.09	-0.08	-1.55	-2.16	2.24	0.45
Six-Class Solution											
1	2.19	0.86	1.68	1.18	1.27	0.51	-1.60	0.85	2.56	-2.24	-1.50
2	0.40	-0.57	-0.82	0.67	-1.55	-0.53	1.50	0.07	-1.52	0.99	1.66
3	-1.46	3.58	-0.25	0.97	0.60	-0.21	-0.50	0.23	1.08	-1.23	0.01
4	1.67	-0.38	-0.85	-2.02	-0.79	0.54	0.37	0.22	0.51	0.65	-2.82
5	1.51	-0.63	-1.05	-1.22	2.15	1.77	-1.76	-2.59	-2.27	2.06	1.16
6	-1.58	-1.11	2.29	1.40	-2.27	-3.37	2.53	2.90	1.81	-2.32	0.63

A classification of the data set in Table 10 indicates that two groups account for 48 per cent of the variation. The Z statistics in Table 11 show the districts in the first group (districts 1, 2, 3, 4, 9 and 16) had relatively few children, and above average proportions of 'other' workers and of households which rented their homes, whereas the second group was characterised by areas with above average percentages of children and of homes being bought on mortgages. With six groups, over 81 per cent of the variation was accounted for. These six social area types were -

1. Areas dominated by rental housing and with few children, (districts 2, 3, 4);
2. 'Average' areas with no outstanding characteristics relative to the city as a whole (districts 6, 8, 15, 17, 20);
3. A single district (9) with a very high percentage of persons aged 15-20 - this is the university area.
4. Areas with few old people and homes owned outright - relatively new outer suburbs (districts 7, 12, 18);
5. Relatively high status residential areas, with above average percentages in occupational categories I and II and below average in III and IV (districts 5, 10, 11, 13, 14, 19); and
6. Relatively low status residential areas with relatively large percentages of rented homes (districts 1, 16; these have a high percentage of tenants of state housing who are not separately identified in the census).

The procedure can be used in cases where the data do not constitute closed number sets. Table 12 gives the information for eight variables for the 22 districts of the city of Whangarei, New Zealand (the data are from the 1971 census of New Zealand, and have been subject to a factorial ecology in Johnston, 1976b). With such data, classification on the raw figures could be misleading, since the mean proportions are much lower for some variables than others. (Recall that in formula (17) the contribution of each column to

Table 12. Characteristics of district populations in Whangarei, 1971

	Variable*							
District	1	2	3	4	5	6	7	8
1	0.18	0.79	0.07	0.14	0.17	0.07	.04	0.02
2	0.12	0.83	0.05	0.15	0.10	0.04	.02	0.12
3	0.17	0.72	0.11	0.24	0.19	0.15	.06	0.03
4	0.23	0.67	0.13	0.54	0.08	0.02	.02	0.39
5	0.21	0.76	0.04	0.16	0.16	0.02	.02	0.03
6	0.20	0.66	0.16	0.32	0.14	0.05	.00	0.07
7	0.20	0.59	0.16	0.26	0.14	0.06	.03	0.03
8	0.29	0.55	0.32	0.37	0.21	0.11	.07	0.04
9	0.24	0.66	0.16	0.27	0.12	0.06	.02	0.08
10	0.33	0.59	0.21	0.35	0.21	0.04	.02	0.12
11	0.40	0.48	0.34	0.64	0.13	0.06	.06	0.09
12	0.18	0.69	0.08	0.21	0.21	0.07	.01	0.09
13	0.31	0.70	0.23	0.35	0.16	0.08	.09	0.15
14	0.16	0.83	0.03	0.17	0.15	0.08	.04	0.06
15	0.16	0.78	0.06	0.24	0.06	0.02	.01	0.23
16	0.20	0.73	0.08	0.21	0.08	0.04	.01	0.11
17	0.17	0.81	0.03	0.13	0.16	0.08	.02	0.09
18	0.08	0.84	0.01	0.08	0.19	0.07	.07	0.04
19	0.16	0.74	0.00	0.67	0.06	0.00	.00	0.68
20	0.16	0.80	0.02	0.09	0.14	0.06	.06	0.12
21	0.17	0.74	0.00	0.13	0.13	0.06	.06	0.13
22	0.20	0.77	0.00	0.19	0.08	0.03	.03	0.20

* Key to variables: 1 proportion of those aged 16+ not married; 2 proportion of one family households; 3 proportion of dwellings that are flats; 4 proportion of dwellings that are rented; 5 proportion of male workers in professional/managerial occupations; 6 proportion of male workers earning $6000+ per annum; 7 proportion of male workers with a University degree; 8 proportion of the population who are Maori.

the I statistic is weighted by the column total.) Thus the data are standardized. Instead of using the normal Z score given by the formula

$$Z_i = (x_i - \bar{x}) / \sigma \qquad (24)$$

which would give negative values in some cases, and so create problems in taking logarithms, the Z scores are transformed into T scores, which have normal distributions with a mean of 50 and standard deviation of 10. The classification proceeds without a standardisation of the data - section IV(ii) - since it is scores relative to the city total rather than the distribution across each district which are relevant.

