INFORMATION THEORY AND CLASSIFICATION

IN GEOGRAPHY.

INFORMATION THEORY AND CLASSIFICATION IN GEOGRAPHY

by

JAMES A. WALSH, B.A.

A Research Paper

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Arts

McMaster University

December, 1975.

MASTER OF ARTS (1975) McMASTER UNIVERSITY
(Geography) Hamilton, Ontario.

TITLE : Information Theory and Classification in Geography.

AUTHOR : James Anthony Walsh, B.A. (University College, Dublin)

SUPERVISOR : Dr. M.J. Webber

NUMBER OF PAGES : v, 156.

SCOPE AND CONTENTS:

In this paper some fundamental concepts of information theory and their potential for classification construction in geography are discussed. The concepts of information and uncertainty are shown to be equivalent. Three different information measures are discussed and the particular situations for which each is appropriate are identified. With this background the available literature on the application of information theory to classification is reviewed and reinterpreted. A classification algorithm for objects characterized by multistate ordinal attributes is presented and tested. Recommendations for further research include the consideration of explicitly spatial information measures and the examination of more general information metrics.

ACKNOWLEDGEMENTS

In writing this paper I have been influenced by the work of many scholars. Their contributions are gratefully acknowledged. Special thanks are extended to my supervisor, Dr. Michael Webber, for his guidance, suggestions and patience. My thanks are also extended to Dr. Martin Taylor for providing the data to test the algorithm developed in chapter three, and to Russ Lee for commenting on earlier drafts of some of the material.

I am also very grateful to Rosanna Wan for her patience and skill in typing this paper.

TABLE OF CONTENTS

INTRODUCTION

From time to time in the historical development of most disciplines some new words or phrases enter their vocabulary. Some of these words, in an unexplained manner, acquire a mythical or "magical" status. Among the words of this type in contemporary human geography are entropy and information theory. The word entropy has a long fascinating history (cf. Dutta 1968), but until 1948 it was largely confined to the physicist's repertoire. Then Shannon in his classic work on information theory (Shannon and Weaver 1949), had the insight to see that the entropy measure of the physicists could also be used as an information measure. Very soon entropy and information were becoming magic words in a number of disciplines, including biology (Quastler 1953), sociology (Van Soest 1954), psychology (Rapoport 1956, Garner 1962), and ecology (Margalef 1958, Pielou 1969). In some cases the pioneers in the field were over-optimistic, as in psychology, (Whitla 1968), and biology (Johnson 1970).

In recent times the concepts of entropy and information have made their way into the literature of economics (Georgescu-Roegen 1971, Mogridge 1972, Marschak 1974), and geography (Wilson 1970, Medvedkov 1970, Lee 1974, Webber 1975). In geography entropy or information measures have been used for two somewhat different purposes. On the one hand, there is the work pioneered by Wilson, which is mostly concerned with generating most likely probability distributions to predict patterns of spatial behavior. This body of work using the entropy maximizing paradigm has been largely

influenced by the ideas of Jaynes (1957) and Tribus (1969).. On the other hand, there are those who use information measures to provide a statistical summary of spatial distributions, (Medvedkov 1967a, b, Gurevich 1969a, b, Garrison and Paulson 1973). In these applications information statistics are used to measure the uncertainty present in distributions, since Shannon's entropy or information measure "is the correct measure of the "amount of uncertainty" in a probability distribution" (Jaynes, 1963a). This paper is concerned with the second type of application of information statistics in geography.

The purpose of this paper is to demonstrate the potentialities of information theory for the construction of classifications in geography. With this purpose in mind the first chapter is devoted to a discussion of the basic concepts of information theory that are necessary for an understanding of the remainder of the paper. In particular, the equivalence between the concepts of information and uncertainty, and the question of which information measure one should use in a particular situation, are examined. In the second chapter the classification problem is stated. Existing information theoretic approaches dealing with this problem are reviewed and modified so that they can be applied to geographic problems. In the third chapter an algorithm for classifying individuals characterized by scores on multistate ordinal attributes is presented. The algorithm is used to discover groups within a sample of 198 individuals. Many possible applications of the algorithm, of interest to behavioral geographers, are indicated. Finally, in chapter four a short summary of the paper is given, along with some conclusions and outlines for further research.

CHAPTER I

INFORMATION THEORY : SOME CONCEPTS AND MEASURES.

Introduction

One of the main concerns of information theorists has been the quantification of information. In general the quantification of information, as outlined in Shannon and Weaver (1949), Goldman (1953), Khinchin (1957) and Kullback (1959) is related to a known set of alternatives - an exhaustive set of mutually exclusive possibilities which is assumed to be known. This set is usually referred to as an ensemble or universe of possible states or outcomes from an experiment. In information theory one is never concerned with the realization of a particular state or outcome as such, but rather with its realization within the universe of all possible states or outcomes. Therefore, information is provided by a selection from the universe of possible states, a selection which reduces the a priori uncertainty of what was going to happen.

The universe of all possible states defines our a priori uncertainty, freedom of choice or doubt. Clearly the larger the universe of possible states the greater will be our a priori uncertainty. Information then may be considered to be that quantity which reduces our uncertainty, curtails our freedom of choice or removes our doubt. Furthermore, associated with the concept of information is an element of surprise, unexpectedness or improbability. The information to be gained from an experiment then depends on the set of possible outcomes and

the likelihood of each. In the words of Rapoport (1954) "the repertoire of the source from which the message is chosen" is all-important when measuring the information content of a message.

The purpose of this chapter is to outline and clarify the basic concepts of information theory that will be necessary for an understanding of the remaining chapters. In later chapters the discussion will focus on measuring the information content of areal distributions and on ways of arranging (classifying) the areal units so as to minimize the uncertainty of the distribution. Here the measurement of the information content of messages and its relation to uncertainty are examined. This is followed by a study of three information measures and a discussion of the appropriate measure to use in a given situation. Information measures are given for objects characterized by a number of non-independent variables. The interrelationships between the concepts of information, entropy and order are discussed and clarified. Finally, the value or usefulness of the information contained in a message is examined.

Information and Uncertainty

Suppose we have a series of observations as a result of experimentation on some phenomenon X. The states that X can attain are assumed to be discrete and to belong to the sequence $x_1, x_2, \ldots\ldots, x_i, \ldots\ldots$ One may think of X as being a message and the x_i as being the symbols or letters that could be used to make up the message. Similarly, X could be the landuse pattern of say some city at a particular time, while the x_i would be the different landuses that could occur at that time. Following Goldman (1953) and Good (1956) the amount of information to be obtained from observing that $X = x_i$ may be defined to be

$$I(X = x_i) = \log \left[\frac{\text{prob } (X = x_i) \text{ after the observation}}{\text{prob } (X = x_i) \text{ before the observation}} \right] \qquad (1.1)$$

Clearly, if there are no errors involved in the observation the numerator in (1.1) is unity, and

$$I(X = x_i) = - \log_k P_i \qquad (1.2)$$

where P_i is the denominator in (1.1) and k is the logarithmic base.

From the definition, (1.1), one can consider the expression $-\log P_i$ to be a measure of the information content or self-information (Ingels 1971) of the message $X = x_i$. Hence the information content of any message is a function of the improbability of its occurrence. Henceforth $-\log P_i$ is considered to be a measure of the amount of potential information, P.I., that can be obtained from the observation that $X = x_i$, since strictly speaking one is concerned with the amount of information that can be obtained, rather than the actual amount obtained from the realization of a particular state x_i.

