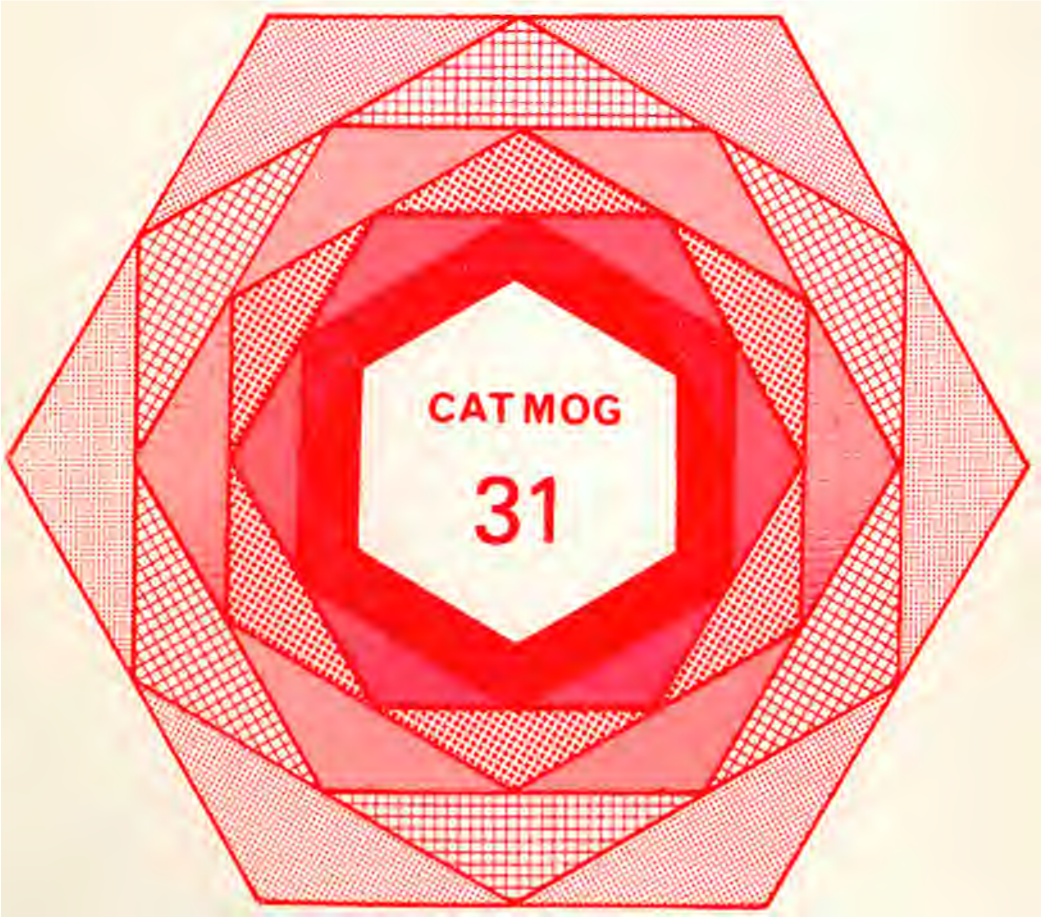
INFORMATION STATISTICS IN GEOGRAPHY

R. W. Thomas



ISBN 0 86094 090 X
ISSN 0306 — 6142

Printed in Greet Britain by Headey Brothers Ltd The Invicta Press Ashford Kent

CATMOG
(Concepts and Techniques in Modern Geography)

CATMOG has been created to fill a teaching need in the field of quantitative methods in undergraduate geography courses. These texts are admirable guides for the 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. An 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 analytical techniques - 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 - 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. Dixon & B. Leach
18. Questionnaires and interviews in geographical research - C. 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 - J.E. Killen
25. Directional statistics - G.L. Gaile & J.E. Burt

continued on inside back cover

CONCEPTS AND TECHNIQUES IN MODERN GEOGRAPHY No. 31

INFORMATION STATISTICS IN GEOGRAPHY

by

R.W. Thomas

(University of Manchester)

CONTENTS

Acknowledgements

Thanks are due to Irene Bristow for typing the manuscript and Graham Bowden for drawing the maps and diagrams.

I. INTRODUCTION

A recurring theme in geographical data analysis is how to isolate and describe properties of spatial distributions. In this context a spatial distribution is taken to be the different values of a variable over a set of regions. For instance, industrial geographers are often presented with the problem of describing regional variations in the distribution of employment in an industry, a problem which requires techniques capable of measuring whether the industry's workforce is excessively concentrated in a few localities or relatively evenly distributed throughout the country. Similarly, social geographers interested in residential segregation (Peach, 1975) need ways of describing the distribution of social and ethnic groups over the census tracts of a city. Traditionally such problems have been tackled using techniques known as Lorenz curves and segregation indices (Duncan and Duncan, 1955) to make the necessary descriptions. However, experience has shown that these methods produce results which are critically dependent on the size, shape and number of regions used in the analysis. This dependence makes it difficult to compare properties of variables distributed over different regional systems. In an effort to overcome at least some of the technical difficulties posed by the more traditional methods, a number of geographers have examined the potential of information statistics as an alternative method of analysing properties of spatial distributions. It is these methods and their simpler geographical applications that are the subject of this monograph.

The mathematics of information theory is closely linked with probability theory and this monograph has been written on the assumption that the reader is familiar with the basic terminology of probability theory such as measuring the probability of an event's occurrence on a scale between 0 and 1. Otherwise, the only background mathematics necessary to understand information theory is a knowledge of how to manipulate logarithms. Indeed, in Section II(ii), the reader will find a description of those uses of logarithms which are peculiar to information theory.

Information statistics originated as part of the development of information theory which is the scientific study of properties of communications systems such as speech, telephony, radio and television. The beginnings of this theory can be traced back to a paper by Hartley (1928) on the transmission of information; however, the major stimulus to its development and application was due to the results obtained by the mathematician Claude Shannon (1948). These results were concerned with the processing and transmission of messages from one place to another. The theory provides engineers with a set of rules which set limits upon the capability of a communication system to transmit given quantities of information. It enables measurement of such quantities as the amount of information a communications channel is capable of transmitting, or the number of different messages that can be formed from a set of coding symbols. In this sense the theory is concerned with quantities of information, but not with the meaning or value human beings might attach to certain messages and thoughts. Thus the theory is incapable of distinguishing between a trivial piece of gossip and, say, a diplomatic message declaring the outbreak of war. The problem of value is left entirely to the recipient of the message.

In a sense, it is just because this special usage of information relates to what could be said rather than what is said that the theory has found general applications outside the field of communications engineering. The term 'what could be said implies that the process of forming messages involves a certain degree of choice. Therefore, it is not surprising to find that the theoretical definition of information is closely bound up with more familiar notions involving chance and probability. It is these links with probability that give information theory its generality as a statistical method of data analysis. Indeed, any set of events whose individual occurrences can be expressed in terms of probabilities falls within the ambit of information theory. In the social sciences the potential of information theory as a method of data analysis was first recognized by psychologists such as Quastler (1955). Geographical interest in the theory was fostered by the work of the economist Henri Theil (1967, 1972) who demonstrated how information concepts could be used to analyse properties of both spatial and temporal data. However, in order to understand such methods it is first necessary to introduce some of the simpler mathematical ideas about probability and information.

II THE INFORMATION CONTENT OF A PROBABILITY DISTRIBUTION

(i) Probability and surprise

Shannon's measure for the entropy, uncertainty, or information content evoked by a probability distribution is fundamental to the understanding of any application of information theory. It is possible to derive the entropy formula in a number of different ways (see Chapman, 1977) but, to avoid confusion, we will concentrate on a simple axiomatic derivation based on the relationship between probability and surprise.

Suppose we use daily rainfall records collected at a meteorological station to obtain the following pair of probabilities: {pw .75, pd = .25} where p_w is the probability that it will rain tomorrow and pd is the probability that it will be dry. If tomorrow turns out to be a wet day, the news of this occurrence will not surprise you a great deal because you already know wet days have a high probability of occurrence. Conversely if tomorrow is a dry day, you will experience a far greater surprise because you know dry days occur relatively infrequently. Thus the degree of surprise evoked by the news that some event has indeed occurred is a function of that event's prior probability of occurrence. We are highly surprised by the occurrence of rare events with low probabilities, but only slightly bemused by the occurrence of events whose probability is close to one (certainty). In other words, the shock is greatest when the unexpected happens. This notion of surprise may be equated with the technical meaning of information. The larger the surpise, the greater the information content or value of the piece of news we have received.

To proceed further it is necessary to give a precise mathematical definition to our intuitive notions about the relationship between surprise and probability. Let the term $S(p_i)$ denote our surprise at the occurrence of some event i with the probability p. Since p is measured on a scale between 0 (impossibility) and 1 (certainty), surprise will be some function of the positive probabilities that occur within these limits. We will

specify three axioms, or rules, which our measure of surprise must satisfy, and then deduce the mathematical function which satisfies these rules.

Our first axiom states that the surprise we experience at the occurrence of a certain event (p_i= 1) is zero, which we write as

$$\text{if } p_i = 1, \text{ then } S(1) = 0.$$

For example, suppose you are told the sum of two throws of a six-sided dice was a number less than 13. You experience no surprise on receiving this news because the event is bound to occur.

The second axiom defines our previous notion that rare events evoke a greater surprise than common events. For example, if our probability distribution describes two possible outcomes, then the occurrence of the event with the lower probability will create the greater surprise. Symbolically, this assertion is written as:

$$\text{if } p_1 > p_2, \text{ then } S(p_1) < S(p_2)$$

This axiom requires us to measure surprise as a decreasing function of probability. Therefore, our function must measure the surprise evoked by a wet day (p_w=.75) as a quantity smaller than the surprise at the occurrence of a dry day (p_d=.25).

The third axiom is the most crucial for our understanding of the mathematical function which measures surprise. For the sake of illustration we must first assume that the occurrences of wet and dry days are independent of one another. That is, we are assuming that today's weather has no influence upon tomorrow's, or any other day's weather. When events are assumed to be independent, the probability that a specified sequence of these events occurs is obtained by multiplying together the probabilities for all the individual events which form the sequence. For instance, the prior probability of a wet day being followed by a dry day, which we will denote by p_{wd}, is obtained from the product

$$p_{wd} = p_w \times p_d = .75 \times .25 = .1875$$

Now suppose a wet day is indeed followed by a dry day. Clearly our surprise at this joint occurrence must be related to this joint probability and would be written as $S(p_{wd})$. Moreover, this joint surprise must itself be equal to the sum of the individual surprises we experienced when first a wet day occurred, $S(p_w)$, and the second surprise we experienced, $S(p_d)$, when a dry day followed. This statement implies that the mathematical function chosen to measure surprise must satisfy the condition

$$S(p_{wd}) = S(p_w) + S(p_d)$$

Notice that this axiom requires us to choose some function which relates the multiplication of individual probabilities on the left-hand side of the expression, since $S(p_{wd}) = S(p_w \times p_d)$, with the *addition* of individual probabilities on the right-hand side. In the arithmetic of the weather problem the condition becomes

$$S(.1875) = S(.75) + S(.25) \qquad (1)$$

It is the *logarithm* of the individual probabilities that is the function which satisfies this condition because the logarithm of the *product* of two numbers is equal to the *sum* of the logarithm of each number. Therefore, if we define $S(p_i) = \log p_i$ and substitute in (1), the condition is satisfied because

$$\log .1875 = \log .75 + \log .25$$
$$-0.7270 = -0.1249 + (-0.6021)$$

Although the definition of surprise as the logarithm of probability satisfies our third axiom, it does *not* satisfy the second axiom which asserted that surprise must be a decreasing function of probability. This failure occurs because the logarithm of a probability increases with the value of the probability. To make surprise a decreasing function of probability we simply define it as the reciprocal of probability, that is

$$S(p_i) = \log (1/p_i) \qquad (2)$$

This revised definition also satisfies the first axiom because a certain event with $p_i = 1$ has an associated surprise of $\log (1/1) = 0$. Thus our definition now satisfies all of our axioms and for the weather problem the individual surprises associated with the occurrence of a wet and a dry day are respectively

$$S(p_w=.75) = \log (1/.75)$$
$$= \log 1.3333 = 0.1249$$

and

$$S(p_d=.25) = \log (1/.25) = 0.6021.$$

(ii) <u>Some properties of logarithms</u>

At this juncture it is worth mentioning a few simple properties of logarithms which are employed in information theory. The first point concerns the writing of logarithms for numbers less than one in a negative form. For example, in the previous section $\log .75$ was written as $-.1249$ and not in the usual logarithmic table form of $\bar{1}.8751$, where the mantissa (.8751) is a positive quantity and the characteristic $(\bar{1})$ is negative. The negative form, which is the more convenient computationally, is obtained simply by subtracting the characteristic from the mantissa, that is

$$\bar{1}.8751 = .8751 - 1 = -.1249$$

In a similar vein it may be noted that our measure of surprise

$$S(p_i) = \log (1/p_i) \qquad (2)$$

may also be written in the form

$$S(p_i) = \log 1 - \log (p_i) ,$$

which simplifies to

$$S(p_i) = -\log (p_i)$$

because the logarithm of one is zero. The quantity $--\log(p_i)$ will always be

positive because the probability, p_i, falls between 0 and 1 giving a negative logarithm which is made positive by the minus sign.