A classification of the districts into two groups accounts for only 30 per cent of the variation, with the major differentiating characteristics (Table 13) being the first three variables, which suggest a separation of

Table 13. Z-Statistics for classification of Whangarei Social Areas

	Variable (for description see Table 12)							
	1	2	3	4	5	6	7	8
Two-Group Solution								
1	2.99	-2.53	3.15	1.68	1.94	2.11	2.26	-0.80
2	-1.62	1.37	-1.71	-0.91	-1.05	-1.15	-1.22	0.44
Three-Group Solution								
1	-1.69	1.24	-1.23	-1.97	1.11	1.03	0.15	-1.51
2	3.60	-2.85	3.50	1.97	1.63	0.91	2.02	-0.51
3	-0.50	0.32	-0.74	1.82	-2.99	-2.33	-2.17	3.02
Five-Group Solution								
1	2.75	-2.48	2.42	2.30	-0.20	0.06	1.01	-0.33
2	1.25	-1.07	2.05	0.55	1.75	3.02	2.72	-0.78
3	0.78	-1.82	1.04	0.18	1.17	-0.14	-1.30	-0.93
4	-0.50	0.32	-0.74	1.82	-2.99	-2.33	-2.17	3.02
5	-1.92	2.61	-1.86	-2.25	0.75	0.19	0.29	-1.01

the urban area into inner-city and other districts (the districts in group 1 are 3, 8, 10, 11, and 13). With just one more group, however, the value of R_s is 72, mainly because of the separation out of a group of five districts (class 3; districts, 4, 15, 16, 19) with high percentages of Maori residents (Table 13). And with five groups, more than 92 per cent of the information content is accounted for, identifying the following social area types:

1. An area with above average proportion of flats, rented dwellings and unmarried adults, and with below average proportions of one family households (district 11);
2. High status residential areas (3, 8, 13);
3. Average areas (6,7, 9, 10, 12);
4. Low status areas with high proportion of Maori (4, 5, 16, 19); and
5. 'Average' areas except for the above average proportion of one family households (1, 2, 5, 14, 17, 18, 20, 21, 22).

In both of these social area classifications, the number of districts to be classified was relatively small, which allows the full data set to be presented. In most studies the number of observations is considerably larger and the ratio of districts to groups much larger too. Nevertheless, these examples illustrate how the classification procedure can be used to identify the main types of social areas within a city, with the Z statistics indicating the major differentiating characteristics for each type.

Although both of the examples presented here relate to urban social areas, the method is equally applicable to a wide range of classification problems in human and physical geography (as well as other disciplines). Applications in economic geography have been suggested earlier. In physical geography, classification of areas according to their plant composition, to

the nutrient characteristics of their soils, to a variety of climatic elements, and so on, or to a combination of such variables, could all be handled by this procedure.

(ii) Typologies as independent variables

In some research exercises, the hypotheses being tested relate some dependent variable to aspects of an area's population structure, and require all of the latter to be incorporated in the analysis. With regard to aggregate analyses of voting behaviour, for example, the researcher may hypothesise: 1) that members of different occupational classes vary in their propensity to vote for a particular party, so that the analysis requires study of all classes to provide a full explanation; and 2) that members of different age groups similarly vary in their propensity to vote for a particular party. Most such analyses use a regression format, but because of the closed number set problem are unable to incorporate all of the classes/age groups. Thus the analyses are incomplete.

The procedure introduced here circumvents that problem by providing a typology of areas that can be used as independent variables, in either an analysis of variance or a regression format. Thus, for an analysis of the percentage voting Labour in each of the wards of the Greater London boroughs in May 1971, to test the hypotheses presented in the preceding paragraph, two classifications were produced to provide the independent variables. A classification of the wards according to their occupational structure provided seven groups accounting for 76 per cent of the variation, and a classification based on age structure resulted in a 12-group solution which accounted for 69 per cent of the variation. These two classifications were used as the independent variables in an analysis of variance of the Labour vote. The socio-economic grouping accounted for 75.7 per cent of the variance; the age structure grouping accounted for 2.5 per cent; and the two groupings together (including their interaction) accounted for 82 per cent. Detailed analysis showed that strongest Labour support came from the dominantly working class wards and in the area of young, working age adult populations.

(iii) Time series analysis

The study of time series by geographers seeks spatial variations in secular patterns of, for example, unemployment and rainfall. Classification of areas with similar trends has been undertaken using the same factorial ecology methods as in social area analysis and it faces the same problems. The procedure outlined here is well suited to the study of time series, however.

As an example, take the study of rainfall trends over a 70-year period at each of 50 stations in the United Kingdom (Gregory, 1975; Johnston, 1981). Did different groups of stations experience similar trends over that period, and were there substantial differences between the groups? To answer this, we have a data matrix comprising 70 (J) columns and 50 (N) rows. The value in each cell is the rainfall at the relevant station in a particular year and the sum of all values along that row is the 70-year rainfall total for the station. If two stations had the same trend over time in annual rainfall, then for each the proportion of the 70-year total which fell in the individual years would be approximately the same. Expressing the rainfall at a station

in each year as a percentage of the total for the period is analogous, therefore, to expressing the population of an area aged 0-14 as a percentage of total population there, or dividing the annual rainfall into its seasonal components (Table 5). Thus the data matrix takes on exactly the same form as those discussed in the previous subsection on social area analysis. Application of the classification procedure to such a matrix groups stations with similar secular trends. Those in the same group will have received similar proportions of their total rainfall in particular years of the sequence.