It was indicated earlier that information is that quantity which reduces our uncertainty. Hence it should be possible to interpret (1.2) as a measure of our uncertainty with respect to x_i. Intuitively it is reasonable to expect that our uncertainty is equivalent to the minimum amount of information needed to select a particular x_i from a set of possible x's. If our uncertainty in any situation is defined to be the minimum number of k - nary questions ($k > 1$, $k \in \mathbb{R}$) needed in order to answer a specific question in the light of specific evidence, then

$-\log_k P_i$ can be shown to measure this uncertainty, where P_i is the probability of occurrence of the event about which we are uncertain.[1]

Suppose some experiment X is repeated N times. Let the outcome of each experiment be denoted by x_i, i = 1, 2, 3, ..., N. Further assume that the outcomes of the N experiments are all different. Then the minimum number of k-nary questions required to determine the particular x_i that represents the outcome of a given experiment is $U = \log_k N$. However, since all the x_i are equally likely to represent the outcome of the experiment, the probability that $X = x_i$ is $P_i = 1/N$. Then by (1.2) the P.I. content of the message $X = x_i$ is $I(X = x_i) = -\log_k P_i$ which, of course, is equal to $\log_k N = U$.

It is not necessary to assume that all the outcomes of an experiment are equiprobable to demonstrate the equivalence between the definitions of information content and uncertainty. For some experiment X let there be m different outcomes. Assume that in a series of N, (N>m), experiments the outcome $X = x_i$ occurs n_i times, where $\sum_{i=1}^{m} n_i = N$, $n_i \geq o$, all i. Our uncertainty in this case is equivalent to the minimum number of questions required to determine which particular x_i represents the outcome of a given experiment. In any sequence of N experiments the number of different outcomes that could each be observed n_i times is N/n_i, i = 1, 2, ..., m. Therefore, to determine the outcome of a particular experiment the minimum number of k-nary questions that it is necessary to ask is $\log_k N/n_i$, where n_i is the number of times the particular outcome

---

[1] The concepts of information and uncertainty have been studied from a different standpoint by De-Groot (1962).

occurs in the sequence of N experiments. Equivalently our uncertainty about the outcome of this particular experiment is $U = \log N - \log n_i$. Conversely, the amount of potential information to be obtained from observing that x_i is the outcome of one performance of the experiment X is $I(X = x_i) = -\log_k n_i/N = \log N - \log n_i = U$.

Hence the definitions of information content and uncertainty are equivalent. Therefore $-\log P_i$ can also be regarded as a measure of the amount of information needed to select a particular x_i from the set of possible x's. It follows immediately from the definition of uncertainty that $-\log P_i$ is the minimum amount of information needed to select the particular x_i. From (1.2) one may observe that the more improbable a particular x_i is the more information that is needed to select it from the universe of possible x's. The foregoing discussion is now summarised in the following list of equivalent interpretations of the expression $-\log P_i$.

$-\log P_i$ = Amount of potential information received from message $X=x_i$ (1.3a)
= Potential information content of message $X=x_i$ (1.3b)
= Uncertainty about content of message $X=x_i$ (1.3c)
= Minimum amount of information needed to select the message $X=x_i$ from the set of all possible messages. (1.3d)

One may observe here that for (1.3a) to be equivalent to (1.3b) it is necessary to assume that the information we are receiving is being transmitted to us in the absence of noise. Throughout the remainder of this paper the first interpretation, (1.3a), will be used. Its equivalence to the other interpretations will be rarely restated, but it will be constantly assumed.

The rationale for using the logarithm of the information function is that it satisfies the following two desiderata. First it allows one to establish a formal equivalence between information and uncertainty measures, as indicated above. Secondly it satisfies the requirement that the P.I. obtained from two independent events be equal to the sum of the P.I. received from each event separately. In fact if one makes the definition of the amount of P.I. to depend solely on probabilities and if one insists on it having the additive property for probabilistically independent events, then minus the logarithm of the probability is the only possible definition, (Good 1956).

Shannon's Information Measure

It has been established that the amount of potential information obtained from observing any state x_i of some phenomenon X can be measured by $-\log P_i$ where P_i is the probability of X being in the state x_i. If all the states that X can assume are a priori equiprobable then clearly the amount of P.I. to be obtained from observing say m states is $-m(\log P_i)$. In general, however, all the states do not have equal prior probabilities, as is evident from the landuse example given earlier. Therefore, the amount of P.I. to be obtained from each state x_i has to be weighted by the probability of its occurrence. We may now define the mean amount of potential information to be obtained from each state x_i in a sequence of x_i's as

$$I_1 = -\sum_i P_i \log P_i \tag{1.4}$$

where $P_i \geq 0$, all i and $\sum_i P_i = 1$.

Expression (1.4) corresponds to Shannon's well known measure of information or entropy. By (1.3d) one may also interpret I_1 as a measure of the minimum amount of information needed, on average, to select a particular x_i from the set of possible x_i's. This type of interpretation led McKay (1950) to suggest the term "selective information" for I_1. From (1.3b), (1.3c) and (1.4) it follows that the mean information content, mean uncertainty and the entropy of a sequence of messages are all equivalent.

The choice of logarithmic base for (1.4) is arbitrary. Information theorists use logarithms to base two, and the information units then are called binary digits or bits. When natural logarithms are used MacDonald (1952) suggested that the information units be called "nits" and for logarithms to base ten it has been suggested that the units be called decimal digits (Good 1953), Hartleys (Abramson 1963) or decits (Pielou 1966a). In the remainder of this paper natural logarithms will be used, unless an occasion demands otherwise.

In applications of I_1 the prior probabilities P_i may be unknown. Then the maximum likelihood estimators of the P_i, $\hat{P}_i = n_i/N$, are used where n_i is the frequency of occurrence of the state x_i in a sample of size N. Then I_1 may be rewritten as

$$\hat{I}_1 = -\sum_i \frac{n_i}{N} \log \frac{n_i}{N} \tag{1.5}$$

where $\sum_i n_i = N$ and $\sum_i \frac{n_i}{N} = 1$.

In this case $\hat{I}_1$ is the geometric mean of the amount of P.I. obtained from observing any state x_i.

This can be easily shown.

Let $P_i = n_i/N$.

The amount of potential information obtained from a particular sample of size N is $-\log \prod_i (P_i)^{n_i}$.