In the monograph, for ease of comparison with standard logarithmic tables, all results are expressed to the base 10. However, in many texts on information theory logarithms to the base 2 are used in computation. The logarithm to the base 2 of any number may be calculated by multiplying its base 10 logarithm by the constant scaling factor

$$k = \log(10)/\log(2) = 3.3223$$

The choice of the base 2 to express results is made because the surprise, or information content, associated with the outcome of an event with prior probability $p = .5$ is measured as

$$S(.5) = k \log_{10}(1/.5)$$
$$= 3.3223 \times .3010 = 1.0000$$

This quantity is said to represent one unit or *bit* of information. Sometimes information theorists present their results in logarithms to the base e 2.7183, and in such circumstances the unit of information is termed the *nit* and occurs for an event where the prior probability of occurrence is $p = 1/2.7183 \quad 0.3678$. Finally, if logarithms to the base 10 are used the unit of information is termed the *decit* and corresponds to an event where $p = 1/10 = .10$.

(iii) <u>Shannon's Entropy</u>

Whereas surprise is a function of a single probability, Shannon's Entropy is a measure of a discrete probability distribution. Entropy is simply the average surprise a probability distribution will evoke. For set of n discrete probabilities, $\{p_i\}$, which by definition sum to one, Shannon's Entropy, H_n, is given by the expression

$$H_n = \sum_{i}^{n} p_i \log (1/p_i) \qquad (3)$$

Alternatively, if $-\log(pi)$ is used as the definition of surprise, Shannon's Entropy may be written in the form

$$H_n = -\sum_{i}^{n} p_i \log p_i \qquad (4)$$

For our weather problem, which is a discrete probability distribution composed of $n = 2$ events, ($i = 1$ = wet, $i = 2$ = dry), the entropy is calculated from formula (3) as

$$H_2 = p_1 \times \log(1/p_1) + p_2 \times \log (1/p_2)$$
$$= (.75 \times .1249) + (.25 \times .6021) = 0.2442.$$

A close examination of this calculation will illustrate the meaning of entropy as average surprise. If we observe the weather for some long period of time we know that a proportion of .75 of the days will be wet and a proportion of .25 days will be dry. -Therefore, we will receive a surprise of 0.1249 on .75 of all days and be surprised by an amount 0.6021 on .25 of all days. Thus, after a large number of days have passed, our average daily surprise will be the sum of the products of the individual surprises and their respective probabilities. On the average, the daily weather will surprise by an amount of 0.2442 and this quantity, the entropy is often termed the amount of uncertainty a set of probabilities are capable of evoking. Another interpretation of H_n is that its value is the average expected information value of a single message confirming the occurrence of one of the defined set of events.

(iv) Maximum entropy

Many practical applications of the entropy idea involve finding the values of a set of n probabilities which make the entropy function take on its maximum value. In other words we often need to know the conditions which create the greatest uncertainty.

For the sake of simplicity we will explain the idea of maximum entropy with reference to the two-event weather problem. For a two event problem the maximum entropy can be found graphically by plotting the value of the entropy function for different pairs of probabilities p_w and p_d. Moreover, because $n=2$, it follows that any value given to p_w automatically defines the value of p_d as $1 - p_w$. Therefore, the x-axis of the graph (see figure 1) is labelled with p_w increasing from 0 to 1, which implies that the corresponding value of p_d used to compute the entropy is $1 - p_w$. Notice that the entropy of our previous example plots on the graph as the point $p_w = .75$ ($p_d = .25$) and $H_2 = .2442$. The graph shows that for a totally arid climate where, $\{p_w = 0, p_d = 1\}$, the entropy is zero because every day is dry and the weather never surprises us. (The entropy of this arid climate is calculated as $H_2 = 0 \log(1/0) + 1 \log(1/1) = 0$, and the term $0 \log(1/0)$ which occurs in relation to the impossible event has a defined value of zero. The entropy gets larger as p_w increases up to a value of .5, but then declines symmetrically to a value of zero for the completely wet climate associated with $\{p_w = 1, p_d = 0\}$. Thus the climate creates the greatest uncertainty when the probabilities for wet and dry days are the same. This result makes sense because, if wet and dry days are equiprobable, we are maximally uncertain about the next day's weather.

We can generalise this result to obtain a simple formula for calculating the maximum entropy for any set of n discrete events. Maximum entropy occurs when the probabilities for all n events are equal, that is when all probabilities have the value

$$p_i = 1/n \qquad (5)$$

Substituting this definition of p_i in formula (3) gives the maximum entropy as

$$\text{max. } H_n = \sum_i^n \frac{1}{n} \log\left(\frac{1}{1/n}\right)$$

$$= \sum_i^n \left(\frac{1}{n}\right) \log n$$

$$= \log n. \qquad (6)$$

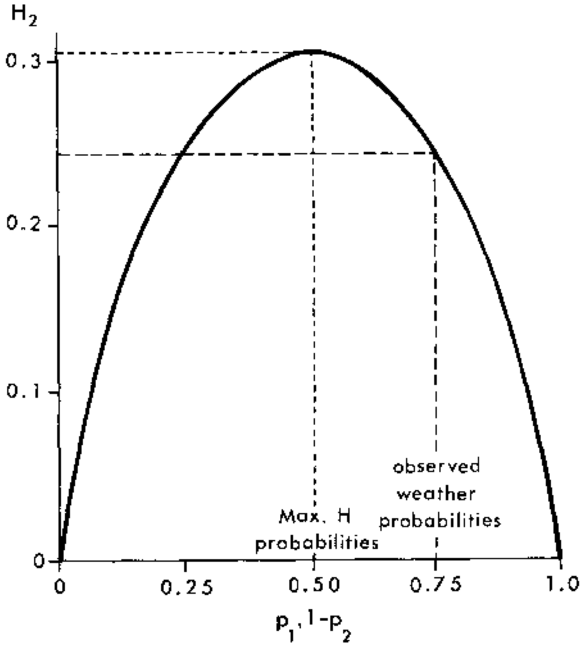


Figure 1. The entropy of all weather forecasts.

III ENTROPY AS A MEASURE OF DIVIDEDNESS

(i) Relative entropy and spatial dispersion

So far we have interpreted entropy purely within the context of information theory. Theil (1972), however, gives a much broader interpretation of the entropy idea which extends beyond the narrow realm of messages and even that of probability. He regards entropy as a general measure of dividedness, capable of representing the extent to which some total population is evenly distributed among its component parts. It is this interpretation that has been applied to a variety of problems in both geography and the social sciences in general. To take a simple example, suppose we want to measure the extent to which a school is racially integrated. If the schoolchildren are divided into two racial groups, black and white, we can measure the proportion of schoolchildren belonging to each group. The entropy of these proportions provides an appropriate measure of racial integration. Its value will be zero when only one racial group is represented in a school and will rise to a maximum of log ($n = 2$) when both groups are equally represented. This type of argument can be applied to any population which can be meaningfully divided amongst a set of parts. In geography, the method is often applied to the dividedness of a population over set of regions, and to illustrate the approach we will examine an analysis of the distribution of the assets of financial corporations in the United States made by Semple (1973).

The simplest application of the entropy idea is to measure either the amount of spatial dispersion or spatial concentration exhibited by a geographical variable whose individual observations, x_i, are the quantity of the variable found in each of n regions. For example, one of Semple's variables was the value of the assets of life insurance companies whose headquarters were located in each of the n 9 major census divisions (see fig. 2) of the United States for 1956. Before analysis can begin these observations (x_i) must be expressed as proportions (p_i) by dividing each value x_i by the sum of all x_i, that is

$$p_i = x_i / \sum_i^n x_i \tag{7}$$

The data are listed in Table 1.

Table 1 The proportions of life insurance company assets in U.S. census divisions, 1956. Source: R.K. Semple (1973)

Census division number, (i see fig. 2)	1	2	3	4	5	6	7	8	9
Assets *, x_i	19.1	50.0	7.1	1.4	1.1	1.3	2.4	.3	1.3
Proportion, p_i	.227	.595	.085	.017	.013	.016	.029	.003	.015

*Assets expressed in billions of dollars

Inspection of these proportions will reveal that life insurance company assets tend to be concentrated in the Northern (i = 1) and Eastern (i = 2) census divisions. To measure their degree of dividedness we first calculate their observed entropy from formula (3), that is

$$H_9 = [p_i \log (1/p_i)] + [p_2 \log (1/p_2)] + , \ldots, +[p_9 \log (1/p_9)]$$

$$= [.227 \log (1/.227)] + [.595 \log (1/.595)] + , \ldots, +[.015 \log (1/.015)]$$

$$= .5341.$$

This quantity lies on a scale which ranges from zero to $\log n$. The minimum value of zero will occur when the variable is present in only one of the regions and therefore one of the proportions will be equal to one and the remainder will all be equal to zero. This limit is equated with maximum spatial concentration of the variable. The maximum dispersion of a set of proportions occurs when each region contains the same amount of the variable, that is when all $p_i = 1/n$. In such circumstances the entropy of the proportions will be at a maximum and may be measured by $\log n$. The degree of dispersion exhibited by an observed set of proportions, $\{p_i\}$, may be measured on a scale between o and 1 by calculating the *relative entropy* of the proportions using the formula

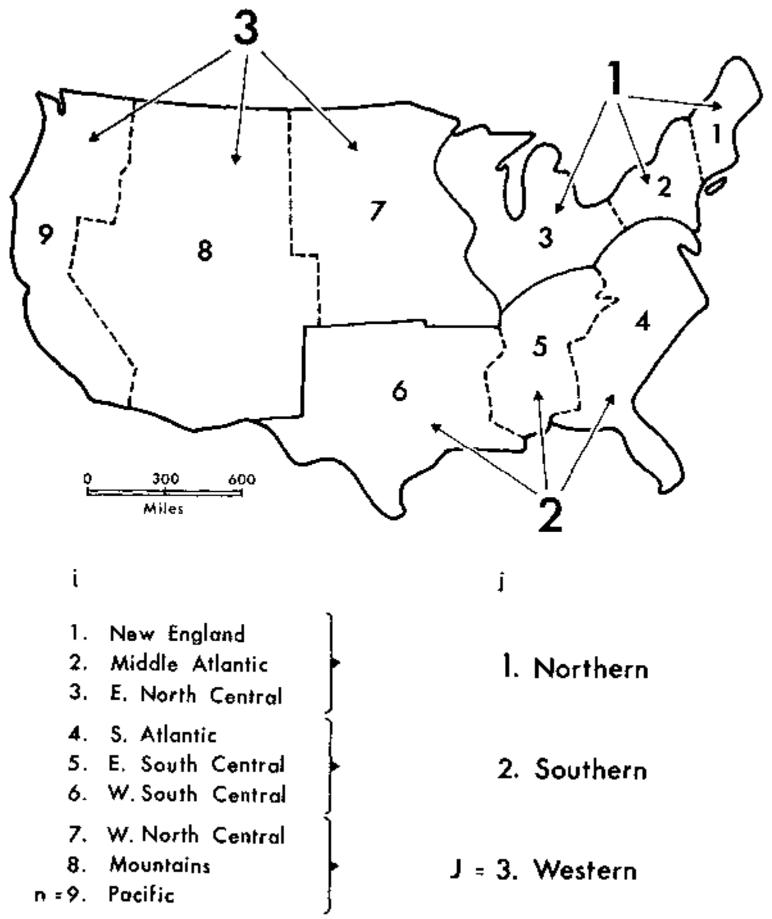


Figure 2. U.S.A., major census divisions and aggregated zones used in Semple's decomposition analysis.