Application of the classification procedure to the 50 x 70 data matrix of annual rainfall totals led to the identification of ten groups of stations, (R_s = 58.9). A map of the group membership indicated clear regional clusters, with few stations not being allocated to the same group as their neighbours. Thus over the period different parts of the United Kingdom experienced different rainfall trends. Graphing the Z statistics illustrates the nature of those trends (see Johnston, 1981) and allows the characteristic trend of each group to be identified. (For an application using unemployment time series, see Johnston, 1983.)

(iv) Trade flows

The methodology reviewed here is clearly applicable in a wide range of research situations in geography, although as yet it has only been used in a few. As already stressed, its particular strength is in the analysis of closed number sets (percentages and proportions). Thus it has been suggested, for example, as a means of classifying districts according to the distribution of votes across a number of candidates. Voting regions are then defined without any of the problems of the factorial ecology procedure (Johnston, 1982).

Data sets which are not closed can also be analysed with this classification procedure (as with the data set in Table 5). This has been illustrated by two studies of trade patterns. For each of eight dates, for example, Semple and Scorrar (1975) analysed the distribution of the value of Canada's imports & exports according to commodity flows, with each of fifty trading partners. There were two matrices for each date, one for exports and one for imports, comprising either 6 or 9 columns (the commodity classifications changed after 1960) plus 50 rows. Classification grouped together Canada's trading partners according to both the volume of trade and its distribution across the commodity classifications. Thus, for example, the partners were classified into five groups according to the destination and composition of exports in 1930 (R_s = 89.0). The first comprised countries (the U.S. and U.K.) receiving large volumes of agricultural and forest products and substantial amounts of animal and of non-ferrous metal products. The second group (mainly countries in Western Europe) received mainly agricultural exports, in substantial amounts, from Canada, whereas the other three comprised countries receiving only small export volumes from Canada, mainly of agricultural products.

The grouping procedure allowed Semple and Scorrar to portray the changing nature of Canadian international trade and the relative importance of the varying trading partners. Similarly, an analysis of trade in the COMECON bloc since 1946 was used to identify groups of countries with similar import and export commodity structures in their exchanges with the Soviet Union. Data were analysed for 1946, 1950, 1960, and 1968 and suggested that by the 1960s three clear groups (the same for both imports and exports) had emerged,

(Semple and Demko, 1977). These were (Table 14): 1) East Germany and Czechoslovakia, which exported large volumes of industrial machinery and equipment and of consumer goods to the Soviet Union and whose imports were dominated by fuels and raw materials; 2) Bulgaria and Hungary, which exported machinery and equipment too, plus a considerable volume of food and consumer goods, and whose imports were dominated by fuels and raw materials and machinery; and 3) Romania and Yugoslavia, whose exports were dominated by fuel, raw materials, and consumer goods, and whose imports were dominated by fuel and materials and equipment. Only Poland moved between groups in the 1960s; in 1960 it was in the second group but in 1968 in the first.

With these studies of trade flows, the classification procedure has categorized countries according to the quantity and the commodity structure of their imports and exports, showing that groups of countries have very similar trading volumes and profiles with a particular partner. In the second example, the three-group solution for 1968 accounted for 79 per cent of the variation among the seven countries in terms of their exports to the Soviet Union, and 77 per cent for their imports. Thus the classification is a simplifying device, clarifying the similarities among groups of places and highlighting the differences between them.

Table 14. Group means in COMECON trade with the Soviet Union, 1968

Commodity	Group* 1 (N=3)	2 (N=2)	3 (N=2)
Exports to the Soviet Union			
Machinery and equipment	0.105	0.048	0.014
Fuel and raw materials	0.028	0.012	0.020
Foodstuffs	0.001	0.032	0.004
Industrial consumer goods	0.045	0.034	0.018
Other	0.026	0.007	0.001
Imports from the Soviet Union			
Machinery and equipment	0.031	0.047	0.014
Fuel and raw materials	0.112	0.070	0.028
Foodstuffs	0.025	0.006	0.003
Industrial consumer goods	0.003	0.004	0.001
Others	0.032	0.010	0.013

* The figures in the table are mean proportions of the total trade analysed. Thus, for example, of the exports to the Soviet Union, on average 10.5 per cent comprised machinery and equipment from each of the countries in group 1.

VI A COMPUTER ALGORITHM

Although the nature of the classification procedure outlined here is straightforward, its application with large data sets is extremely tedious if undertaken by either hand or calculator. Once the data matrix has been assembled and expressed in the needed format (as in Table 5), the steps are as outlined in Figure 5.