If every state had the same probability P of occurrence then the mean amount of P.I. to be obtained from the observation of any state would be -log P, by (1.2). Then if we ask what function f satisfies the condition that

$$I(X) = -\log P = -\log f\left(\prod_i (P)^{n_i}\right)$$

$$= -\log f\,(P^N)$$

the answer is, of course, that f is the N^{th} root. Therefore, the mean amount of potential information obtained from each observation in the sample is

$$I(X) = -\log \prod_i \left((P_i)^{n_i/N} \right)$$

$$= -\sum_i n_i/N \log (n_i/N)$$

$$= \hat{I}_1 .$$

The sampling distribution of $\hat{I}_1$ has been studied by Basharin (1959) and Bowman et. al. (1971). Basharin showed that $\hat{I}_1$ is a biased, consistent, asymptotically normal estimate of I_1, and that its mean and variance have the following values,

$$E(\hat{I}_1) = I_1 - \frac{S-1}{2N} \log_2 e + \Theta(N^{-2}) \tag{1.6}$$

$$\text{Var}(\hat{I}_1) = \frac{1}{N}\left[\sum_{i=1}^{S} \frac{n_i}{N}\left(\log \frac{n_i}{N}\right)^2 - (I_1)^2\right] + \Theta(N^{-2}) \tag{1.7}$$

where S is the number of states that X can assume and $\Theta(N^{-2})$ designates a quantity of order N^{-a}, $a > 0$. When the S states are equally likely $\hat{I}_1$ has been shown to be χ^2 distributed, (Bowman et al. 1971). Formula (1.7) can be used in a test for equality of I_1 from two samples following a method proposed by Hutcheson (1970). However, the fact that in calculating the variance formula it is necessary to assume that the number of states that can occur in the sampling universe is known, may limit the applicability of this procedure to geographic situations.

It has been suggested by Pielou (1966a) that $\hat{I}_1$ can never be regarded as strictly equal to I_1. The reason offered is that to do so would involve making two assumptions, first that the population at hand is a sample from some undefined, conceptually infinite population, and second that it is an exactly representative sample. These assumptions are made when invoking the Law of Large Numbers to equate the probabilities P_i with $\hat{P}_i = n_i/N$, (Von Mises 1957). Here it is felt that this mode of reasoning is conditional on one's interpretation of what is meant by the term probability. The mode of reasoning that Pielou exemplifies follows from adopting an objectivist and relative frequency interpretation of probability. However, if one subscribes to the subjective school of probability (Jeffreys 1939, Good 1950, Savage 1954, Barnard 1962) where a probability statement is considered to represent one's degree of belief in propositions, then one does not have to appeal

to the Law of Large Numbers nor to vague undefined infinite populations. The probabilities may still be defined in relative frequency terms, but now one can interpret the observed relative frequencies as being the true population probabilities consistent with what information is available. These probabilities can then be used to calculate $\hat{I}_1$ from the sample data at hand, and $\hat{I}_1$ will be equal to I_1, since now the probabilities are being defined subjectively as $P_i = \hat{P}_i = n_i/N$, representing our belief that the system being studied is in certain states. Hence $\hat{I}_1$ can be equal to I_1, if the probabilities are defined subjectively.

Brillouin's Information Measure

An alternative way of measuring the mean information content of a message was proposed by Brillouin (1956). He considered a situation where N different things might happen. He also assumed that each of the N possible outcomes were equally probable a priori. The information content of the event that actually happened then was defined as

$$I = K \log N, \quad K \text{ a constant.} \tag{1.8}$$

Of course, this definition is mathematically equivalent to the one given earlier in (1.2).

In the light of this definition of the information content of an event one can easily derive a measure of the mean information content of a particular outcome from a series of experiments. Assume an experiment X is performed N times and that for each experiment there are m possible outcomes. The outcome x_i occurs n_i times such that $\sum_{i=1}^{m} n_i = N$. The number of different permutations of the N outcomes subject to the

condition that the i^{th} outcome occurs n_i times, all i, is

$$W = N! / \prod_{i=1}^{m} n_i!$$

When each of these permutations are equally probable the potential information content of the one that is realized is B = log W. Then, of course, the mean information content of each of the N outcomes that occur is

$$I_2 = B/N = \frac{1}{N} [\log N! - \sum_{i=1}^{m} \log n_i!] \qquad (1.9)$$

If, and only if, all the n_i are large Stirling's approximation, $\log x! \simeq x \log x - x$, may be applied to (1.9). Then a good approximation of I_2 for large N is

$$\tilde{I}_2 = - \sum_{i} (n_i/N) \log (n_i/N) . \qquad (1.10)$$

Otherwise, in order to evaluate (1.9) the more complete form of Stirling's approximation, $\log x! \simeq x \log x - x + 0.5 \log (2\pi x)$, should be applied, or (1.9) may be calculated directly from the table of $\log x!$, $x = 1, 2, \ldots, 1000$, given by Pearson and Hartley, (1954).

At this point one may observe that while the formulae for $\hat{I}_1$ and $\tilde{I}_2$ are identical, quite different interpretations are given to the results depending on whether $\hat{I}_1$ or $\tilde{I}_2$ is being calculated. $\hat{I}_1$ provides an estimate (albeit a biased one) of I_1, while $\tilde{I}_2$ provides an approximation to I_2. Thus whenever the expression $- \sum_{i} n_i/N \log n_i/N$ is used it should always be indicated whether it is being used as an estimator or as an approximation measure. Unfortunately, in the literature this distinction is rarely made explicit.

In the literature (Margalef 1958, Pielou 1967, 1969, Lloyd et. al. (1968), Orloci 1968, 1970) it has been suggested that I_2 is the appropriate measure to use to determine the mean information content of any event in a population of events which is completely sampled, since Shannon's measure I_1 is, strictly speaking, defined only for an infinite population. This question of when to use the Shannon or Brillouin measure is now examined in more detail.

Comparison of the Shannon and Brillouin Measures

The Brillouin information measure, I_2, is always less than or equal to the Shannon measure, I_1, or its estimate $\hat{I}_1$.

$$\hat{I}_1 - I_2 = -\sum_i \frac{n_i}{N} \log \frac{n_i}{N} - \frac{1}{N} \log \frac{N!}{n_1!n_2!\ldots n_m!}$$

$$= -\frac{1}{N} \left[\log \frac{N!}{n_1!n_2!\ldots n_m!} + \sum_i n_i \log \frac{n_i}{N} \right]$$

$$= -\frac{1}{N} \log \left[\frac{N!}{n_1!n_2!\ldots n_m!} \cdot \prod_{i=1}^{m} \left(\frac{n_i}{N}\right)^{n_i} \right]$$

The expression in square brackets is equivalent to a multinomial probability (Feller 1968, p.168) and is, therefore, necessarily less than or equal to unity.

Hence $$\hat{I}_1 - I_2 \geq 0. \qquad (1.11)$$

However, the expression in brackets is equal to unity only when one of the n_i is equal to N and all the other n_i's are zero. This corresponds to a situation where the same result is achieved from each performance of an experiment. In this situation the potential information content of any

result is zero, whether it is measured by I_1 or I_2. Hence, $\hat{I}_1 \geq I_2$ with equality if, and only if, one event occurs with probability one and the others with probability zero.

Given this mathematical result the next intriguing question is how or why does it arise?. The answer to this question is especially interesting when it is remembered that both I_1 and I_2 are, by definition, measures of the mean information content of a message selected from a set of possible messages. The distinction between the Shannon and Brillouin methods of calculating the mean information content of a message lies in the way the probabilities, that a particular message represents a certain event, are calculated. In Shannon's measure it is assumed that these probabilities are known a priori, cf (1.1) and (1.4), and that they remain the same for each message. Hence they must be defined in some way external to the events being studied. However, for Brillouin's measure the probabilities for each message are calculated from the events being studied. The probabilities are defined in relative frequency terms, but these relative frequencies cannot be determined until all the events, whose mean information content we wish to measure have already occurred. Therefore, when we apply Brillouin's measure we already know the total number of events that have occurred and the relative frequencies of the different types of events. It is precisely because we possess this structural information that I_2 is less than I_1. To see this consider a typical situation to which Brillouin's measure would be applied.