$$\text{Rel. } H_n = \frac{\sum_i^n p_i \log (1/p_i)}{\log n} \tag{8}$$

which is the ratio between the observed entropy of the proportions and their maximum possible entropy. For the life insurance assets proportion formula (7) gives the relative entropy as

$$\text{Rel. } H_9 = \frac{.5341}{\log 9} = \frac{.5341}{.9541} = .5598,$$

which indicates that life insurance assets achieve a degree of dispersion approximately 56% of the maximum possible. Depending on the idea we wish to test, it is sometimes preferable to express our results on a scale where maximum spatial concentration coincides with the upper limit of one. To achieve this transformation we compute an index known in information theory as the *redundancy* which is simply the complement of the relative entropy, that is

$$\text{Redundancy}= 1 - \text{Rel. } H_n$$

Thus for our assets problem the degree of spatial concentration is measured as

$$\text{Redundancy}= 1-.5598 = .4402.$$

Spatial dispersion indices are descriptive measures used to compare differences in dispersion between different sets of proportions. Semple used information statistics to test the idea that the distribution of a nation's wealth would become more evenly shared among the nation's regions as the national economy grows and matures. This suggestion would lead us to expect the distribution of the assets of financial companies to become more dispersed during the studied period. Table 2 lists the value of the relative entropy of the proportions for three classes of financial company (life insurance, banking and utilities) for the years 1956 and 1971. It can be seen that for each of the three classes the relative entropy of the asset proportions showed a small increase during the study period indicating a slight tendency towards increased spatial dispersion. These results give some credance to Semple's ideas, although the changes in the relative entropies are small and are indicative only of minor modifications to the spatial distribution of assets.

Table 2: Dispersion analysis of U.S. financial corporations, 1956-71. Adapted from R.K. Semple (1973) p. 317

Class	Year	n	Observed Entropy	Maximum Entropy	Relative Entropy
Life	1956	9	.5341	.9541	.5598
Insurance	1971	9	.5515	.9541	.5780
Banking	1956	9	.5741	.9541	.6017
	1971	9	.5998	.9541	.6287
Utilities	1956	9	.5467	.9541	.8730
	1971	9	.5707	.9541	.5982

(ii) The entropy decomposition theorem

An advantage of relative entropy over the traditional indices of spatial dispersion is that its value is *invariant* with the value of n, the number of regions. By invariant we mean that an observed relative entropy of say .5c) is indicative of the same degree of dispersion irrespective of whether the observed proportions were derived from a system of n = 10 or n = 100 regions. In this sense relative entropy is said to be a dimensionless index and this property enables the index to be used to compare degrees of dispersion on different sized regional systems. However, it must be emphasized that this property does not mean that relative entropy is insensitive to the particular shapes and sizes of the regions used as a basis for collecting the observed proportions. For example, if the relative entropy of life insurance assets had been calculated from proportions derived from individual states instead of major census regions we would no doubt obtain a different result. The difference between these two relative entropies would be due solely to differences between the way in which the two regional systems partition the distribution of life insurance assets. This effect is a manifestation of the scale problem which has an unspecified influence on many forms of spatial analysis. Fortunately, a theoretical result known as the entropy decomposition theorem allows us to measure directly the way in which differences in degrees of diversification are attributable to different regional data collection systems. The use of information statistics to measure such scale and aggregation effects is generally termed *decomposition analysis.*

Table 3. Notation for entropy decompositions using Semple's 1956 life insurance proportions

Symbols		Values								
Subscripts	j	1			2			3		
	i	1	2	3	4	5	6		8	9
	$i\epsilon j$, n_j	$i\epsilon 1$		n_1	$i\epsilon 2$		n_2	$i\epsilon 3$		n_3
Proportions	p_j	.907			.046			.047		
	$p_j(i)$	.227	.595	.085	.017	.013	.016	.029	.003	.015

As a preliminary to decomposition analysis, regional systems must be delimited for the different scales of analysis. Figure 2 illustrates the two scales of analysis used by Semple to analyse the dispersions of financial company assets. The i = 1,2,..../, 9(n = 9) U.S. census divisions have each been assigned to one of i = i, 2, = 3) larger regions termed the Northern, Southern and Western regions respectively. The census divisions are termed the *sub-regional* scale and the groups of census divisions are termed the *regional scale.* Notice Semple assigned equal numbers of sub-regions to each region, and this design feature is a necessary condition of the description of decomposition analysis that follows.

The definition of a two-tier regional system requires us to introduce aggregation notation in order to identify proportions belonging to different scales (see Table 3). The proportion of life insurance assets located in each sub-region (i) are retermed $p_j(i)$. For Semple's data, $p_2(4) = .017$ refers to the proportion of life insurance company assets controlled from the S. Atlantic sub-region (i = 4) which is a part of the Southern region (j = 2). To calculate the proportion of the variable found in each region (j) we simply sum the sub-region proportions assigned to the region. Symbolically, we write this summation as

$$p_j = \sum_{i \varepsilon j}^{n_j} p_j(i) \qquad (9)$$

where, $i \varepsilon j$ denotes the value of i which is the first element (ε) of set j, and n_j is the value of i which forms the last element of set j (see Table 3). For example, the proportion of financial company assets controlled from the Southern region (p_2) is calculated from formula (9) as

$$p_2 = \sum_{i=4}^{6} p_2(i)$$

$$= p_2(4) + p_2(5) + p_2(6)$$

$$= .017 + .013 + .016 = .046.$$

Notice that this method of aggregating sub-regional proportions, $(p_j(i))$, ensures that the sum of the regional proportions is one, that is

$$\sum_{j}^{J} p_j = 1.$$

Decomposition analysis involves separating the observed entropy of all the sub-regional proportions into two components which measure the contribution of each regional scale of the observed degree of diversification. The entropy of all sub-regional proportions, $p_j(i)$ is given by

$$H_n = \sum_{j}^{J} \sum_{i \varepsilon j}^{n_j} p_j(i) \log [1/p_j(i)] \qquad (10)$$

and is referred to as the total observed entropy. For the life insurance assets formula (10) gives the result

$$H_9 = [p_1(1) \log (1/p_1(1)) + p_1(2) \log (1/p_1(2)) + p_1(3) \log (1/p_1(3))]$$

$$+[p_2(4) \log (1/p_2(4)) + p_2(5) \log (1/p_2(5)) + p_2(6) \log (1/p_2(6))]$$

$$+[p_3(7) \log (1/p_3(7)) + p_3(8) \log (1/p_3(8)) + p_3(9) \log (1/p_3(9))]$$

$$= [.227 \log (1/.227) + .595 \log (1/.595) + .085 \log (1/.085)]$$

$$+[.017 \log (1/.017) + .013 \log (1/.013) + .016 \log (1/.016)]$$

$$+[.029 \log (1/.029) + .003 \log (1/.003) + .015 \log (1/.015)]$$

$$= .5341.$$

Notice this is the same calculation we made for the observed entropy on p. 7 and here we have simply added the subscript j to denote that each sub-regional proportion is a member of one of the regional proportions.

The *entropy decomposition theorem* states that formula (10) can be re-written in terms of both the proportions $p_j(i)$ and p_j as

$$H_n = \sum_{j=1}^{J} p_j \log (1/p_j) + \sum_{j=1}^{J} \left[p_j \sum_{i \varepsilon j}^{n_i} \left(\frac{p_j(i)}{p_j}\right) \log \left(\frac{p_j}{p_j(i)}\right)\right]. \qquad (11)$$

The interested reader is referred to Theil (1972) for a proof of the result which depends on the additivity of information as defined by the third axiom of surprise (p. 5). This lengthy formula is composed of two expressions on either side of the addition sign. The expression on the left-hand side is termed the *between* region entropy and is denoted by H_J, while the expression on the right-hand side is termed the average entropy *within* regions and is denoted by $H_{n/J}$. This terminology allows the entropy decomposition to be written much more simply as

$$H_n = H_J + H_{n/J}. \qquad (12)$$

The meaning of these terms will become clear if we calculate their observed values. The between region entropy is simply the observed entropy of the regional proportions and, for the life insurance problem, is calculated as

$$H_J = \sum_{j=1}^{3} p_j \log (1/p_j)$$

$$= p_1 \log (1/p_1) + p_2 \log (1/p_2) + p_3 \log (1/p_3)$$

$$= .907 \log (1/.907) + .046 \log (1/.046) + .047 \log (1/.047)$$

$$= .1624$$

This quantity reflects the degree of dividedness that is observed at the regional scale.

The average within region entropy is calculated by expanding the term on the right-hand side of the addition sign in formula (11) to give

$$H_{9/3} = p_{(j=1)}\left[\left(\frac{p_1(1)}{p_1}\right)\log\left(\frac{p_1}{p_1(1)}\right) + \left(\frac{p_1(2)}{p_1}\right)\log\left(\frac{p_1}{p_1(2)}\right) + \left(\frac{p_1(3)}{p_1}\right)\log\left(\frac{p_1}{p_1(3)}\right)\right]$$

$$= p_2\left[\left(\frac{p_2(4)}{p_2}\right)\log\left(\frac{p_2}{p_2(4)}\right) + \left(\frac{p_2(5)}{p_2}\right)\log\left(\frac{p_2}{p_2(5)}\right) + \left(\frac{p_2(6)}{p_2}\right)\log\left(\frac{p_2}{p_2(6)}\right)\right]$$

$$= p_3\left[\left(\frac{p_3(7)}{p_3}\right)\log\left(\frac{p_3}{p_3(7)}\right) + \left(\frac{p_3(8)}{p_3}\right)\log\left(\frac{p_3}{p_3(8)}\right) + \left(\frac{p_3(9)}{p_3}\right)\log\left(\frac{p_3}{p_3(9)}\right)\right]$$

$$\begin{aligned} H_{9/3} &= .907\left[\frac{.227}{.907}\log\left(\frac{.907}{.227}\right) + \frac{.595}{.907}\log\left(\frac{.907}{.595}\right) + \frac{.085}{.907}\log\left(\frac{.907}{.085}\right)\right] \\ &+ .046\left[\frac{.017}{.046}\log\left(\frac{.046}{.017}\right) + \frac{.013}{.046}\log\left(\frac{.046}{.013}\right) + \frac{.016}{.046}\log\left(\frac{.046}{.016}\right)\right] \\ &+ .047\left[\frac{.029}{.047}\log\left(\frac{.047}{.029}\right) + \frac{.003}{.047}\log\left(\frac{.047}{.003}\right) + \frac{.015}{.047}\log\left(\frac{.047}{.015}\right)\right] \\ &= .907\ [.3669] + .046\ [.4744] + .047\ [.3641] \\ &\approx .3328 + .0218 + .0171 \\ &= .3717 \end{aligned}$$

Inspection of this expansion will help to clarify the meaning of the term *average entropy within regions.* Inside each pair of square brackets each region is treated as a separate data set comprised of its 3 sub-regions. The division of each sub-regional proportion by its regional proportion $(p_j(i)/p.)$, converts the sub-regional proportions into a distinct set of proportions which sum to one. For example, the Northern region (J=1) is treated as a separate data set with life insurance proportions

$$\left\{\frac{p_1^{(1)}}{p_1} = \frac{.227}{.907} = .250;\quad \frac{p_1^{(2)}}{p_1} = \frac{.595}{.907} = .656;\quad \frac{p_1^{(3)}}{p_1} = \frac{.085}{.907} = .094\right\}$$

and the reader can easily check these converted values sum to one. The calculations inside the square brackets give the entropy of the converted proportions within each of the regions. Thus the within region entropy of .03669 obtained for the Northern Region 0=1) is indicative of the degree of dividedness of the three life insurance proportions forming that region. Taken together, the three within region entropies (.3669, .4744, 3641) may be used to compare degrees of dividedness in each region. The *average* within region entropy of .3717 is then obtained as the sum of each within region entropy weighted by its degree of importance as measured by p_j. Notice that, although the dividedness of life insurance proportions is greatest in the Southern region where the within region entropy is .4744, this region only contributes .1)218 to the average within region entropy because the proportion of all life insurance assets controlled from the South is low $(P2=.046)$. Thus our final value of .3717 reflects the average dividedness of the proportions within the regions.