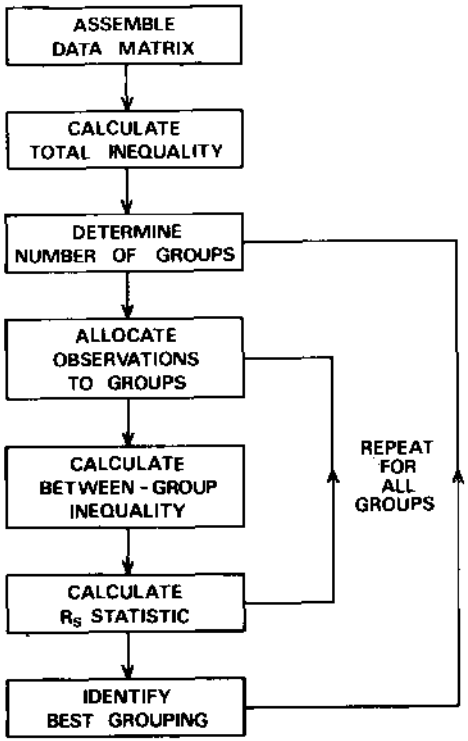


Fig. 5 A flow-diagram for the classification procedure.

For a classification involving more than a small number of observation units the number of times that the inner of the two loops in Figure 5 is traversed is extremely large, because of the very large number of ways in which N individuals can be combined into K groups. A computer program clearly is needed to speed up the process and eliminate the tedious calculations. Nevertheless, even a fast computer will take a long time to evaluate all of the possible groupings in a substantial problem and so methods have been developed to speed up the process. The program produced for the classification procedure (Semple, Youngmann and Zeller, 1972) makes an initial allocation of N individuals to K groups and calculates the R_s statistic. (An initial configuration can be read in, indicating the proposed group for each individual). It then moves each individual into each of the other groups in turn, calculating the R_s statistic and comparing this with the one previously computed. If the new R_s statistic exceeds the previous one, the grouping is retained; if it does not, the individual is returned to its original group. This procedure, seeking a better R_s statistic, continues until no move from a particular grouping produces an increase in the R_s statistic; when this occurs, the best grouping has been found. The program finds the best grouping for each value of K lying between an inputted minimum and maximum.

The program developed by Semple, Youngmann and Zeller (1972) has been modified slightly to allow data to be percentaged across rows and to calculate Z statistics; otherwise, it is unaltered in its basic features. It is made operational by a control card, as follows.

Control Card

Column	Mnemonic	Format	Description
1-16	PRNM	16A1	Title for the run; any alphanumeric characters
17-20	JØBS	I4	Number of observations (rows)
21-24	NVAR	I4	Number of variables (columns of the distribution)
25-26	MINNG	I2	The minimum number of groups to be extracted (if zero, the program inserts 1)
27-28	MAXNG	I2	The maximum number of groups to be extracted (if zero, the program inserts JOBS)
29-30	INTMED	I2	If INTMED = 1, the results of all groupings are printed
31-32	IPER	I2	If IPER = 1 the data are standardised to proportions of the row totals prior to analysis.
33-34	IMEM	I2	If IMEM = 1, a prior grouping is input to the program to speed operations.
37-45	SUB	F9.2	The value to be substituted for zeros in the data (necessary because the log of zero cannot be taken). If not input the program substitutes 1.0
46-55	SINC	F10.2	The grand total. This is calculated in the program if omitted.

For the data set in Table 3A, the full set up was

```
TRIALDATASETFOUR0004000302040001000000000.010000000.00
403030
453520
303535
253540
```

(Statement 103 was changed for this to 3F2.0)

Input Channels

The version of the program included here includes two input channels, 4 and 5. Channel 4 is used for the data set and Channel 5 for the control cards. It is, of course, possible to put the control cards in front of the data set and use one channel only.

A Priori Grouping

The initial stage of the program involves allocating the observation units to an initial set of groups, prior to the switching procedure which searches for the optimal group configuration. With large data sets, the initial allocation (which divides the observations into equal-sized groups according to their numerical position in the data set) may be a long way from the optimum. The search for the optimum is then expensive in computer time.

To reduce the search time, the analyst may wish to suggest an initial configuration that is intuitively close to the optimum. For this, the value of IMEM is set to 1 on the first control card. A further (set of) control card(s) on Channel 5 is then required. This gives the suggested grouping of the observations (the format is 2513). Thus with ten observations and three groups, with observations 1, 3, 5 and 7 in the first group, 2, 4, and 9 in the second, and 6, 8, 10 in the third, that control card would read

0010020010020010030010030020003

Note that:

a) this procedure can only operate for the smallest number of groups asked for (i.e. MINNG on Control Card 1); and

b) if the control cards and the data set are in one file on the same channel, the additional control card comes after the data set.

Logged Data

As laid out in the text above, the calculations use base 2 logarithms. In the FORTRAN program these are produced using the function ALOG2.

Some compilers lack this function. To obtain base 2 logarithms, two possibilities exist.

a) Obtain base 10 logarithms and multiply by a constant, so that the base 2 logarithm of X (L2X) is

```
L2X = ALOG(X) * 3.3223
```

b) Use reciprocals, so that

```
ALG2 = 1.0 / ALOG (2.0)
```

and when ALOG2 (X) is required

```
L2X = ALG2 * ALOG (X)
```


The Output from the program includes the following

A. General

1. The data set
2. A statement DATA IN PROPORTIONS if IPER = 1
3. The data transformed into proportions if IPER = 1
4. The value of SING
5. The total means for each column - note that these are not the means of the raw or transformed data matrix but of the matrix in which cell values are expressed as proportions of SINC. For ease of presentation, they are multiplied by 100.
6. The standard deviations of the total means.