An experiment X for which there are m different outcomes is performed N times. The first outcome occurs n_1 times, the second n_2 times and so on such that $\sum_{i=1}^{m} n_i = N$. Then Brillouin's measure is, by (1.9)

$$I_2 = \frac{1}{N} \left[\log N! - \sum_{i=1}^{m} \log n_i! \right]$$

It is known that for large values of N, I_2 provides a good approximation to Shannon's measure, $I_1 = - \sum_i P_i \log P_i$, where each P_i is estimated by $\hat{P}_i = n_i/N$. When we estimate each P_i by $\hat{P}_i = n_i/N$ we are in fact assuming that our prior probability of the occurrence of each of the outcomes of type i does not change throughout the sequence of experiments. However, if our sample forms a complete population, then before we proceed to calculate the information content of a message about the outcome of any experiment we possess some structural information, i.e. we know the total number of experiments that have been performed and also the frequencies of each of the different outcomes. This information must be taken into account when we calculate the probabilities that are used to determine the information content of each outcome. In this situation the sampling procedure is without replacement and the probability of any outcome, as the sequence of experiments proceeds, becomes dependent on the outcomes of the experiments that have already been performed. If after performing some number, say K (K < N), of the experiments it is found that the frequencies of the different outcomes are given by the sequence $\{k_i\}_{i=1}^{m}$, ($k_i \leq n_i$, all i) such that $\sum_{i=1}^{m} k_i = K$, then there are only $n_i - k_i$ outcomes of type i from the remaining N - K experiments. Then the probability of the $K + 1^{st}$ outcome being of type i is $\frac{n_i - k_i}{N - K}$ rather than n_i/N, and the information content of this outcome is

$$- \log \frac{n_i - k_i}{N - K} = \log (N - K) - \log (n_i - k_i).$$

Using this method to calculate the information content of each outcome it follows that the total information content of the N outcomes is

$$\sum_{K=0}^{N-1} \log (N-K) - \sum_{i=1}^{m} \sum_{k_i=0}^{n_i-1} \log (n_i - k_i) \qquad (1.12)$$

$$= \log N! - \sum_{i=1}^{m} \log n_i!$$

$$= B, \text{ (cf. example in Appendix A).}$$

As usual the mean information content of each outcome to the experiment is $I_2 = B/N$.

From this derivation of Brillouin's measure we may conclude that I_2 measures the mean information content of any event in a finite population of events when the whole population is sampled. If, however, the events being studied form a random sample from a very much larger population then the size of the population and the frequencies of the different types of events within it may be unknown. Lacking this information we can only assume that the probability P_i of any event being of type i will be the same for all events. Then the mean information content of each event in the large population will be measured by $I_1 = - \sum_i P_i \log P_i$ where the P_i are defined exogeneously or calculated from the relative frequencies in a random sample from the population. From (1.12) observe that when K = 0 the information content of the outcome of the first experiment is $-\log n_i/N$ and on average this value is $- \sum_i n_i/N \log n_i/N = \hat{I}_1$, the estimate of the mean information content of any outcome as measured by Shannon. Therefore, the mean information content of the first outcome or observation in a sample is the same by either measure. Hence Shannon's measure, I_1, may

be applied to calculate the mean information content of the first observation in a collection that may be a population itself or a sample from a very much larger or conceptually infinite population. When the collection represents a sample from a larger population I_1 provides a measure of the mean information content of each observation in the population. Whether or not it is necessary to assume that the sample is from a conceptually infinite population depends on whether one is taking a relative frequency measure to represent a subjective or an objective probability.

Earlier it was shown that I_1 is also a measure of our mean uncertainty about the outcome of any experiment. So also is Brillouin's measure, I_2. Both I_1 and I_2 measure the minimum number of questions we need to ask, on average, to determine the outcome of any experiment. Our uncertainty about the outcome of the first experiment, on average, is the same whether it is measured by I_1 or I_2. In a completely sampled population, however, after each experiment the uncertainty about the outcomes of the remaining ones is reduced, by the information obtained from identifying the outcomes of the experiments that have occurred. Some further properties of I_2 are examined now.

The expected reduction in the information content of a sample when one randomly chosen observation is omitted is the same for either measure, (Baer, 1953).

Consider two samples, one of size N and the other of size N-1. The total information content of the sample of size N may be written as

$$B^N = \sum_{i=1}^{m} \frac{n_i}{N} \log \frac{N!}{n_1! n_2! \ldots\ldots n_m!} \quad .$$

The probability that the omitted observation represents the occurrence of an event of type i is $P_i = n_i/N$. The expected amount of information to be

obtained from the sample of size N-1 then is

$$E(B^{N-1}) = \sum_{i=1}^{m} \frac{n_i}{N} \log \frac{(N-1)!}{n_1!n_2!\ldots(n_i-1)!\ldots n_m!} \quad .$$

The expected difference in the information content of the samples is

$$\begin{aligned} E(\Delta B) &= E(B^N) - E(B^{N-1}) \\ &= \sum_i \frac{n_i}{N} \log \left[\frac{N!}{n_1!n_2!\ldots n_m!} \cdot \frac{n_1!n_2!\ldots(n_i-1)!\ldots n_m!}{(N-1)!} \right] \\ &= -\sum_i \frac{n_i}{N} \log \frac{n_i}{N} \quad . \end{aligned} \tag{1.13}$$

A consequence of the previous result is that the expected effect of omitting a randomly chosen observation from a sample is to decrease the sample average amount of potential information per observation, when the information is measured by I_2.

The mean information content of any observation in the sample of size N is $I_2^N = B^N/N$. Similarly, the expected mean information content of any observation in the sample of size N-1 is $E(I_2^{N-1}) = E(B^{N-1})/N-1$. Then the expected difference in the mean information content of each observation is

$$\begin{aligned} E(\Delta I_2) &= I_2^N - E(I_2^{N-1}) \\ &= \frac{B^N}{N} - \frac{E(B^{N-1})}{N-1} \\ &= \frac{E(\Delta B)}{N-1} - \frac{B^N}{N(N-1)} \\ &= -\frac{1}{N(N-1)} \log \left[\frac{N!}{n_1!n_2!\ldots n_m!} \cdot \prod_{i=1}^{m} \left(\frac{n_i}{N} \right)^{n_i} \right] . \end{aligned}$$

As before the expression in square brackets corresponds to a multinomial probability and is, therefore, less than or equal to unity. Hence

$$E(\Delta I_2) \geq 0 \tag{1.14}$$

with equality if, and only if, $n_i = N$ for one i, say i_0, and $n_i = 0$ for all $i \neq i_0$. In terms of uncertainty the last result may be interpreted as follows: on the average our mean uncertainty about the type of any event decreases as the number of events in the sample decreases.

If the Shannon estimator, $\hat{I}_1$, were used the result of disregarding a randomly chosen observation may be either an increase or a decrease in the average amount of potential information to be obtained from any observation. The result depends on the frequency of occurrence of the type of event represented by the disregarded observation. If the frequency is high the result is an $\hat{I}_1' > \hat{I}_1$ where $\hat{I}_1'$ is an estimate of the average amount of potential information to be obtained from any observation in the sample of size N-1. Conversely, if the disregarded observation is atypical in the sense that the event that it represents has a low frequency then the result is $\hat{I}_1' < \hat{I}_1$.