To check our arithmetic we can substitute our results in formula (12) to get

$$H_n = H_J + H_{n/J} \qquad (12)$$

$$.5342 = .1624 + .3717,$$

which shows that the decomposition theorem has been satisfied.

To obtain indices of dispersion for the various scales the condition under which the various entropies take on maximum values must be established. In fact the expressions in the decomposition (12) all take on a maximum value when all the proportions in each sub-region are equal, that is when all

$$p_j(i) = 1/n. \qquad (13)$$

When all $pj(i)$ are equal it follows that all pj must also be equal because each regional proportion is an aggregate of the same number (n/J) of sub-regional proportions. Therefore, maximum dispersion at the sub-regional scale occurs simultaneously with maximum dispersion at the regional scale, when, by definition, all

$$p_j = 1/J. \qquad (14)$$

From formulas (13) and (14) it is apparent that maximum dispersion of all proportions, $\{p_j(i)\}$, is measured by $\log n$ and the maximum dispersion between the regional proportion, $\{p_j\}$, is given by $\log J$. Finally, within any region j there are n/j proportions of the form $p_j(i)/p_j$ and when these proportions are equal the maximum dispersion within the region will be given by $\log (n/J)$. Since all regions are made up of the same number (n/J) of proportions, the maximum *average* dispersion within regions is also measured by $\log (n/J)$. Notice that taken together the entropies associated with maximum dispersion satisfy the decomposition theorem (12) because

$$\log n = \log J + \log (n/J),$$

and therefore

$$\max H_n = \max H_J + \max H_{n/J}.$$

Conversely, the opposite extreme of maximum spatial concentration occurs when one of the sub-regional proportions is equal to one and the remainder are all equal to zero. In this case all the terms in the decomposition (11) will also take on a value of zero.

These boundary conditions of the entropy decomposition allow us to construct two further relative entropies which measure the degree of dispersion observed at each regional scale on a scale from zero to one. The *degree of dispersion* exhibited by the observed set of regional proportion, $\{p_j\}$, is measured by the between region relative entropy given by

$$\text{rel. } H_J = H_J/\log J. \qquad (15)$$

For the life insurance proportions formula (15) gives the result

$$\begin{aligned} \text{rel. } H_3 &= .1624/\log 3 \\ &= .1624/.4771 \approx .3406, \end{aligned}$$

which indicates a high degree of spatial concentration at this scale. The average dispersion within regions exhibited by the sets of proportions, $\{p_j(i)/p_j\}$, is measured by the average within region relative entropy given by the ratio

$$\text{rel. } H_{n/J} = H_{n/J} \;/\; \log\,(n/J). \tag{16}$$

For the life insurance proportions formula (16) is evaluated as

$$\begin{aligned}\text{rel. } H_{9/3} &= .3717/\log\,(9/3)\\ &= .3717/.4771 = .7791.\end{aligned}$$

which is indicative of a tendency for life insurance assets to be dispersed within the regions.

Table 11.4 lists the results of Semple's decomposition analyses for the distribution of the assets of three types of financial company in 1956 and 1971. Notice that, unlike life insurance assets, both banking and utilities exhibit higher degrees of dispersion between regions than within region. The results illustrate some quite significant distributional changes between 1956 and 1971 which were not apparent in the initial aggregated analysis (Table 2). For example, the total relative entropy (rel. H_n) indicates little change in the distribution of utility assets between 1956 and 1971. However, this result disguises two counterbalancing changes which are revealed by the decomposition analysis. While the between region dispersion of utility assets increased quite markedly from .6344 to .7397, this change was cancelled out by an increased concentration of assets within the regions.

Table 4: Entropy decomposition analysis of U.S. financial corporations, 1956-71. Adapted from R.K. Semple 1973) p.317.

Class	Year	Relative Entropies		
		rel. H_J	rel. $H_{n/J}$	rel. H_n
Life Insurance	1956 1971	.3406 .3615	.7791 .7946	.5598 .5780
Banking	1956 1971	.6750 .7129	.5285 .5444	.6017 .6287
Utilities	1956 1971	.6344 .7397	.5117 .4567	.5730 .5982

(iii) Further decompositions

So far we have used the decomposition theorem to give detailed descriptions about the form of a spatial distribution at two regional scales. However, the theorem can easily be adapted to handle more complex partitions of single variables. One of the interesting geographical extensions is the case where the variable is partitioned over a set of regions and a set of categories. For instance, the value of the sales of an industry can be partitioned among various companies that form the industry and the regions, or markets where these companies sell their final product. Table 5 lists some data for an industry made up of $j = 1, m$ (=2) companies and $i = 1, n$ (=2) marketing regions. Therefore, the p_{ij} each measure the proportion of the industries total sales made by the j^{th} company in the i^{th} marketing region. The observed entropy of these n x m = k proportions is measured by

Table 5: Partition of an industry's sales by company and region

j → m

		1	2	$p_{i.}$
p_{ij} = i ↓ n	1	.2	.4	.6
	2	.1	.3	.4
$p_{.j}$		.3	.7	

$$H_k = \sum_i^n \sum_j^m p_{ij} \quad \log\,(1/p_{ij}), \tag{17}$$

which for the example proportion yields a value of H_4 = .555 out of a maximum possible dividedness of max H_4 = log 4 = .602

It is possible to decompose H_k in two ways. The first of these makes use of the terms p_j which are formed by summing the elements p_{ij} in each column of the table and measure the proportion of the industry's total sales made by the j^{th} company. The decomposition of H_k now takes the usual form

$$H_k = \sum_j^m p_{.j} \;\log\,(1/p_{.j}) + \sum_j^m p_{.j} \sum_i^n \frac{p_{ij}}{p_{.j}} \log\left(\frac{p_{.j}}{p_{ij}}\right) \tag{18}$$

$$= \quad H_m \; + \; H_{k/m}. \tag{19}$$

The term H_m is the between company entropy and measures the dividedness of the industry's sales among the m companies. Its value ranges from zero, which represents a monopoly industry where one company makes all sales, to a maximum given by max H_m = log(m) which occurs in a perfectly competitive industry where each company makes an equal proportion of the industry's total sales. The term $H_{k/m}$ is the average entropy of each company's sales within the n marketing regions. This term takes on a minimum value of zero when each company sells all its output to a single region, and reaches a maximum value of max $H_{k/m}$ = log (k/m) when all companies sell an equal proportion of their output to each region. The results of this decomposition for the example problem are listed in Table 6, and show that the degree of competition between companies for the industry's sales (rel. H_m = .884) is less than the average degree of competition between companies for sales within the market regions (rel. $H_{k/m}$ = .964).

Table 6: Decomposition of the industry's sales problem.

Source	Observed Entropy	Maximum Entropy	Relative Entropy
H_m	.265	.301	.884
$H_{k/m}$	.290	.301	.964
H_k	.555	.602	.923
H_n	.292	.301	.971
$H_{k/n}$	.264	.301	.876.
H_k	.555	.602	.923

The second decomposition of H_k is based on the terms $p_{i.}$, which are formed by summing the elements p_{ij} in each row of the table (see Table 5), and measure the proportion of the industry's total sales which are made in the i^{th} region. The decomposition is given by

$$H_k = \sum_i^n p_{i.} \log (1/p_{i.}) + \sum_i^n p_{i.} \sum_j^m \frac{p_{ij}}{p_{i.}} \log\left(\frac{p_{i.}}{p_{ij}}\right) \qquad (20)$$

$$= H_n + H_{k/n} \qquad (21)$$

The interpretation of this decomposition is more familiar. The term H_n is the between region entropy and measures the degree of dividenss of the industry's sales among the n regions on the usual scale from zero representing maximum spatial concentration to max H_n = log n representing maximum spatial dispersion. The term $H_{k/n}$ is the average entropy, or dividedness, of each region's sales among the m companies. Its value is zero when each region is supplied by a single company, and reaches a maximum of max.$H_{k/n}$ = log (k/n) when each region receives equal proportions of sales from each company. The results of this decomposition (Table 6) show that the degree of dispersion of the industry's sales between the regions (rel. H_n = .971) is greater than the dispersion of the regional sales among the companies (rel. $H_{n/k}$ = .876).

Clearly, many other geographical problems are suited to this form of decomposition analysis. Degrees of dispersion of social variables grouped by class and region, or political variables such as voting behaviour classed by party and constituency can easily be analysed within this decomposition framework. Moreover, it is not necessary to restrict the decomposition to one set of regions and one set of categories. Theil (1967, Ch. 8) presents an analysis of the distribution of car sales in the U.S.A. decomposed into two regional scales (major census division and state) and two industrial categories (company and make of car). These four-way decompositions are not described here because their explanation requires the use of rather lengthy and cumbersome notation. However, we may note that those more complex partitions are based on the same principle as the two-way decompositions and simply involve a further separation of the between and within category entropies to account for the effect of the additional classes.

The reader who is familiar with inferential statistics may have noticed the entropy decomposition analysis is similar in style to the classical analysis of variance technique. A method which involves separating the total variation in a set of sample observations distributed over a set of regions into a component which measures variation that occurs within the regions and a second component that measures variation that occurs between the regions. By testing the ratio of the between to within components for statistical significance, analysis of variance is capable of testing hypotheses concerning the significance of the differences between the regional means. For example, the method can be used to establish whether differences between mean wheat yields on various soil types are significantly different from one another or merely due to chance variations in yield. Although entropy decomposition analysis does not allow for formal statistical testing in the manner of analysis of variance, it does give a descriptive evaluation of the relative importance of the between and within region effects.

IV COMPARING TWO DISTRIBUTIONS

So far our use of information statistics has been confined to the analysis of a single variable divided into a set of proportions, {pi}. This restriction is imposed by Shannon's entropy which is capable of measuring only the degree of correspondence between the observed proportions and the idealised distribution of equal proportions. However, in many cases it is more realistic to compare the observed distribution with a second variable, distributed over the same set of regions. For example, to assess a country's population distribution we might decide to compare the proportion of the total population in each county with the proportion of the country's national area occupied by each county. In this case an even population distribution would occur when each county's population proportion was equal to its area proportion. This problem, which involves two variables - population and area, cannot be tackled using entropy statistics. Instead, the description requires the use of an information statistic devised by Kullback (1959) which is variously termed information gain, expected information or directed divergence.

(i) Information gain : theoretical derivation

The index of information gain will be derived using our previous definition for the surprise at the outcome of an event given by s(p) = log (1/p), where p is the prior probability of occurrence. Recall that the function s(p) also measures the information contained in the message which tells us of the event's occurrence. This information content is large when p is small because we have received news of a rare event. Conversely, if p is close to one the information content of the message tends to zero because news about a near certain event contains little new information. Now, suppose we are told, not that an event has occurred, but that the event's *prior probability* of occurrence, 0, has changed to some new *posterior probability* whose value is q. What is the information content of such a message? Clearly the problem requires us to find a function to define a quantity $s(q:p)$ which we will use to measure the information content of the message which transforms the probability p into some new value, q.

To obtain a suitable definition for $S(q:p)$ it is necessary to proceed under the assumption that the event ultimately does occur. The starting point for our analysis is the prior probability p, and the end point is the news that the event has occurred. Consider two different routes which both begin and end in this way. Route one consists of two messages; the first message informs us that the prior probability has changed to some new value q. and the second message tells us the event occurred with the revised probability q. This first route may be illustrated by a weather forecaster who changes his mind. He usually assigns the occurrence of a dry day the probability $P = .25$ which is based on their past frequency, however, because of prevailing anticyclonic conditions he decides to revise tomorrow's probability for dry weather to $q = .75$. His suspicions are confirmed when tomorrow turns out to be dry. To the recipient of these messages the information content of this route is made up of two components namely the information content of the initial message transforming p to q denoted by $s(q:p) = S(.75: .25)$, plus the information content of the second message confirming the occurrence of a dry day and denoted by $S(q) = s(.75)$. The second route consists of a single message stating that the event occurred with the initial probability, p. In this instance the recipient does not hear the revised forecast, and, therefore, the information content of the dry day's occurrence is measured by $s(p) = s(.25)$.