B. For each grouping

1. The group number
2. For each member
   - i) its observation number
   - ii) its data - expressed as a proportion of SINC and multiplied by 100
   - iii) the group means - as for the total means
   - iv) the Z statistics

(Note that running this program with small numbers of observations occasionally produces difficulties. In such cases, it may be necessary to put the data into Double Precision).

BIBLIOGRAPHY

1. Information Theory and Classification

Chapman, G.P. (1977). *Human and Environmental Systems: A Geographer's Appraisal.* (Academic Press, London).

Forrest, J. and Johnston, R.J. (1981). On the characterization of residential areas according to age structure. *Urban Geography,* 2, 31-40.

Johnston, R.J. (1979). On the characterization of urban social areas. *Tijdschrift voor Economische en Sociale Geografie,* 70, 232-238.

Johnston, R.J. (1979b). The homes of callers to a telephone counselling service. *New Zealand Geographer,* 35, 34-40.

Johnston, R.J. (1981). Regarding the delimitation of regions according to climatic fluctuations. *Archives for Meteorology, Geophysics and Bioclimatology,* Series B, 29, 215-228.

Johnston, R.J. (1982). The definition of voting regions in multi-party contests. *European Journal of Political Research,* 8, 293-304.

Johnston, R.J. (1983). An inductive approach to the study of spatial variations in unemployment trends. Regional *studies,* 17, 105-112.

Orloci, L. (1968). Information analysis in phytosociology: partition, classification and prediction. *Journal of Theoretical Biology,* 20, 271-284.

Semple, R.K. and Demko, G.J. (1977). An information-theoretic analysis: an application to Soviet-COMECON trade flows. *Geographical Analysis,* 9, 51-63.

Semple, R.K. and Gauthier, H.L. (1972). Spatial-temporal trends in income inequalities in Brazil. *Geographical Analysis,* 4, 169-179.

Semple, R.K. and Golledge, R.G. (1970). An analysis of entropy changes in a settlement pattern over time. *Economic Geography,* 46, 157-160.

Semple, R.K. and Scorrar, D.A. (1975). Canadian international trade. *The Canadian Geographer,* 19, 135-148.

Semple, R.K. and Wang, L.H. (1971). A geographical analysis of changing redundancy in inter-urban transportation links. *Geografiska Annaler,* 53B, 1-5.

Semple, R.K., Youngmann, C.E. and Zeller, R.E. (1972). *Economic Regionalization and Information Theory: An Ohio Example.* Discussion Paper 28 (Department of Geography, Ohio State University, Columbus).

Thomas, R.W. (1981). *Information Statistics in Geography.* CATMOG 31, (Geo Books, Norwich).

Williams, W.T., Lambert, J.M. and Lance, G.N. (1966). Multivariate methods in plant ecology. V Similarity analyses and information-analysis. *Journal of Ecology,* 54, 427-445.

2. General References

Beaumont, J.R. and Gatrell, A.G. (1982). *An Introduction of Q-Analysis.* CATMOG 34, (Geo Books, Norwich).

Berry, B.J.L. ed. (1971). *Comparative Factorial Ecology. Economic Geography,* 47.

Blalock, H.M. (1960). *Social* ***Statistics.*** (McGraw Hill, New York).

Everett, B. (1974). *Cluster Analysis.* (Heinemann, London).

Evans, I.S. and Jones, K. (1981). Ratios and closed number systems. In: N. Wrigley and R.J. Bennett, eds, ***Quantitative*** *Geography,* 123-134. (Routledge and Kegan Paul, London).

Gatrell, A.G. (1981). On the structure of urban social areas using Q-analysis. *Transactions, Institute of British Geographers,* NS6, 228-245.

Johnston, R.J. (1976a). *Classification in Geography.* CATMOG 6, (Geo Books, Norwich).

Johnston, R.J. (1977). Principal components analysis in geographical research: some problems and issues. *South African Geographical Journal,* 59, 30-44.

Johnston, R.J. (1978). *Multivariate Statistical Analysis in Geography.* (Longman, London).

Lankford, P.H. and Semple, R.K. (1973). Classification in geography. *Geographia Polonica,* 25, 7-30.

Mather, P.M. (1976). *Computational Methods of Multivariate Analysis in Physical Geography.* (Wiley, London).

Openshaw, S. (1983). Classifying data for counties and districts. In: D. Rhind, ed., *census user's Handbook,* (Methuen, London).

Semple, R.K., Casetti, E. and King, L.J. (1969). *The Determinants of Optimal Number of Groupings in Classification Problems.* Discussion Paper 10. (Department of Geography, Ohio State University, Columbus).

Silk, J. (1981). *The Analysis of Variance.* CATMOG 30, (Geo Books, Norwich).

Taylor, P.J. (1978). *Quantitative Methods in Geography.* (Houghton Mifflin, Boston).

Theil, H. (1972). Statistical *Decomposition Analysis.* (North-Holland, London).

Wishart, D. (1978). *CLUSTAN 9C User Manual.* (Program Library Unit, Edinburgh).

3. Other Works Cited

Gregory, S. (1975). On the delimitation of regional patterns of recent climatic fluctuations. ***weather,*** 30, 276-287.