Finally, the sensitivity of the two information measures to the sample size, N, is examined. By definition if the probability distribution remains constant Shannon's measure, $I_1 = -\sum_i P_i \log P_i$, is not affected by varying N. However, Brillouin's measure, I_2, is an increasing function of N for all n_i and N satisfying the following condition,

$$\frac{N}{\prod_i n_i} < \left(\frac{2\pi}{e}\right)^{m-1}$$

where m is the number of different types of events that occur and e is the natural logarithmic base. When this condition is satisfied I_2 increases at a decreasing rate with respect to N.

To see this write

$$I_2 = \frac{1}{N} [\log N! - \sum_i \log NP_i!] \text{ , where } \sum_i NP_i = N .$$

Applying the extended version of Stirling's approximation,

$$\log x! \simeq (x + \frac{1}{2}) \log x - x + \frac{1}{2} \log (2\pi) ,$$

I_2 may be rewritten as

$$I_2 = \frac{1}{N} [(N + \frac{1}{2}) \log N - N + \frac{1}{2} \log (2\pi) - \sum_{i=1}^{m} ((NP_i + \frac{1}{2}) \log NP_i - NP_i + \frac{1}{2} \log (2\pi))]$$

$$= - \sum_i P_i \log P_i - (\frac{m-1}{2N}) \log 2\pi N - \frac{1}{2N} \sum_i \log P_i .$$

Then $$\frac{\delta I_2}{\delta N} = (\frac{m-1}{2N^2}) \{ \log (2\pi N) - 1 \} + \frac{1}{2N^2} \sum_i \log P_i$$

This derivative is strictly positive when

$$(m-1) \log_e (2\pi N) - (m-1) + \sum_{i=1}^{m} \log_e P_i > 0$$

$$\Longleftrightarrow \log (2\pi N)^{m-1} + \log \prod_{i=1}^{m} P_i > (m-1)$$

$$\Longleftrightarrow \log \{ (2\pi N)^{m-1} . \prod_{i=1}^{m} P_i \} > m-1$$

$$\Longleftrightarrow \quad \{ (2\pi N)^{m-1} \cdot \prod_i P_i \} > e^{(m-1)}$$

$$\Longleftrightarrow \quad \left(\frac{2\pi N}{e}\right)^{m-1} > \frac{1}{\prod_i P_i}$$

Write $P_i = n_i/N$ where n_i varies proportionately with N.

Then $$\left(\frac{2\pi N}{e}\right)^{m-1} > \frac{N^m}{\prod_i n_i}$$

or $$\frac{N}{\prod_i n_i} < \left(\frac{2\pi}{e}\right)^{m-1} . \qquad (1.15)$$

The reason that I_2 increases with N when (1.15) holds is because the information about the sample size and the relative frequencies decreases in its importance as N becomes large.

To examine the sensitivity of I_2 to N three relative frequency distributions have been studied for various values of N ranging from N = 20 to N = 2000. The situation may be envisaged as one where an experiment X is being performed N times. There are six possible outcomes to each experiment. It is assumed that the relative frequency of occurrence of the different outcomes, $n_i/N, i=1,\ldots 6$, is independent of the the sample size. To determine if the sensitivity of I_2 to N is a function of the shape (i.e. smooth versus peaked) of the relative frequency distribution the measure I_2 has been calculated at various levels of N for three distributions ranging in shape from an almost uniform (A) to a very peaked one (C). The relative frequencies

corresponding to each distribution are expressed in decimal form in Table 1.

DISTRIBUTION	OUTCOME					
	1	2	3	4	5	6
A	0.10	0.20	0.20	0.20	0.20	0.10
B	0.05	0.05	0.10	0.65	0.10	0.05
C	0.00	0.05	0.05	0.08	0.05	0.05

TABLE 1. Relative Frequency Distributions.

The computed values of I_2 for each of the three distributions at different levels of N are listed in Table 2 and shown graphically on Fig. 1.

DISTRIBUTION N	A	B	C
20	0.6132*	0.3995	0.2533
40	0.6683	0.4411	0.2822
60	0.6984	0.4594	0.2952
80	0.7045	0.4700	0.3029
100	0.7131	0.4770	0.3071
120	0.7191	0.4820	0.3117
140	0.7230	0.4858	0.3145
160	0.7272	0.4888	0.3166
180	0.7300	0.4916	0.3184
200	0.7339	0.4932	0.3199
300	0.7398	0.4996	0.3247
400	0.7439	0.5030	0.3273
500	0.7465	0.5053	0.3290
1000	0.7520	0.5103	0.3329
1500	0.7541	0.5121	0.3342
2000	0.7552	0.5131	0.3349

TABLE 2. Brillouin information measures for distributions A, B, and C.

*all the information units are decits.