It should be apparent that the information content of both these routes is the same because they end with the same piece of information, the occurrence of a dry day. Symbolically, we can write route one's equality with route two as

$$\underset{\text{Route 1}}{S(q:p) + S(q)} = \underset{\text{Route 2}}{S(p)\cdot} \tag{22}$$

For the weather forecasting probabilities this assertion reads

$$S(.75 : .25) + S(.75) = S(.25).$$

In expression (22) we already know the values of the two terms s(p) and $S(q)$ which measure the information content of an event occurring with a single probability. They are defined in the usual way as

$$S(p) = \log (1/p) \text{ and } S(q) = \log (1/q).$$

This knowledge enables us to rearrange expression (22) so that the unknown term s(q:p) is defined by the two known terms s(p) and s(1) that is

$$\begin{aligned} S(q:p) &= S(p) - S(q) \\ &= \log (1/p) - \log (1/q) \end{aligned} \tag{23}$$

Thus the information content of the message which transformed the probability of a dry day from $p = .25$ to $q = .75$ is measured as

$$\begin{aligned} S(.75 : .25) &= \log (1/.25) - \log (1/.75) \\ &= \log (4) - \log (1.3333) \\ &= .6021 - .1249 = .4772. \end{aligned}$$

A more economical definition of $S(q:p)$ is obtained if we use the simple relationship which states that the result of subtracting the logarithms of two numbers is equal to the logarithm of the number obtained from the division of the first number by the second number, that is $\log (a) - \log (b) = \log (a/b)$. This relationship enables formula (23) to be written as

$$\begin{aligned} S(q:p) &= \log\left(\frac{1/p}{1/q}\right) \\ &= \log (q/p) \end{aligned} \tag{24}$$

and allows the information content of the weather forecast transformation to be obtained more quickly as

$$\begin{aligned} S(.75:.25) &= \log (.75/.25) \\ &= \log (3) = .4772 \end{aligned}$$

Notice that in this example the information content of the transformation is a positive quantity because the likelihood of a dry day has been increased from $p = .25$ to $q = .75$. Indeed for all cases were $p<q$ the information gained will be positive. Conversely, when $p>q$ the information content of the transformation will be negative because the event is less likely to occur. Finally for the case where $p = q$ the value of $\log(q/p)$ is zero indicating that there is no information content in an unchanged situation.

(ii) Expected information gain

We can extent this argument to obtain a measure of the expected information content of a set of messages which transform the values of a prior probability distribution, {pi}, into a new set of values, $\{qi\}$. To exemplify the derivation of expected or average information gain we will use the simple two event weather forecasting problem. When the dry day probability was transformed from $p_2 = .25$ to $q_2 = .75$, it follows that the wet day probability was also revised from $p_1 = .75$ to $q_1 = .25$ because in a two event problem $p_1 = 1 - p_2$. Therefore, if the day turned out to be wet, the information content of the wet probability transformation would be measured as

$$\begin{aligned} S(.25 : .75) &= \log (.25/.75) \\ &= \log (0.3333) = -.5182 \end{aligned}$$

To derive $s(qi:pi) = -.5182$ and $S(q2:p2) \quad .4772$ it was necessary to assume that the events they describe ultimately occurred. Consequently, we do not know which transformation is appropriate until the day's weather is known. Nevertheless, we do know that the information measured by $\log(q_1/p_1)$ is likely to be received with probability q_1, while the information measured by $\log (q_2/p_2)$ is likely to be received with probability q_2. Hence the expected, or average, information content of the messages which transformed the prior probabilities, $\{p_i\}$, to the posterior probabilities $\{q_i\}$ can be expressed as the sum of each individual information transformation weighted by its posterior probability, that is

$$I(q:p) = q_1 \log (q_1/p_1) + q_2 \log(q_2/p_2) \qquad (25)$$

Thus, on the average, the information content of the weather forecast transformations will be

$$\begin{aligned} I(q:p) &= .25 \log(.25/.75) + .75 \log (.75/.25) \\ &= -.1296 + .3579 = .2283. \end{aligned}$$

Clearly the expected information gain defined by formula (25) need not be restricted to the two event case, and the generalised formula for the expected information gained from a set of n messages which transform a set of n prior probabilities, $\{p_i\}$, into a set of n posterior probabilities, $\{q_i\}$, is given by

$$I(q:p) = \sum_i^n q_i \log (q_i/p_i) \qquad (26)$$

The most important property of formula (26) is that its value is *always positive* despite the fact that the individual terms, qi log(qi/pi) may be either positive, when q_i is greater than p_i, or negative, when q_i is less than its original value p. To prove that the expected information gain is always positive is beyond the scope of this monograph, however, the interested reader is referred to Theil (1972, p.59) who proves that the sum of the positive terms, qi log(qi/pi), will always exceed the sum of the negative terms. The one exception to this rule is the case where each prior probability is equal to its transformed posterior probability for all n events, that is when $p_i = q_i$ for $i = 1, 2, \ldots\ldots, n$. In such a circumstance the expected information gain is zero because all $q_i/p_i = 1$ and therefore the term $\log(q_i/p_i) = \log (1)$ is by definition always equal to zero. Thus this case, where the priors remain unchanged by the transformation, represents the lower limiting case of zero expected information gain. Unfortunately, the expected information gain defined by formula (26) has no finite upper limit. The absence of such a limit is due to the case where the prior probability of some event is specified as zero and the transformation raises this estimate to a positive probability q. When an impossible event is deemed possible by the transformation the prior probability of zero has been increased by a factor of infinity $(q_i/0 = \infty)$ and the term $\log \infty = \infty$ reflects the fact that we are infinitely surprised by this message. Similarly, the expected information gain is also infinite if any pi is transformed from zero to some positive probability.

Taken together these boundary conditions imply that expected information gain is a general measure of the average degree of difference between two sets of n probabilities. The index takes on a value of zero when the two distributions are identical and becomes increasingly large and positive as the differences become more pronounced.

(iii) Expected information gain as an index of spatial concentration

In essence we have derived expected information gain as a measure of the goodness-of-fit between a prior and a posterior distribution. The poorer fit the larger is the value of the index. To convert expected information gain into an index of spatial concentration, we again work with proportions distributed over n regions rather than probabilities. For such applications it is usual to give the prior proportions values which correspond to the most appropriate definition of maximal spatial dispersion over the n regions and to define the posterior proportions as the observed proportion of the variable located within each region. Given these definitions, expected information gain will be zero when each posterior proportion is equal to its corresponding prior proportion and is indicative of maximum spatial dispersion. Expected information gain becomes increasingly large and positive as the degree of spatial concentration exhibited by the posterior proportions increases. Unfortunately, there is no general formula for the maximum value of expected information gain because this maximum is specific to the problem in hand and depends both on the values given to the prior proportions and the number of regions, n. All we can say about the upper limit is that, as long as all prior proportions are greater than zero, expected information gain will tend to some finite maximum value as one of the posterior proportions tends to a value of one. Thus, under the stated conditions, the maximum coincides with maximum spatial concentration when the variable tends to be located in a single region.

The absence of a general maximum for expected information gain means that the comparative value of the index is more restricted than relative entropy. As a rule of thumb, it is only sensible to use expected information gain to compare degrees of spatial concentration either between different variables (α^1, α^2, etc.) and the same priors, or between the same variable when the posterior proportions are measured at different points in time. Such comparisons are valid because the priors remain constant throughout the analysis. A second point of difference between relative entropy and expected information gain is that the former measures increasing spatial dispersion whereas, the, latter measures increasing spatial concentration.

(iv) Voting in Liverpool (1974)

To illustrate the use of expected information gain as a measure of spatial concentration, we present an analysis of voting patterns in Liverpool during the October 1974 General Election. The British electoral system, with its series of winner takes all constituency elections, does not necessarily guarantee that the number of Parliamentary seats won by each political party is in accordance with their share of the total national vote. Indeed, in the General Election of February, 1974 the Labour Party won the greatest number of Parliamentary seats while the Conservative Party polled the greatest number of votes. Such inequities are due to differences in the size of constituency electorates and differences in the geographical distribution of the votes for each party. In general a party whose support is fairly evenly dispersed over all constituencies will tend to win relatively more seats than a party whose vote is concentrated in a few constituencies. The following analysis attempts to identify these inequalities.

The results of the eleven constituency (fig. 3a) elections in Liverpool are listed in Table 7. The voting left Labour with nine Liverpool members and the Conservatives with two. To understand how voting behaviour in Liverpool differed from U.K. voting tendencies we will derive the sets of prior proportions, $\{p_{ij}\}$, $\{p_{i.}\}$ and $\{p_{.j}\}$ from two simplifying assumptions (see Table 8). The values $p_{.j}$ are the proportion of the U.K. electorate which voted for each of the $j = 1, 2, \ldots\ldots, m = (4)$ parties in October 1974. Thus $p_{.1} = .261$ of all electors voted Conservative while $p_{.2} = .286$ of all electors voted Labour, etc. The row totals, $\{p_{i.}\}$, are simply the proportion of the total Liverpool electorate (635,083) found in each constituency. For example, the proportion of Liverpool electors found in Crosby $(i = 1)$ is obtained from Table 7 as

$$p_{1.} = 78{,}583/635083 = .124$$

The individual priors, p_{ij}, for each voting category (j) in each constituency (i) are derived from the assumption that the constituency electors will vote in concordance with the U.K. voting proportions, $p_{.j}$. The values, $\{p_{ij}\}$, listed in Table 8 are obtained, therefore, from the proportional relationship.

$$p_{ij} = p_{i.} \times p_{.j} \qquad (27)$$

For instance, if Liverpool votes according to national trends, the expected proportion of Conservative voters in Crosby is obtained from (27) as

$$p_{11} = p_{1.} \times p_{.1} = .124 \times .261 = .032.$$

The posterior proportions $\{q_{ij}\}$, $\{q_{i.}\}$, and $\{q_{.j}\}$, - see Table 9 -, are all obtained directly from the voting data. The values $\{q_{ij}\}$ are the proportion of the total Liverpool electorate (635083) voting for party j in the ith constituency. Thus the proportion of Liverpool electors voting Labour $(j = 2)$ in Kirkdale $(i = 4)$ is obtained from Table 7 as

$$q_{42} = 17{,}686/635{,}083 = .028$$

Similarly, the values $\{q_{.j}\}$ are the proportion of Liverpool electors who voted for the jth party. For example, $q_{.3} = 54{,}025/635{,}083 = .085$ is the proportion of all Liverpool electors who voted Liberal in October 1974. Lastly, the set $\{q_{i.}\}$ is the proportion of the Liverpool electorate found in the ith constituency. Notice that in this problem the values $\{p_{i.}\}$ and $\{q_{i.}\}$ are identical, however in general terms this equality is not a necessary condition of the analysis.

Our voting proportions each have two subscripts $(i$ and $j)$ and therefore, the total information gained in the transformation of $\{p_{ij}\}$ into $\{q_{ij}\}$ is measured as

$$I(q:p) = \sum_{i}^{n} \sum_{j}^{m} q_{ij} \log(q_{ij}/p_{ij}) \qquad (20)$$

The reader may wish to check that for the Liverpool voting problem formula (28) takes on a value of .020 (using logarithms to the base 10). This result measures the difference between expected national average voting proportions for parties in Liverpool constituencies and the observed party voting proportions in those constituencies. Because expected information gain has no general maximum value this result is only of passing interest. However, by decomposing this result into between and within category components a more detailed interpretation may be given.