Grigg, D.B. (1965). The logic of regional systems. *Annals, Association of American Geographers,* 55, 465-491.

Hoskins, W.G. (1949). The Leicestershire crop returns of 1801. *Transactions of the Leicestershire Archaeological Society,* 24, 127-153.

Johnston, R.J. (1976b). Residential area characteristics. In: D.T. Herbert and R.J. Johnston, eds., *Social* ***Areas*** *in Cities,* 1, 193-236, (John Wiley, London).

Johnston, R.J. (1980). *City and* ***Society.*** (Penguin, London).

Knox, P.L. (1982). *Urban social Geography.* (Longman, London).

Thomas, D. (1963). *Agriculture in Wales during the Napoleonic wars.* (University of Wales, Cardiff).

Turner, M. (1981). Arable in England and Wales: estimates for the 1801 crop return. *Journal of Historical Geography,* 7, 291-302.

APPENDIX I: PROGRAM LISTING

```
C
C      INFORMATION CLASSIFICATION
C      WRITTEN BY SEMPLE, YOUNGMANN AND ZELLER
C      MODIFIED BY JOHNSTON
C      INQUIRIES TO JOHNSTON DEPT GEOGRAPHY UNIV SHEFFIELD
C
       DIMENSION PRNM(16)
       DIMENSION SD(20),TM(20)
       COMMON IOBS,NVAR,IOUT,ID,MAXNG,MINNG,SUB,IAGG,IPER,IMEM,IN,IX,SINC
      1,INTMED

C      PARAMETERS ARE
C       COLS  1-16 PRNM    ALPHANUMERIC TITLE
C       COLS 17-20 JOBS    NUMBER OF OBSERVATIONS (INTEGER)
C       COLS 21-24 NVAR    NUMBER OF VARIABLES (INTEGER)
C       COLS 25-26 MINNG   MINIMUM NUMBER OF GROUPS (INTEGER:DEFAULT 1)
C       COLS 27-28 MAXNG   MAXIMUM NUMBER OF GROUPS (INTEGER:DEFAULT JOBS)
C       COLS 29-30 INTMED  IF SET AT 1, ALL INTERMEDIATE GROUPS PRINTED
C       COLS 31-32 IPER    IF SET AT 1,DATA CONVERTED TO PROPORTIONS OF ROW TOTAL
C       COLS 33-34 IMEM    IF SET AT 1, A PRIOR GROUPING IS INPUT
C       COLS 37-45  SUB    VALUE SUBSTITUTED FOR ZERO DATA
C                          (FORMAT F9.2) (DEFAULT 1.0)
C       COLS 46-55 SINC    THE GRAND TOTAL (IF OMITTED, DATA ARE SUMMED IN
C                          PROGRAM)
C                                                    IO=1
C      TWO INPUT CHANNELS ARE USED
C      CHANNEL 5 CARRIES THE PARAMETER VALUES
C      CHANNEL 4 CARRIES THE DATA SET
C
    1  CONTINUE
C
C      READ PARAMETERS
C
       READ (5,40) PRNM,IOBS,NVAR,MINNG,MAXNG,INTMED,IPER,IMEM,ID,SUB,SINC
C
C      SET DEFAULTS
C
       IAGG = 1
       IF (SINC.LE.0.) IAGG = 0
       IF (SUB.LE.0.) SUB = 1.0
       IF (MINNG.LE.0.) MINNG = 1
       IF (MAXNG.LE.0.) MAXNG = IOBS
       IF (MAXNG.LT.MINNG) MINNG = 1
       WRITE (6, 50) SUB
    2  CALL INFOST(IDATA,IMEMB1,IMEMB2,IY,IYR,INR,INGRP,ISUMX,IGMEAN,INAM
      1E)
       GO TO 1
   40  FORMAT (16A1,2I4,6I2,F9.2,F10.2,I2)
   50  FORMAT (1H ,27HSUBSTITUTE FOR ZERO DATA IS,F10.2)
       STOP
       END
```



```
C
C      SUBROUTINE INFOST
C
       SUBROUTINE INFOST(IDATA,IMEMB1,IMEMB2,IY,IYR,INR,INGRP,ISUMX,IGMEA
      1N,INAME)
       DIMENSION DATA(99,20),MEMB1(511),MEMB2(511),Y(20) YR(99,20)
       DIMENSION NR(511),NGRP(20),SUMX(20,20),GMEAN(20,20)
       DIMENSION SD(20),TM(20)
       COMMON IOBS,NVAR,IOUT,ID,MAXNG,MINNG,SUB,IAGG,IPER,IMEM,IN,IX,SINC
      1,INTMED
       OBS = FLOAT(IOBS)