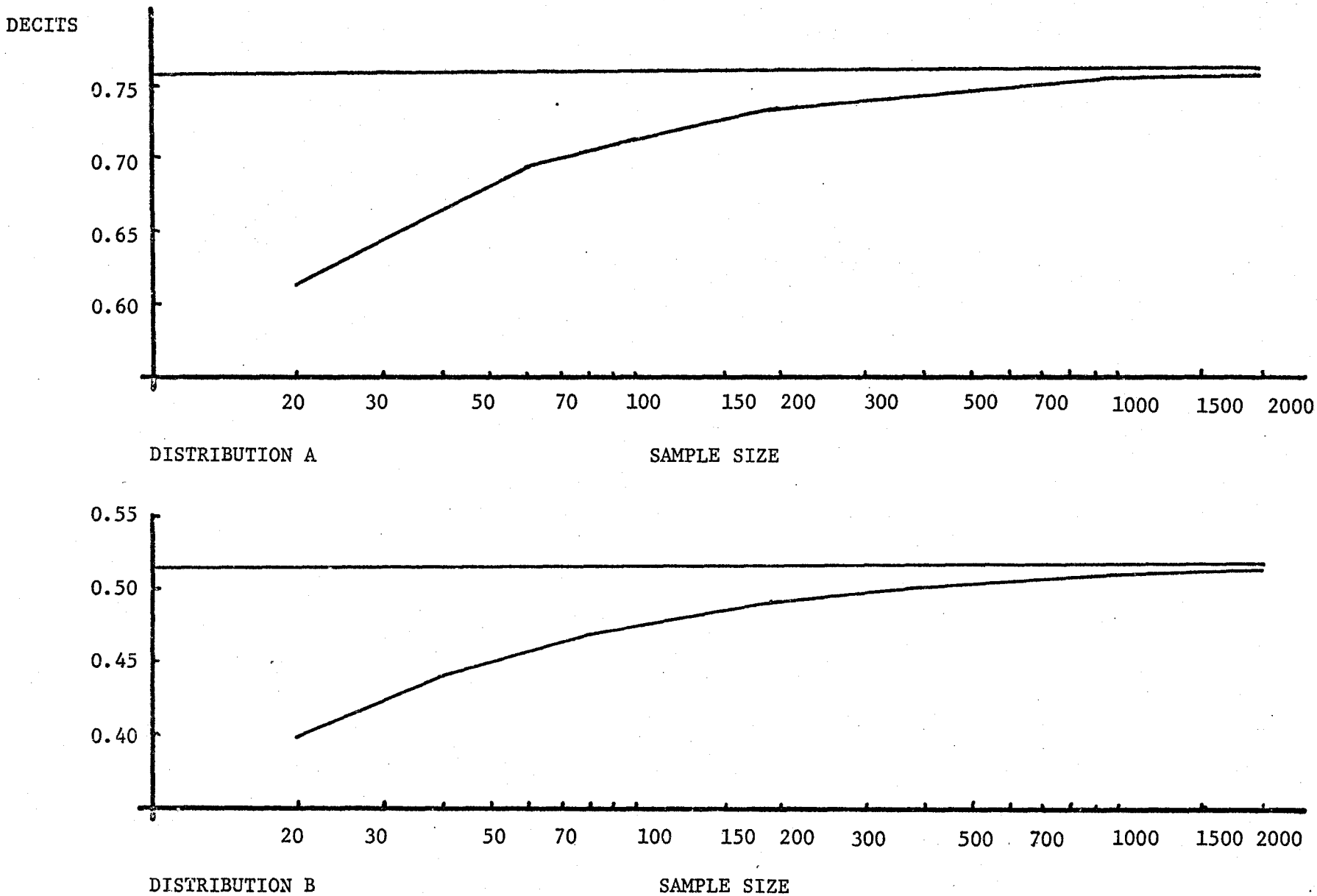

FIGURE 1 Sensitivity of Brillouin's Information Measure to Sample Size, for distributions A and B

The horizontal line in each graph is the measure obtained from Shannon's formula.

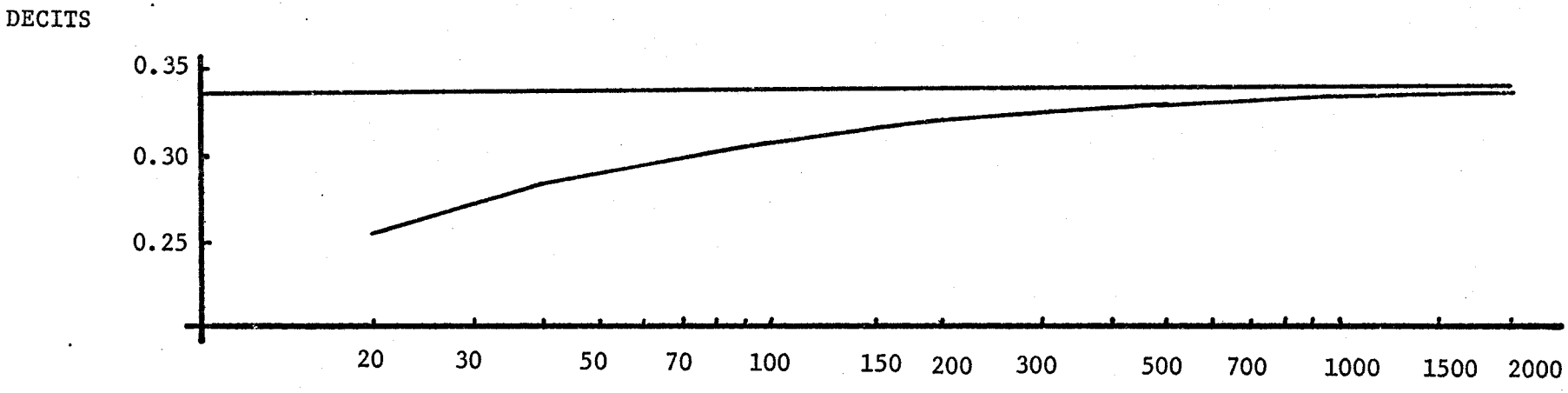

FIGURE 1 cont'd. DISTRIBUTION C SAMPLE SIZE

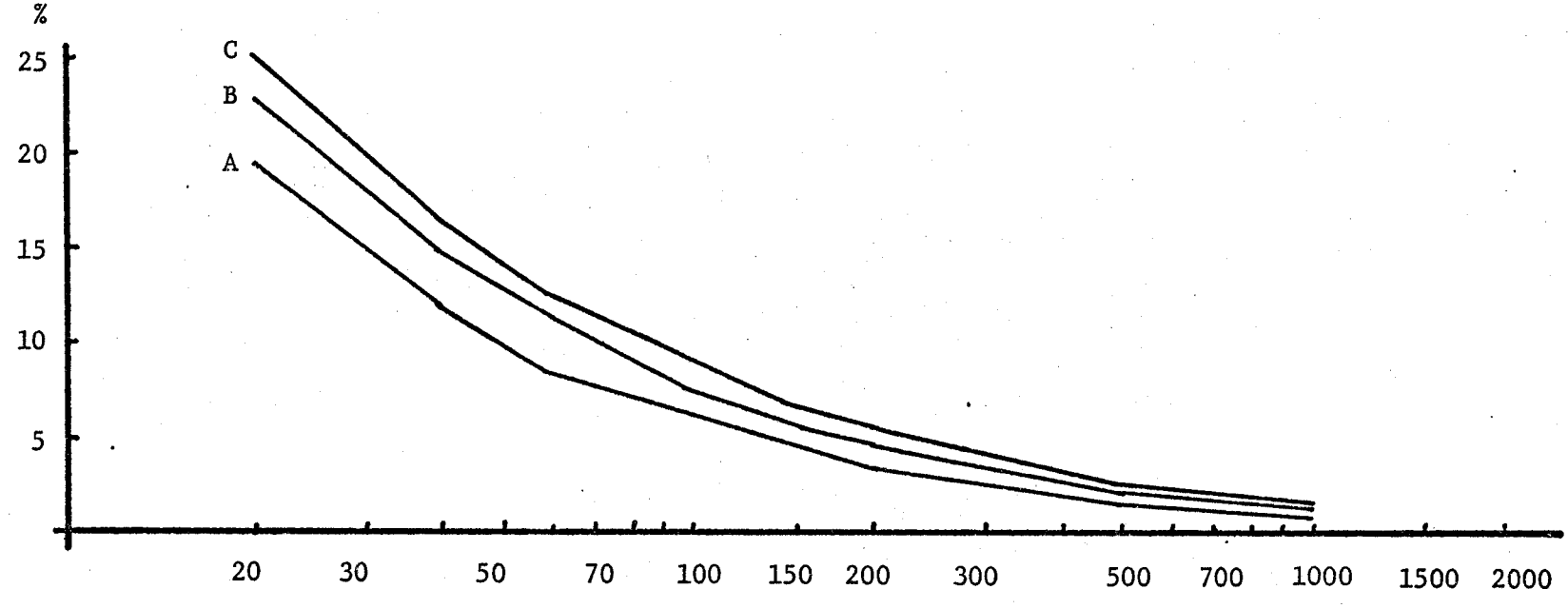

FIGURE 2 Structural Information Indices for distributions A, B and C.