Table 7. Election results in Liverpool, February (1974)

		$j \rightarrow$ 1	2	3	m 4	
	Party	Con	Lab	Lib	No vote*	Total Electorate
	Constituency					
$i \downarrow$	1. Crosby	29,764	17,589	10,429	20,801	78,583
	2. Bootle	10,743	27,633	4,266	21,571	64,213
	3. Scotland/ Exchange	2,234	15,154	944	16,804	35,136
	4. Kirkdale	8,305	17,686	2,908	16,533	45,332
	5. Walton	10,706	20,568	4,221	16,456	51,951
	6. N. Derby	11,445	23,964	4,215	19,243	58,867
	7. Edge Hill	5,208	13,023	6,852	15,879	40,962
	8. Toxteth	8,062	15,312	3,176	19,324	45,874
	9. Wavertree	18,971	16,216	6,193	18,323	59,703
	10. Garston	24,557	27,857	5,865	22,723	81,002
	11. Huyton	15,517	31,750	4,956	21,237	73,460
	Totals	145,412	226,752	54,025	208,894	635,083

* Category includes votes for small parties.

The reader may have noticed that the Liverpool voting problem is similar in structure to the company sales problem described in Section II (iv). Indeed, the only difference between the two problems is that in formula (28) terms of the form $q_{ij} \log(q_{ij}/p_{ij})$ have replaced the terms $p_{ij} \log(1/p_{ij})$ in formula (17). Moreover, by making the same replacement of terms in formula (18) we obtain the following decomposition of formula (28)

$$I_k(q:p) = \sum_{j}^{m} q_{.j} \log(q_{.j}/p_{.j}) + \sum_{j}^{m} q_{.j} \sum_{i}^{n} \frac{q_{ij}}{q_{.j}} \log\left(\frac{q_{ij}/q_{.j}}{p_{ij}/p_{.j}}\right) \qquad (29)$$

where $k = n \times m$. The term on the left hand side of the addition sign is the between category (party) information gain while the term on the right-hand side is the average information gain for categories (parties) within the n constituencies (regions). As usual formula (29) can be written more

Table 8. Prior proportions (pig) for the Liverpool voting problem.

Constituency	$j \rightarrow$ 1 Con	2 Lab	3 Lib	m 4 No Vote	$p_{i.}$
i 1	.032	.035	.017	.040	.124
2	.026	.029	.013	.032	.101
3	.014	.016	.007	.018	.055
4	.019	.020	.009	.023	.071
5	.021	.023	.011	.026	.082
6	.024	.027	.012	.030	.093
7	.017	.018	.009	.020	.064
8	.019	.021	.010	.023	.072
9	.025	.027	.013	.030	.094
10	.033	.037	.017	.041	.128
n 11	.030	.033	.015	.037	.116
$p_{.j}$	.261	.286	.133	.320	1.000

economically in the notation

$$I_k (q:p) = I_m (q:p) + I_{k/m}(q:p). \qquad (30)$$

Calculations of the terms in formula (30) for the Liverpool voting data produces the decomposition

$$+.020 = .009 + .011,$$

which shows that the within party information gain is slightly more important than the between party effect.

To gain a more detailed understanding of these results it is necessary to examine the individual terms which sum to make the two major components of the decomposition. The between party information gain of $I_m(q:p)$ is a measure of the difference between the proportion of votes cast for political parties in all Liverpool constituencies and the proportion of

Table 9. Posterior proportions (q„) in the Liverpool voting problem

Constituency	$j \rightarrow$ 1 Con	2 Lab	3 Lib	m 4 No Vote	$q_{i.}$
i 1	.047	.028	.016	.033	.124
2	.017	.043	.007	.034	.101
3	.003	.024	.002	.026	.055
4	.013	.028	.004	.026	.071
5	.017	.032	.007	.026	.082
6	.018	.038	.007	.030	.093
7	.008	.020	.011	.025	.064
8	.013	.024	.055	.030	.082
9	.030	.026	.009	.029	.094
10	.039	.044	.009	.036	.128
n 11	.024	.050	.008	.034	.116
$q_{.j}$	.229	.357	.085	.329	1.000

votes cast for those parties in the U.K. as a whole. Inspection of formula (291 will show that this between oartv effect is the sum of $m(=4)$ terms of the form $q_{.j} \log(q_{.j}/p_{.j})$ which, in this instance, give the result

$$I_m(q:p) = -.013+.035-.017+.004 = +.009.$$

These individual terms measure the extent to which the proportion voting for the jth party in Liverpool was either above or below the national voting proportion for that party. Thus, during the October, 1974 election, relatively more people voted Labour (+.035) in Liverpool than in the country at large, while relatively fewer people voted Conservative (-.013) or Liberal (-.017). The proportion who did not vote in Liverpool was close to the national proportion not voting (+.004).

The average within party information gain of $I_{k/m}(q:p) = .011$ may also be interpreted by reference to its component parts. Inspection of formula (29) will reveal that this result is composed of $m = 4$ main terms of the

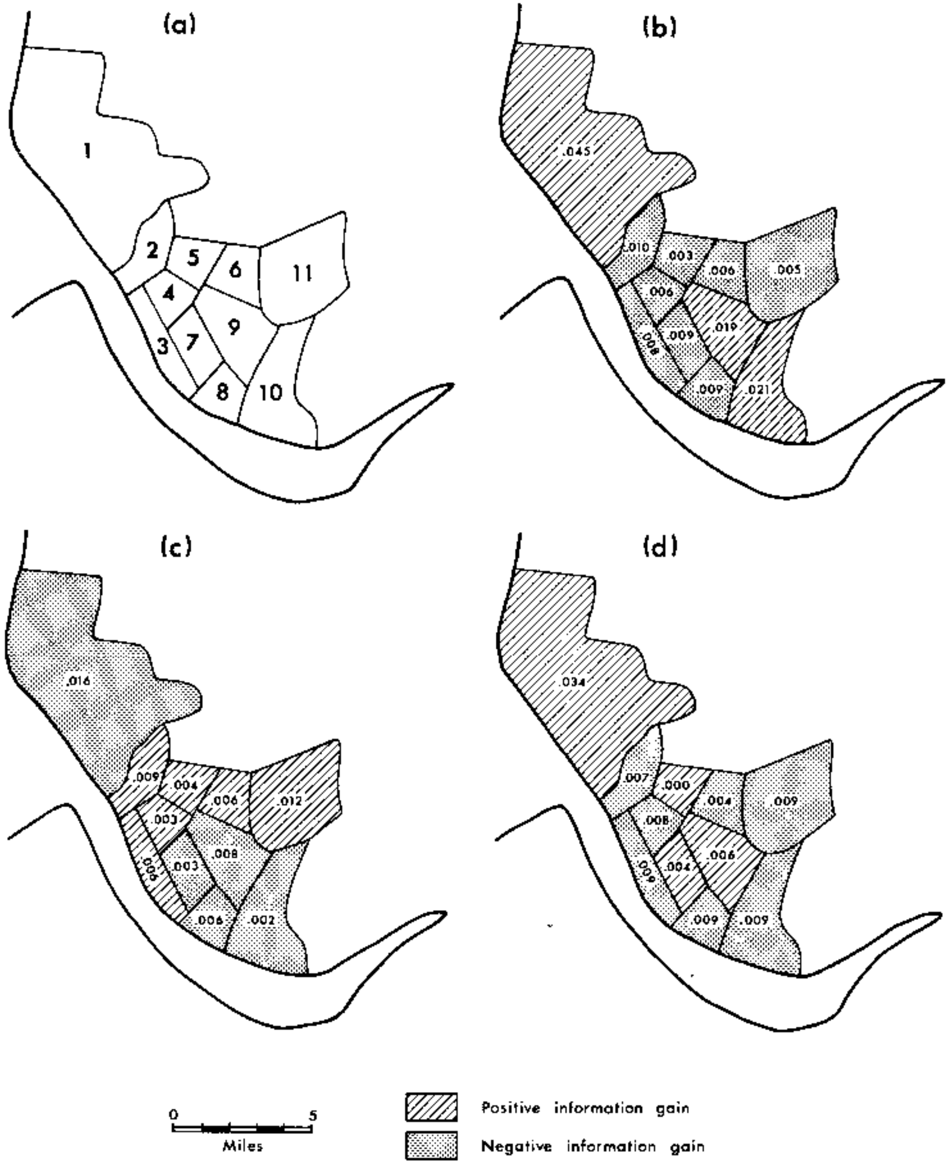


Figure 3. Voting in Liverpool (October, 1974). (a) Constituency numbers (i). (b) Information gain for Conservative voters. $\{(q_{i1}/q_{.1})\ \log\ [(q_{i1}/q_{.1})/p_{i1}/p_{.1})]\}$. (c) Information gain for Labour voters, $\{(q_{i2}/q_{.2})\ \log\ [(q_{i2}/q_{.2})/p_{i2}/p_{.2})]\}$. (d) Information gain for Liberal voters, $\{(q_{i3}/q_{.3})\ \log\ [(q_{i3}/q_{.3})/(p_{i3}/p_{.3}]\}$.

form $q_{.j} \sum_{i}^{n} \frac{q_{ij}}{q_{.j}} \log\left(\frac{q_{ij}/q_{.j}}{p_{ij}/p_{.j}}\right)$ each measuring the relative concentration of the vote for the jth party. The terms in the summation over the n = 11 constituencies each measure the information difference between the observed proportion of voters for jth party in the ith constituency, $q_{ij}/q_{.j}$, and the expected, national proportion for that party and constituency, $p_{ij}/p_{.j}$. For example, this information difference for Conservative $(j = 1)$ voting in Crosby $(i = 1)$ is calculated as

$$\frac{q_{11}}{q_{.1}} \log\left[\frac{q_{11}/q_{.1}}{p_{11}/p_{.1}}\right] = \frac{.047}{.229} \log\left[\frac{.047/.229}{.032/.261}\right] = +.045,$$

and is a measure of the extent to which the proportion of Conservative votes in Crosby is greater than the national proportion of Conservative votes. These terms may be positive or negative depending on whether the constituency proportion is above or below the national proportion. Mapping these terms for all constituencies gives an indication of regional variations in the proportion of the vote for each party. The map of these terms for the Conservative party (fig. 3b) shows a tendency for Conservative votes to be over-represented in suburban constituencies at the expense of inner city constituencies. The pattern is reversed for the Labour differences (fig. 3c) with over-representation in the inner city constituencies at the expense of suburban constituencies, while the pattern of Liberal differences (fig. 3d) is more haphazard.

The average within party information gain for this problem may now be represented in the form

$$\begin{aligned} I_{k/m} &= .229\ [.029] + .357\ [.005] + .085\ [.025] + .329\ [.002] \\ &= .006 + .002 + .002 + .001 = .011. \end{aligned}$$

The terms inside the square brackets are the sum of the individual constituency differences for each party and measure the local concentration of the vote for each party. Thus the concentration of the Conservative (.029) and Liberal (.025) votes is relatively high, while the concentration of Labour (.0o5) and no voting (.002) is relatively low. These last results help to explain the unrepresentative assignment of Parliamentary seats in Liverpool. In the Election the Labour party won 9 seats and the Conservatives 2 seats, whereas under a system of proportional representation Labour would have won 6 seats, the Conservatives 4 seats and the Liberals 1 seat. Such a malapportionment of seats occurs partly because the Conservatives vote is concentrated in relatively few constituencies. Thus, although this concentration allowed the Conservatives to win a couple of seats with large majorities, the majority of seats were lost to the more evenly distributed voting power of the Labour party.

The importance of the concentration of Conservative votes is further reinforced when the measures of party voting concentrations are weighted by the observed voting proportions, $q_{.j}$, to give the final four (m) terms which make up the average information gain within parties. Notice from the calculation above that the weighted Conservative concentration of .229 × .029 = .006 makes up over half the average concentration of all parties given by $I_{k/m}(q:p)$ = .011.

(v) Models of contingency table data

So far our use of information statistics has been purely descriptive, that is, we have used the various statistics to measure the departure of observed proportions from some idealised distribution. However, when observed data is representable in the form of a contingency table, such as the posterior proportions listed in Table 9, a more sophisticated form of analysis may be employed. This method is generally termed categorical data analysis and the usefulness to the geographer of this approach has recently been expounded by Wrigley (1976, 1979), Upton and Fingleton (1979, 1980) and Stapleton (1980). A recent text by Gokhale and Kullback (1978) develops methods of categorical data analysis which make specific use of information theory, and what follows is a brief outline of the logic of these methods.