C
C      READ AND PRINT DATA
C
       DO 2 I=1,IOBS
       READ (4,103) (DATA(I,J),J=1,NVAR)
       WRITE (6,102) I,(DATA(I,J),J=1,NVAR)
C
C      REPLACE ZEROS
C
       DO 1 J=1,NVAR
    1  IF (DATA(I,J).LE.0.) DATA(I,J) = SUB
    2  CONTINUE
       SINC = 0.0
C
C      CHECK IF TO TAKE PROPORTIONS ACROSS ROWS
C
       IF (IPER.NE.1) GO TO 6
       WRITE (6,100)
       DO 5 I=1,IOBS
C
C      SUM ROW
C
       XTQ = 0.
       DO 3 J=1,NVAR
    3  XTQ = XTQ + DATA(I,J)
C
C      CHANGE TO PROPORTIONS
C
       DO 4 J=1,NVAR
    4  DATA(I,J) = DATA(I,J) / XTQ
C
C      WRITE PROPORTIONS
C
       WRITE (6,102) I, (DATA(I,J),J=1,NVAR)
    5  CONTINUE
    6  CONTINUE
C
C      SUM MATRIX
C
       DO 7 J=1,IOBS
       DO 7 I=1,NVAR
    7  SINC = SINC + DATA(J,I)
       WRITE (6, 101) SINC
C
C      EXPRESS CELL VALUES AS PROPORTION OF GRAND TOTAL
C
       DO 8 I=1,IOBS
       DO 8 J=1,NVAR
    8  DATA(I,J) = (DATA(I,J)/SINC) * 100.0
```

```
C
C     FIND MEANS AND STANDARD DEVIATIONS
C
      DO 9 I=1,NVAR
      SD(I) = 0.
    9 TM(I) = 0.
      DO 10 I=1,IOBS
      DO 10 J=1,NVAR
      TM(J) = TM(J) + DATA(I,J)
   10 SD(J) = SD(J) + DATA(I,J)**2
      DO 11 J=1,NVAR
      TM(J) = TM(J)/OBS
   11 SD(J) = SQRT( (SD(J)/OBS) - TM(J)**2)
      WRITE (6, 104) (TM(J), J=1,NVAR)
      WRITE (6, 105) (SD(J), J=1,NVAR)
C
C     FIND TOTAL INEQUALITY
C
      TINEQ = 0.0
      DO 12 JJ=1,NVAR
   12 Y(JJ) = 0.
      DO 13 JJ=1,NVAR
      DO 13 II=1,IOBS
   13 Y(JJ) = Y(JJ) + DATA(II,JJ)
      DO 15 J=1,NVAR
      X = 0.
      DO 14 I=1,IOBS
   14 X = X + (DATA(I,J)/(Y(J)) * ALOG2(OBS * DATA(I,J)/Y(J))
   15 TINEQ = TINEQ + Y(J) * X
C
C     LOOP GROUP SIZES FROM MINIMUM TO MAXIMUM
C
      STATMX = 0.
      DO 44 NG=MINNG,MAXNG
      IXOUT = 0
C
C     CHECK IF INITIAL GROUP TO BE READ IN
C
      IF (IMEM.EQ.1) GO TO 20
   16 I = IOBS / NG
      L = 1
      M = 1
      DO 18 K=1,NG
      DO 17 J=L,M
   17 MEMB1(J) = K
      L = M+1
   18 M = M+I
      IF (L.GT.IOBS) GO TO 23
      DO 19 I=L,IOBS
   19 MEMB1(I) = K
C
C     READ PRIOR GROUPING
C     THE INITIAL GROUP MEMBERSHIP OF EACH OBSERVATION IS READ IN
C     IN A VECTOR (LENGTH=JOBS) THE FORMAT OF THE VECTOR IS 2513
C
   20 CONTINUE
      IF (IMEM.EQ.1.AND.IX.NE.1) READ(5,106) (MEMB1(I),I=1,IOBS)
   21 IX = 1
      GO TO 23
```



```
   22 IMEM = 0
C
C     SWITCH GROUPS TO FIND OPTIMUM
C
   23 OLSTAT = 0.
      ICNT = 0
   24 IND = 0
      DO 41 I=1,IOBS
      DO 40 J=1,NG
      ISTORE = MEMB1(I)
      MEMB1(I) = J
      IND = IND + 1
C
C     CALCULATE BETWEEN-REGION INEQUALITY
C
      DO 29 II=1,NG
      DO 25 K=1,NVAR
   25 YR(II,K) = 0.
      NR(II) = 0
      DO 27 JJ=1,IOBS
      IF (MEMB1(JJ).NE.II) GO TO 27
      DO 26 K=1,NVAR
   26 YR(II,K) = YR(II,K) + DATA(JJ,K)
      NR(II) = NR(II) + 1
   27 CONTINUE
      DO 28 JJ=1,NVAR
   28 YR(II,JJ) = YR(II,JJ) / Y(JJ)
   29 CONTINUE
      BINEQ = 0.
      DO 31 JJ=1,NVAR
      BINEQ2 = 0.
      DO 30 II=1,NG
      IF (NR(II).EQ.0) GO TO 30
      BINEQ2 = BINEQ2 + YR(II,JJ) * ALOG2(YR(II,JJ) * OBS/NR(II))
   30 CONTINUE