From these data it is clear that I_2 is a monotonically increasing function of log N, rising rapidly at first for N up to about 200 and then levelling off to converge to I_1. To determine how the sensitivity of I_2 to N is affected by the shape of the distribution it is necessary to establish a method for comparing the results obtained from each distribution. It has been shown that for most distributions $\hat{I}_1 > I_2$, (1.11) above, and that $\hat{I}_1$ is unaffected by varying the sample size if the relative frequency distribution is held constant. By utilizing these properties of $\hat{I}_1$ it is possible to define a structural information coefficient, S.I., for each distribution at various levels of N. By the structural information content of a population is meant the information about its size, N, and the relative frequency of occurrence of different types of events in it. For comparative purposes the coefficient may be expressed in percentage terms as

$$S.I. = \left(\frac{\hat{I}_1 - I_2}{\hat{I}_1}\right) \times 100\% \quad , \quad \hat{I}_1 \neq 0. \tag{1.16}$$

Taking the relative frequencies as subjective probabilities the Shannon measure I_1 has been calculated for each distribution. Here it is assumed that the probabilities of the different outcomes remain the same for each experiment. At all levels of N the values of I_1 for the three distributions are

$$I_1^A = 0.7572 \quad , \quad I_1^B = 0.5168 \quad , \quad I_1^C = 0.3377 \quad .$$

where the superscripts A, B and C refer to the corresponding distribution. Using these values of I_1 the structural information coefficient for each of the distributions has been calculated at different levels of N. The

results are listed in Table 3 and appear on Figure 2. The results indicate clearly that the sensitivity of I_2 to N is a function of the shape of the distribution. Specifically, the more peaked is the distribution the greater the sensitivity of I_2 to N. The structural information will be at least 5% in samples of size up to 130 when the distribution is nearly uniform, like A, and in samples of size up to 200 when the distribution is peaked, like C above.

DISTRIBUTION N	A	B	C
20	19.23	22.69	24.99
40	11.97	14.64	16.43
60	8.27	11.10	12.58
80	7.20	9.05	10.30
100	6.08	7.70	9.05
160	4.21	5.42	6.25
200	3.33	4.56	5.27
500	1.67	2.21	2.57
1000	0.93	1.06	1.42

TABLE 3. Structural Information Coefficients.

The interpretational significance of these properties of I_2 can be illustrated by some examples. Suppose some experiment X is performed N times. There are five possible outcomes to each experiment. In one sequence of twenty experiments each of the five outcomes occur equally often. For this sequence

$$I_1^{(1)} = 0.6990 \quad , \quad I_2^{(1)} = 0.5743 \ .$$

In a second sequence of twenty experiments it is observed that only four different outcomes occur and that these are equiprobable. The corresponding information measures are

$$I_1^{(2)} = 0.6021 \quad , \quad I_2^{(2)} = 0.5035 \quad .$$

In a third sequence of two hundred experiments it is again observed that there are only four different equiprobable outcomes. The information measures are

$$I_1^{(3)} = 0.6021 \quad , \quad I_2^{(3)} = 0.5848 \quad .$$

If it were assumed that I_1 and I_2 were both measures of our mean uncertainty about the outcome of any experiment (without specifying the nature of the sample data), then $I_1^{(2)} < I_1^{(1)}$ and $I_2^{(2)} < I_2^{(1)}$ would be intuitively reasonable results. So also would be the result $I_1^{(3)} < I_1^{(1)}$, but $I_2^{(3)} > I_2^{(1)}$ would appear to be counter-intuitive, since it would imply that our uncertainty as to the outcome of any experiment is less after performing twenty experiments from which there are five equiprobable outcomes than after performing two hundred experiments which indicate that there are four equally likely outcomes to any one experiment. Given the possibility of results of this type Peet (1974) suggested that Brillouin's measure would be unacceptable for measuring heterogeneity in ecological distributions.

If, however, the interpretations given to I_1 and I_2 in this paper are adopted, then the seemingly internal inconsistencies in the results from I_2 vanish. When I_2 is interpreted as a measure of our mean uncertainty about some event in a completely sampled population of events

and I_1 is interpreted as a measure of our mean uncertainty about some event in a large population of events, from which we have only a sample then all the information measures calculated above vary in the way that one would expect them to. The reason why the result $I_2^{(1)} < I_2^{(3)}$ is obtained is because the structural information in the first sequence is of greater importance, since the sample is small. As the sample size increases the structural information declines in importance, and hence I_2 increases.

The fact that I_2 increases with N may be a disadvantage if it is desired to compare the information measures obtained from different studies. In ecological situations, however, Pielou (1967) has argued that the information or uncertainty measure (which she uses to measure diversity) should be a function of the sample size, and hence she advocates the use of Brillouin's measure. The sample size effect may be overcome to a large extent by using a relative information index, rather than the actual information measure. The proposed relative information index is

$$RI_2 = I_2/I_{2\,Max} \tag{1.17}$$

$$\text{where } I_{2\,Max} = \frac{1}{N} \log \frac{N!}{\{[\frac{N}{m}]!\}^{m-r} \cdot \{([\frac{N}{m}] + 1)!\}^{r}} \tag{1.18}$$

and N is the sample (population) size,
m is the number of different outcomes observed,
[N/m] is the integer part of N/m,
$r = m\,(N/m - [N/m])$.
When N is an exact multiple of m, r = 0.

The index RI_2 ranges between zero and unity. When $RI_2 = 0$ we obtain no information from a message, because we were already certain about its contents. $RI_2 = 1$ when we are maximally uncertain about the content of a particular message. Hence the index RI_2 may be used to provide an indication of the importance of the constraints that are operating in a particular experimental environment, (Gurevich, 1969a, b).

The advantage of using the relative information index in conjuction with the actual information measure may be illustrated by the following example. In a population of twenty events five different types of event are represented and each occurs four times. In this case our mean uncertainty about the type of any event is $I_2^{(1)} = 0.5743$. In another population of 200 events it is observed that four different types of event each occur 38 times, while a fifth type of event occurs 48 times. Our mean uncertainty about the type of any event in this population is $I_2^{(2)} = 0.6746$. $I_2^{(2)}$ is greater than $I_2^{(1)}$ because the structural information in the second sample is less useful to us. In the first population there is a uniform distribution of the different types of events, while in the second population the distribution is almost uniform. This difference in the uniformity of the distributions is reflected by the relative information indices, which are $RI_2^{(1)} = 1.000$ and $RI_2^{(2)} = 0.99697$ respectively. Hence the relative information index provides a useful tool for comparing the internal uniformity or homogeneity of completely sampled populations.

Clearly for large populations, from which we may have only sample data to use when estimating the probabilities of the different types of events occurring, the relative information index is

$$RI_1 = I_1/I_{1\,Max} \tag{1.19}$$

where $I_{1\,Max} = \log m^*$

and m^* is the number of different types of events that occur in the total population of events.

Good's Information Measure

In the Shannon estimator, $\hat{I}_1$, an objectivist is assuming that n_i/N provides a good estimate of the true population probability P_i. This will only be the case if each of the n_i are large. Furthermore, in the Shannon and Brillouin measures it is assumed that the number of outcomes that can occur from an experiment X is known a priori. Oftentimes this may not be the case. Good (1953) has provided an information measure for these circumstances. His approach can be summarised as follows.