In their book Gokhale & Kullback analyse a contingency table recording the occurrence of leukemia mortality among survivors of A-bomb explosions in Japan who were exposed to different levels of radiation. A simplified notational format for this data is listed in Table 10, where q_{ij} denotes the probability of the presence (i = 1) or absence (i = 2) of leukemia mortality among survivors exposed to different levels of radiation (j). The aim of categorical analysis is to formulate a model to predict a corresponding set of theoretical probabilities, $\{p_{ij}\}$, based solely upon the information contained in the contingency table. The most usual way to solve this problem is to construct a *log-linear model* which predicts the logarithm of the cell probability (p_{ij}) in terms of some linear combination of the marginal probabilities $\{q_{i.}\}$ and $\{q_{.j}\}$ and/or the observed probabilities, $\{q_{ij}\}$. For example, the most general form of the log-linear model for a two-way (variable) contingency table is given by

$$\log_e (p_{ij}) = u + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \qquad (31)$$

where u is a scalar roughly related to the average cell probability, λ_i^A is a parameter measuring the contribution of variable A (leukemia) to the predicted cell probability, λ_j^B measures to contribution of variable B (radiation), while λ_{ij}^{AB} refer to the joint (interaction) effects of categories belonging to each variable on the cell probability. The values given to these parameters are estimated from the observed data in the contingency table, however, space prevents a full discussion of the various estimation procedures and the interested reader is referred to Upton and Fingleton (1979) for an introduction to this topic.

The model described by equation (31) is termed a *saturated* log-linear model because it contains a parameter to describe every element of data in the contingency table. For this reason the model will give a perfect description of the data. The object of fitting this model is to obtain a yardstick for assessing the relative importance of all the possible parameters. These saturated estimates enable the quick formulation of unsaturated models which contain fewer than the maximum number of parameters. The aim of the whole exercise is to find the unsaturated model with the least number of parameters which adequately fits the data. For example, one unsaturated model for the leukemia data is given by

Table 10. Leukemia problem : data and parameter format

a) Observed probabilities

B
Radiation Exposure
(low to high)

		i \ j	1	2	m	
A	Leukemia Dead	1	q_{11}	q_{12}	q_{1m}	$q_{1.}$
	Mortality Alive	n	q_{n1}	q_{n2}	q_{nm}	$q_{n.}$
			$q_{.1}$	$q_{.2}$	$q_{.m}$	

b) Predicted probabilities and log-linear parameters

B
j →

		1	2	m	
A ↓ i	1	p_{11} λ_{11}^{AB}	p_{12} λ_{12}^{AB}	p_{1m} λ_{1m}^{AB}	λ_1^A
	n	p_{n1} λ_{n1}^{AB}	p_{n2} λ_{n2}^{AB}	p_{nm} λ_{nm}^{AB}	λ_n^A
		λ_1^B	λ_2^B	λ_m^B	

$$\log_e (p_{ij}) = u + \lambda_i^A + \lambda_j^B. \qquad (32)$$

This equation is termed the general independence model because the prediction depends only on the marginal leukemia and radiation probabilities and is therefore independent of any joint interactions, λ_{ij}^{AB}, which have been dropped from the model equation. The log-linear format allows for a wide variety of different parameter combinations to be tested, and how these combinations are chosen depends on both the number of variables in the contingency table and whether these variables are factors (causes) or responses.

In our example the leukemia variable is a response to the radiation variable because radiation can cause leukemia but not vice versa. More generally, the number of variables defines the maximum number of parameters that may be included in the model while the factor/response relationships indicate which of these parameters should be included. If an age category variable, c, is added to the leukemia contingency table then the saturated model takes the form

$$\log_e(p_{ijk}) = u + \lambda^{A}_{i} + \lambda^{B}_{j} + \lambda^{C}_{k} + \lambda^{AB}_{ij} + \lambda^{AC}_{ik} + \lambda^{BC}_{jk} + \lambda^{ABC}_{ijk} \qquad (33)$$

where k is a subscript identifying the individual age categories. However, because A is a response variable and 13 and c are factor variables the number of parameters in the saturated models may be substantially reduced. It is customary to include those joint parameters which measure interactions between factor variables, while joint parameters measuring interaction between a response variable and one or more of the factor variables are excluded from the model specification. For this reason an appropriate model for analysing the leukemia table is

$$\log_e(p_{ijk}) = u + \lambda^{A}_{i} + \lambda^{B}_{j} + \lambda^{C}_{k} + \lambda^{BC}_{jk} \qquad (34)$$

By fitting models similar to equation (34) to the three-way leukemia data Gokhale and Kullback (1978) were able to test the statistical significance of radiation exposure and age upon leukemia mortality. Their main conclusions were that while radiation exposure (λ^{B}_{j}) was the primary factor determining leukemia mortality certain combinations of age and radiation (λ^{AB}_{jk}) produced significant mortality in excess of the radiation effect. However, age (λ^{C}_{k}) on its own was not a significant component of mortality.

Information statistics have a number of important roles to play in the fitting and testing of log-linear models. Gokhale and Kullback (1978) have shown that finding the predicted probabilities, $\{p_{ij}\}$, which best fit an observed set of contingency table probabilities, $\{q_{ij}\}$, is equivalent to the problem of finding the values of p_{ij} which minimise the value of information gain given by

$$I(q:p) = \sum_{i}\sum_{j} q_{ij} \log(q_{ij}/p_{ij}). \qquad (35)$$

Recall that $I(q:p)$ tends to zero when all $q_{ij} = p_{ij}$. Furthermore it can be proved that, irrespective of the number of specified parameters, the equation which minimises the value of $I(q:p)$ will always be some form of the log-linear model. For this reason a natural strategy for estimating the parameter values in a log-linear model is to adopt a fitting procedure which identifies the parameter estimates which minimise information gain.

Information statistics can also be used to construct significant tests both for different model predictions and for their individual parameter estimates. For example, if we define $q_{ij}.n=x_{ij}$ and $p_{ij}\ .n=\hat{x}_{ij}$, then x_{ij} and $\hat{x}_{ij}$ and the observed and predicted cell frequencies respectively. Using these frequencies we can define a version of information gained termed minimum discriminant information and given by

$$2I(x : \hat{x}) = 2 \sum_{i}\sum_{j} x_{ij} \log(x_{ij}/\hat{x}_{ij}) \qquad (36)$$

which measures the degree of fit between the two sets of frequencies. This M.D.I. statistic is asymptotically distributed as chi-square with appropriate degrees of freedom and may be used to test the null hypothesis of no significant difference between the observed and predicted frequencies.

The significance tests for model parameters are made on differences in the value of the M.D.I. statistic obtained by fitting models with different parameter sets to the observed contingency table. For example, let $\hat{x}_a$ represent the prediction obtained from a model with the parameter set *a*, and let $\hat{x}_b$ denote the predictions obtained from an enlarged model containing additional parameters to those specified in *a*, then the goodness-of-fit between these models and the data is denoted by the M.D.I. statistics $2I(x : \hat{x}_a)$ and $2I(x : \hat{x}_b)$. Clearly, the value of $2I(x : \hat{x}_b)$ will be less than $2I(x : \hat{x}_a)$ because model *b* has more parameters and will therefore give a better fit to the data. The difference between these two statistics is given by the quantity

$$2I(\hat{x}_b : \hat{x}_a) = 2I(x : \hat{x}_a) - 2I(x : \hat{x}_b), \qquad (37)$$

which measures the amount of information accounted for by the additional parameters included in model *b*. The larger this difference, given the appropriate degrees of freedom, the greater is the explanatory power of the additional parameters. Fortunately, $2I(\hat{x}_b : \hat{x}_a)$ is also approximately distributed as chi-square and may be used to test the null hypothesis of no significant difference between the predictions of parameter set a and the improved predictions with the additional parameters in b. By testing differently specified models in sequence it is possible to use the statistic (37) to work out the significance of each variable's parameters as an explanation of the observed cell frequencies.

The application of log-linear models to geographical problems is a recent phenomena. It seems likely that their usage will increase because they provide a powerful method of evaluating alternative hypotheses about the factor combinations which explain observed cell frequencies. In addition to their technical elegance, log-linear models are representations of the most ubiquitious form of geographical data, the contingency table. Thus there is no shortage of problems which are suited to this style of analysis.

V APPLICATIONS AND DEVELOPMENTS

So far we have sketched out some of the simpler geographical applications of Shannon's entropy and the index of information gain. In this final section we take a wider view of the subject and trace the progress of information theory in geography during the recent past together with some of the technical difficulties which have emerged.

A dominant field of research has been the application of information statistics to measure various types of geographical concentration and dispersion. The early studies in this field often took their inspiration from the Theil's (1967) book, *'Economics and Information Theory'* and I and A Horowitz's (1968, 1970) studies of competition and concentration in the brewing industry. Early geographical adaptations of these methods included Semple's (1973) study of the spatial concentration of corporate headquarters and Garrison and Paulson's (1972, 1973) analyses of the effect of water availability on the concentration of manufacturing activity in the Tennessee Valley region. In a similar vein Chapman (1970, 1973) has used information gain and its decompositions to measure changes in the regional population distribution of the U.K. and U.S.A. during the past hundred years. As an alternative to the analysis of regional proportions, both Berry and Schwind (1969) and Chapman (1973, 1977) have proposed ways in which information statistics can be applied to migration and spatial interaction matrices to provide measures of the relative location of the places where the flows begin and end.

A number of geographers have used information statistics in conjunction with more familiar techniques in an effort to understand certain spatial problems. Marchand (1972) devised three versions of Shannon's redundancy, termed internal, structural and global redundancy, which he used to measure the relationship between natural vegetation types and cash crop production in the Andes and the Llanos of Venezuela. In a second study Marchand (1975) attempted a regionalization of Venezuela using variables measuring regional levels of industrial and agricultural activity as a basis for the classification. Briefly, the method adopted was to use factor analysis to group the variables and then to apply redundancy indices to assess the relative importance of each group in each region. The redundancy values suggested that an effective regionalization could have been achieved with a great reduction in the number of variables used in the original analysis. A more ambitious study of agricultural systems was undertaken by Chapman (1974, 1977) who analysed the responses of Indian farmers to the effects of climatic conditions on their rice yields. A game theoretic framework was adopted which involved the construction of matrices which contain for instance, the probability of an unirrigated crop surviving a drought. Entropy statistics were calculated from such data to provide measures of the farmer's uncertainty about the occurrence of different rice yields in the face of different climatic conditions and different management decisions.

Most of the studies described so far use data which are in the form of sets of regional proportions. An important technical problem which influences the results of such analyses is the way in which the boundaries used to define the regional data collection units condition the values of the proportions used in the analysis. Clearly, different regional partitions of the same study area will produce different sets of proportions which in turn will produce different values of the observed entropy or information gain. Geographers, such as Batty (1974, 1976), have often shown an awareness of this potential source of error but relatively little has been done to rectify the matter. However, in a recent paper, Batty and Sammons (1979) propose a method of modifying Shannon's entropy in accordance with the particular regional partition used in the analysis. They begin by arguing that Shannon's entropy is appropriate in its usual form only when the study area has been partitioned into equal sized regions. This condition ensures that the values of the observed proportions are independent of the partition. However, when the regions are of unequal size the proportions are to some extent dependent on the partition and therefore it is necessary to use a

modified version of Shannon's formula. Batty and Sammons' modified formula makes use of the ratios $\Delta x_i/\Delta x$, where Δx_i is the area of the ith region expressed as a proportion of the entire study area, and $\Delta x = 1/n$ where n is the total number of regions. Therefore, Δx is an idealised equal area region. Without stating the proof, the corrected entropy, $\bar{H}$, for the effect of unequal sized regions is given by

$$\bar{H} = H + \sum_{i}^{n} p_i \log (\Delta x_i/\Delta x) \tag{38}$$

where H is Shannon's entropy calculated in the usual way and p_i is the proportion of the variable found in the ith region. In conclusion Batty and Sammons argue that in future studies $\bar{H}$ should be used in preference to H and suggest that the ratio $\bar{H}/H$ could in some cases be used to identify the most suitable partition of the study area.