      BINEQ = BINEQ + Y(JJ) * BINEQ2
   31 CONTINUE
C
C     CALCULATE RS STATISTIC
C
      IF (TINEQ.GT.0.) GO TO 32
      PINEQ = 0.
      IF (BINEQ.EQ.0.) PINEQ = 100.0
      IXOUT = 1
      GO TO 33
   32 PINEQ = BINEQ / TINEQ * 100.0
C
C     EVALUATE NEW GROUP AGAINST PREVIOUS
C
   33 IF (PINEQ.GT.OLSTAT) GO TO 34
      IND = IND - 1
      MEMB1(I) = ISTORE
      PINEQ = OLSTAT
      IF (NG.GT.MINNG) GO TO 39
C
C     SUMS FOR OPTIMAL GROUPS
C
   34 DO 35 I2=1,NG
      NGRP(I2) = 0
      DO 35 J2=1,NVAR
```

```
   35 SUMX(I2,J2) = 0.
      DO 37 I2=1,IOBS
      DO 36 J2=1,NVAR
      LT = MEMB1(I2)
   36 SUMX(LT,J2) = SUMX(LT,J2) + DATA(I2,J2)
   37 NGRP(LT) = NGRP(LT) + 1
      DO 38 I2=1,NG
      DO 38 J2=1,NVAR
   38 SUMX(I2,J2) = SUMX(I2,J2) / NGRP(I2)
   39 OLSTAT = PINEQ
   40 CONTINUE
   41 CONTINUE
      IF (IND.EQ.0) ICNT = ICNT + 1
      IF (ICNT.LT.3) GO TO 24
      IF (INTMED.EQ.1) CALL RITE (MEMB1,DATA,NG,SUMX,TM,SD)
      IF (IXOUT.NE.1) WRITE(6,107) NG,PINEQ
      IF (IXOUT.EQ.1) WRITE(6,108) NG,PINEQ,TINEQ,BINEQ
      IF (NG.GT.MINNG.AND.PINEQ.LE.STATMX) GO TO 44
C
C     CALCULATE GROUPS MEANS AND S D S
C
      NGMAX = NG
      DO 42 I=1,IOBS
   42 MEMB2(I) = MEMB1(I)
      DO 43 I=1,NG
      DO 43 J=1,NVAR
   43 GMEAN(I,J) = SUMX(I,J)
      STATMX = PINEQ
   44 CONTINUE
      WRITE (6,109) NGMAX,STATMX
      CALL RITE (MEMB2,DATA,NGMAX,GMEAN,TM,SD)
  100 FORMAT (1H0,19HDATA IN PROPORTIONS)
  101 FORMAT (1H0,15HMATRIX TOTAL IS,F10.5)
  102 FORMAT (1H ,I5,10F10.2/1X,5F10.2)
  103 FORMAT (10F8.0)
  104 FORMAT (1H0,11HTOTAL MEANS//2(1X,10F10.6/))
  105 FORMAT (1H0,19HSTANDARD DEVIATIONS//2(1X,10F10.6/))
  106 FORMAT (2513)
  107 FORMAT (1H0,I5,8H CLASSES,F6.2,19H PER CENT EXPLAINED)
  108 FORMAT (1H0,I5,8H CLASSES,F6.2,9H PER CENT, F6.2,6H TOTAL,
     1F6.2,8H BETWEEN)
  109 FORMAT (1H0,22HOPTIMAL CLASSIFICATION,I5,8H CLASSES,F6.2
     1,18H PER CENT EXPLAINED)
      RETURN
      END
C
C     SUBROUTINE RITE
C
      SUBROUTINE RITE (MEMB2,DATA,NG,GMEAN,TM,SD)
      DIMENSION MEMB2(511),DATA(99,20),GMEAN(20,20)
      DIMENSION SD(20),TM(20)
      DIMENSION ZT(20),TT(20)
      DIMENSION SE(20)
      COMMON IOBS,NVAR,IOUT,ID,MAXNG,MINNG,SUB,IAGG,IPER,IMEM,IN,IX,SINC
     1,INTMED
C
C     WRITE GROUP MEMBERS
C
      DO 3 I=1,NG
      KK = 0
```



```
      WRITE (6,100) I
      DO 1 J=1,IOBS
      IF (MEMB2(J).EQ.I.AND.ID.EQ.1) WRITE(6,101) (DATA(J,K),K=1,NVAR)
      IF (MEMB2(J).EQ.I.AND.ID.NE.1) WRITE(6,102) J,
     1(DATA(J,K),K=1,NVAR)
      IF (MEMB2(J).EQ.I.AND.ID.NE.1) KK = KK+1
    1 CONTINUE
      WRITE (6,104) (GMEAN(I,J),J=1,NVAR)
      DIV = FLOAT(KK)
      DO 2 J=1,NVAR
      SE(J) = SD(J) / SQRT(DIV)
      TT(J) = GMEAN(I,J) - TM(J)
C
C     CALCULATE Z VALUES
C
    2 ZT(J) = TT(J) / SE(J)
      WRITE (6,105) (ZT(J),J=1,NVAR)
  100 FORMAT (1H0,5HCLASS,I4)
  101 FORMAT (1H ,20F6.3)
  102 FORMAT (1H ,11HOBSERVATION,I4,2(20F6.2/))
  103 FORMAT (1H0,7HZ TESTS//1X,20F6.2)
  104 FORMAT (1H0,11HGROUP MEANS/1X,20F6.3)
  105 FORMAT (1H0,7HZ TESTS//1X,20F6.2)
    3 CONTINUE
      RETURN
      END
      FINISH
```