A related field of geographical inquiry where information statistics are beginning to make an impact is point pattern analysis. Broadly, this technique involves placing a grid of cells over the map of some point distribution and then analysing properties of the set of observations formed by counting the number of points in each cell. Medvedkov (1967, 1970) was the first geographer to devise entropy statistics to measure the degree of organisation present in point patterns and subsequently his methods have been used to analyse settlement patterns by Semple and Golledge (1970). More recently, Thomas and Reeve (1976) have demonstrated that Medvedkov's methods need to be modified to account for variations in the average number of points per cell between different patterns. Ecologists too use point pattern methods to describe properties of plant distributions and, among others, Bowman et al. (1971), Margalef (1958) and Pielou (1966, 1967 and 1977) have devised information statistics which may be used to measure properties such as the diversity of species in a plant population. A further aspect of point pattern analysis is the study of maps coloured black or white according to whether the variable is present or absent in each cell. Gatrell (1977) has examined the relationship between Shannon's redundancy and some simple spatial autocorrelation coefficients for a simulated set of such 'binary' maps. He demonstrated a parabolic relationship between the two indices such that high redundancy values were related to both high positive spatial autocorrelation (cells of the same colour tending to be near neighbours) and high negative spatial autocorrelation (cells of different colours tending to be near neighbours). This result leads him to suggest that redundancy is a useful surrogate for spatial autocorrelation. However, here it is important to realise that, whereas spatial autocorrelation coefficients measure the spatial arrangement of a geographical variable, information statistics only measure some property of the frequency distribution of a spatial variable. In fact Clayton and Lamprecht (1974) have conducted an empirical study of how various information statistics respond to changes in the arrangement of a pattern.

For the most part we have been concerned with applications of information statistics as index numbers for measuring properties of patterns and distributions. However, in the section on log-linear modelling we touched upon the use of these statistics in the construction of data models and it is in the field of mathematical modelling that we find the other major geographical application of information theory. Most of these applications are in the field of urban and regional modelling and originate with the work of Wilson (1967, 1970). He demonstrated that entropy - maximising principles could usefully be built into existing urban planning models which predicted variables such as the pattern of journey-to-work trips in a city or the number of households in urban regions. For example, his journey-to-work model predicts the number of work-trips between urban regions which both maximises observer's uncertainty (entropy) about which individuals made which trip, and satisfies given information about the numbers of jobs and workers in each region and the cost of travelling between regions. This entropy maximising solution with maximum uncertainty may be interpreted in another way. In a statistical sense it is said to be the most *likely* solution because it gives the individuals the maximum possible freedom to choose between work-trip routes given the information about costs, jobs and workers' residences. The entropy maximising principle was a significant improvement on earlier work because it improved both the theoretical elegance of these models and their capacity to fit observed distribution.

The derivation of predictive equations for such planning models requires knowledge of the complex mathematical topic of constrained maximisation which is outside the scope of this monograph. However, in a recent mathematical treatment of information theory in geography Webber (1979) has shown that the mathematical principles needed to derive the entropy-maximising urban models are simple logical extensions of those principles we used to derive the index numbers. Indeed, for the more mathematically inclined reader a number of texts have been listed in the bibliography which treat the analytical aspects of information theory in more detail than has been attempted here.

BIBLIOGRAPHY

A. THEORETICAL

Attneave, F. (1959), *Applications of Information Theory to Psychology.* (New York : Holt).

Batty, M. (1972), 'Entropy and spatial geometry', *Area* 4, 230-6.

Batty, M. (1974), 'Spatial Entropy', *Geographical Analysis,* 6, 1-32.

Batty, M. (1976), 'Entropy in spatial aggregation', *Geographical Analysis,* 8, 1-21.

Batty, M. and Sammons, R. (1979), 'A conjecture on the use of Shannon's formula for measuring spatial information', *Geographical Analysis,* 11, 305-10

Brillouin, L. (1956), *Science and Information Theory,* (New York : Academic Press).

Chapman, G.P. (1977), *Human and environomental systems : A Geographer's Appraisal.* (New York : Academic Press).

Clayton, C. and Lamprecht, J.L. (1974), 'Information Theory and Geographical differentiation', *Geografiska Annaler,* 56B, 78-89.

Duncan, O.D. and Duncan, B. (1955) 'A methodological analysis of segregation indices', *American Sociological Review,* 20, 210-17.

Fast, J.D. (1970), *Entropy,* (London : Macmillan).

Guiasu, S. (1977), *Information Theory with applications,* (New York : McGraw-Hill).

Gokhale, D.C. and Kullback, S. (1978), *The information in contingency tables,* (New York : Marcel Dekker).

Hartley, R.V.L. (1928), 'Transmission of information', *Bell System Technical Journal,* 7, 535-63.

Hobson, A. and Cheng, B.K. (1973), 'A comparison of the Shannon and Kullback information measures,' *Journal of Statistical Physics,* 7, 301-10.

Jones, D.S. (1979), *Elementary Information Theory,* (Oxford : Clarendon).

Khinchin, A.I. (1957), *Mathematical foundations of Information Theory,* (New York : Dover).

Kullback, S. (1959), *Information Theory and Statistics,* (New York : Wiley).

Margalef, D.R. (1958) 'Information Theory in Ecology', *General Systems,* 3, 36-71.

Pielou, E.C. (1977), *An introduction to mathematical ecology,* (New York : Wiley).

Quastler, H. (1955), *Information theory in Psychology,* (Glencoe : Free Press).

Reyni, A. (1965), 'On the foundations of information theory', *Review of the International Statistical Institute,* 33, 1-14.

Reza, F.M. (1961), *An introduction to Information Theory,* (New York : McGraw-Hill).

Shannon, C.E. (1948), 'A mathematical theory of communication', *Bell System Technical Journal,* 27, 379-423, 623-656. (This article is reprinted in Shannon, C.E. and Weaver, W. (1949) *The mathematical theory of communication,* Urbanna : University of Illinois Press).

Sheppard, E.S. (1976), 'Entropy, theory construction and spatial analysis', *Environment and Planning* A, 7, 279-91.

Stapleton, C.M. (1980), 'Limitations of log-linear models in geography', *Transactions, Institute of British Geographers (New Series),* 5, 502-508.

Theil, H. (1967), *Economics and Information Theory,* (Amsterdam : North Holland).

Theil, H. (1972), *Statistical decomposition analysis,* (Amsterdam : North Holland).

Upton, G.J.G. and Fingleton, B. (1979), 'Log-linear models in geography', *Transactions, Institute of British Geographers (New Series),* 4, 103-115.

Upton, G.J.G. and Fingleton, B. (1980), 'A rejoinder to comments by Dr Wrigley', *Transactions, Institute of British Geographers (New Series),* 5, 118-122.

Walsh, J.A. and Webber, M.J. (1977), 'Information Theory : Some concepts and measures', *Environment and Planning A,* 9, 395-417.

Webber, M.J. (1978), *Information Theory and urban spatial structure,* (London : Croom Helm).

Wilson, A.G. (1967), 'A statistical theory of spatial trip distribution models', *Transportation Research,* 1, 253-69.

Wilson, A.G. (1970), *Entropy in urban and regional modelling.,* (London : Pion).

Wrigley, N. (1976), *CATMOG No. 10 : An introduction to the use of logit models in geography,* (Norwich : Geo Abstracts).

Wrigley, N. (1979), 'Developments in the statistical analysis of categorical data', *Progress in Human Geography,* 3, 317-57.

B. Applications

Berry, B.J.L. and Schwind, P.J. (1969), 'Information and entropy in migrant flows', *Geographical Analysis,* 1, 5-14.

Bowman, K.D., Shenton, L.R., Hutcheson, K. and Odum, E.P. (1971), 'Comments on the distribution of indices of many species' in G.P. Patil, E.C. Pielon and W.E. Waters (eds.), *Many Species Populations, Ecosystems, and Systems Analysis, 3. Proceedings of the Symposium on Statistical Ecology,* (Pennsylvania State University), 315-66.

Chapman, G.P. (1970), The application of information theory to the analysis of population distributions in space', *Economic Geography,* 46, 317-31.

Chapman, G.P. (1973), The spatial organizations of the population of the United States and England and Wales', *Economic Geography,* 49, 325-43.

Chapman, G.P. (1974), 'Perception and regulation : a case study of the farmers in Bihar', *Transactions, Institute of British Geographers,* 62, 328-43.

Connelly, D. (1972), 'Information theory and geomorphology', in R. Chorley (ed.), *Spatial Analysis in Geomorphology,* (New York : Harper & Row), 91-108.

Finkelstein, M.O. and Friedberg, R.M. (1967), The application of an entropy theory of concentration to the Clayton Act', *The Yale Law Review,* 76, 677-717.

Garrison, C.B. and Paulson, A.S. (1972), 'Effect of water availability on manufacturing activity in the Tennessee Valley Region', *Water Resources Research,* 8, 301-16.

Garrison, C.B. and Paulson, A.S. (1973), An entropy measure of the geographic concentration of economic activity', *Economic Geography,* 49, 319-324.

Gatrell, A.C. (1977), 'Complexity and redundancy in binary maps', *Geographical Analysis,* 9, 27-41.

Gurevich, B.L. (1969), 'Measures of feature based and areal differentiation and their use in city services', *Soviet Geography : Review and Translation,* 10, 380-6.

Gurevich, B.L. (1969), 'Geographical differentiation and its measures in a discrete system', *Soviet Geography : Review and Translation,* 10, 387-413.

Horowitz, I. (1970), 'Employment concentration in the Common Market : an entropy approach', *Journal of the Royal Statistical Society, Series A,* Part 3, 463-79.

Horowitz, I. and Horowitz, A. (1968), 'Entropy, Markov processes and competition in the brewing industry', *Journal of Industrial Economics,* 16, 196-211.

Horowitz, I. and Horowitz, A. (1970), 'Structural changes in the brewing industry', *Applied Economics,* 2, 1-13.

Hutcheson, K. (1970), 'A test for comparing diversities based on the Shannon formula', *Journal of Theoretical Biology,* 29, 151-4.

Lloyd, M., Zar, J.G. and Karr, J.R. (1968), 'One the calculation of information theoretical measures of diversity', *The American Midland Naturalist,* 79, 257-72.

Marchand, B. (1972), 'Information theory and Geography', *Geographical Analysis,* 4, 234-57.

Marchand, B. (1975), 'On the information content of regional maps : the concept of geographical redundancy', *Economic Geography,* 51, 117-27.

Medvedkov, Y. (1967), 'The regular component in settlement patterns as shown on a map', *Soviet Geography: Review and Translation,* 8, 150-68.

Medvedkov, Y. (1967), The concept of entropy in settlement pattern analysis', *Papers, Regional Science Association,* 46, 307-16.

Medvedkov, Y. (1970), 'Entropy, an assessment of its potentialities in Geography', *Economic Geography,* 46, 307-16.

Peach, C. (1975), *Urban social segregation,* (London : Longman).

Pielou, E.C. (1966), 'Shannon's formula as a measure of specific diversity - its use and misuse', *American Naturalist,* 100, 463-5.

Pielou, E.C. (1966), 'The measurement of diversity in different types of biological collections', *Journal of Theoretical Biology,* 13, 131-44.

Pielou, E.C. (1967), 'The use of information theory in the study of the diversity of populations', *Proceedings, 5th Berkley Symposium Statistics and Probability,* 4, 163-77.

Semple, R.K. (1973), 'Recent trends in the spatial concentration of corporate headquarters', *Economic Geography,* 49, 309-18.

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, 169-80.

Semple, R.K. and Golledge, R.G. (1970), 'An analysis of entropy changes in a settlement pattern over time', *Economic Geography,* 46, 157-60.

Theil, H. and Finizza, A.J. (1971), 'A note on the measurement of racial integration of schools by means of informational concepts', *Journal of Mathematical Sociology,* 1, 187-93.

Thomas, R.W. and Reeve, D.E. (1976), 'The role of Bose-Einstein statistics in point pattern analysis', *Geographical Analysis,* 8, 113-36.

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

This series, Concepts and Techniques in Modern Geography
is produced by the Study Group in Quantitative Methods, of
the Institute of British Geographers.
For details of membership of the Study Group, write to
the Institute of British Geographers, 1 Kensington Gore,
London, S.W.7.
The series is published by Geo Abstracts, University of
East Anglia, Norwich, NR4 7TJ, to whom all other enquiries
should be addressed.