Take a random sample from the outcomes of an experiment X. Let S_r denote the number of outcomes that occur r times in the sample such that $\sum_r S_r = N$, where N is the sample size. Let q_r be the unknown population proportion of an arbitrary outcome that occurs r times in the sample. Then Good has proved that Shannon's measure I_1 is given exactly by

$$I_3 = \frac{1}{N}\sum_r E(S_r)\left\{\frac{1}{r+1} + \frac{1}{r+2} + \ldots + \frac{1}{N-r} - \frac{d}{dr}\log E(S_r) - E(\log_e (1-q_r))\right\} \tag{1.20}$$

Here $I_3 = I_1$ is the mean amount of P.I. obtained from any outcome of an experiment conducted a very large number of times. To estimate I_3 from sample data it is necessary to "smooth" the sequence $S_1, S_2, \ldots S_r, \ldots$ and replace it by the sequence $S_1', S_2', \ldots, S_r', \ldots$; see Good (p. 242-3) for a description of the smoothing methods. Using these smoothed values

then as an estimator of I_3 or I_1 we can write

$$\hat{I}_3 = \log N - \frac{1}{N} \sum_r r\, S'_r \left(g_r + \frac{d}{dr} \log S'_r\right) \tag{1.21}$$

where $g_r = \sum_{j=1}^{r} \frac{1}{j} - \gamma$

and $\gamma = 0.577215 \ldots\ldots$, is the Euler Mascheroni constant.

For large values of r the factor $\left(g_r + \frac{d}{dr} \log S'_r\right)$ may be replaced by log r to a good approximation. Good (1953) has provided some examples to illustrate the steps involved in computing this measure.

Non Probabilistic Measures of Information

The three information measures, I_1 I_2, I_3, discussed above are based on probability notions. In fact until 1962 it was generally accepted that information theory was a branch of probability theory, see for instance Good (1956), Khinchin (1957). In 1962 the mathematical foundations of information theory were defined without the use of probability notions by Ingarden and Urbanik, thereby showing that the dependence of information on probability could be reversed and that the two notions are equivalent in the abstract. They have demonstrated the formal equivalence of the abstract notions of probability and information within a measure theoretic framework. This approach has been further developed in a series of subsequent papers, for a complete set of references see Urbanik (1972). In this paper, however, the discussion will be limited to applications of the probabilistically defined information measures given earlier.

The Appropriate Measure In Different Geographic Situations

The appropriate measure to apply to a particular set of data will depend on whether the data represent, on the one hand a population in itself or, on the other hand, a random sample from some conceptually larger population. If the events being studied form a random sample from some larger population then the appropriate measure to use is Shannon's, I_1, or its estimate $\hat{I}_1$. The measure that is calculated represents our mean uncertainty about any event in the total population, relative to what is known. On the other hand, if the events being studied represent a complete population of events then the appropriate measure is Brillouin's, I_2. The measure that is calculated then represents our mean uncertainty about any event in the population relative to what we know about the population. For large, completely sampled, populations of events Brillouin's measure may be approximated by $\tilde{I}_2$. $\tilde{I}_2$ will always be an overapproximation and for populations smaller than 200 the approximation will be greater than the true value by at least 5%, unless the distribution is completely uniform, cf data in Table 2. Hence for small populations the mean information measure should be calculated directly from I_2. Generally as the population becomes very large one can only collect data on a sample from it. The information measure to be used in this situation is $\hat{I}_1 = -\sum_i {}^{n}i/N \log {}^{n}i/N$ which provides an estimate, albeit a biased one, of the mean information content of each event in the population. The variance of the estimate obtained from any sample can be calculated from (1.7) and the estimates obtained from different samples may be compared by Hutcheson's (1970) method. If in a random sample from a larger population some types of events are only rarely represented or, perhaps, it is felt that types of events that occur in the

population are not represented at all in the sample, then the appropriate information measure is Good's, I_3, or its estimator, $\hat{I}_3$.

The following example may help to illustrate the type of situation for which each of the measures is appropriate. Suppose we wish to measure the diversity of urban land use patterns. We equate the diversity of a pattern with the amount of information needed to describe it, and hence the more diverse the pattern the more information that will be needed to describe it. We study one large urban region, say Toronto. Denote the total area of the urban region by N, the number of landunits devoted to each landuse i by n_i such that $\sum_i n_i = N$. If we are only interested in the diversity of the Toronto landuse pattern then we use Brillouin's measure, I_2, and calculate our mean uncertainty about the landuse type that occurs in any given land unit. If our interest lies in estimating the diversity of urban landuse patterns in general then we would use Shannon's estimate $\hat{I}_1$ where the probabilities are estimated by $\hat{P}_i = n_i/N$. This measure estimates our mean uncertainty about the landuse type that will occur in any landunit of any urban region.

Suppose we are interested in measuring the diversity of the horticultural or market gardening zone found around many urban regions. We could take a sample region and calculate, say, the number of landunits devoted to the cultivation of each plant or vegetable. We may find that in our sample region some plants or vegetables are not widely grown. This factor may cause us to doubt if we have considered all the plants that are likely to be grown in these types of areas. To estimate the plant and vegetable diversity of these regions in general the appropriate measure to use is Good's estimate, $\hat{I}_3$.

The advantages of using $\hat{I}_3$ are that it does not require that the population proportions of land devoted to each plant i, say, be estimated by n_i/N and it is not necessary to know all the different plants that occur in these regions in general. A disadvantage of $\hat{I}_3$ is that it requires a large sample, since large values of S_r are required to permit acceptable smoothing of the sequence $S_1, S_2, \ldots, S_r, \ldots$.

Multivariate Information Measures

The three information measures (Shannon's, Brillouin's and Good's) discussed above have been defined strictly for univariate distributions. In geographic applications of these measures it is more likely that we will be dealing with phenomena characterized by more than one variable. Without assuming that the variables are independent we can define a multivariate counterpart of I_1 and I_2.

Suppose we have a collection of n objects, each characterized by a vector containing s entries. The s entries in each vector represent the values assumed by a particular object on s variables $X_1, X_2, \ldots, X_s$. It is assumed that each object is characterized by the same s variables. For example, the objects could represent census tracts within an urban area and the s variables could represent s different landuses types that occur in the urban area. Then each census tract would be characterized by a vector with each entry representing the number of units of a particular landuse that occur in that tract. Alternatively, the objects could be individuals with the variables representing questions (to which there are a fixed number of possible responses) designed to determine, say, some activity patterns of people. Each individual would be characterized by a vector, each entry in it

representing the response of the individual to a particular question, where the responses are coded nominally.

When the n objects form a random sample from a much larger population of objects the mean amount of information needed to select the vector that characterizes a particular object in the population can be shown to be

$$I_1(X_1, X_2, \ldots, X_s) = I_1(X_1) + I_1(X_2|X_1) + I_1(X_3|X_1, X_2) + \ldots + I_1(X_s|X_1, X_2, \ldots X_{s-1}). \quad (1.22)$$

$$= - \sum_{k_1 k_2 \ldots k_s}^{r_1 r_2 \ldots r_s} P(x_{1k_1}, x_{2k_2}, \ldots, x_{sk_s}) \cdot \log P(x_{1k_1}, x_{2k_2}, \ldots, x_{sk_s}) , \quad (1.23)$$

where for any ℓ, $2 \leq \ell \leq s$

$$I_1(X_\ell|X_1, \ldots, X_{\ell-1}) = - \sum_{k_1 k_2 \ldots k_\ell}^{r_1 r_2 \ldots r_\ell} P(x_{1k_1}, x_{2k_2}, \ldots x_{\ell k_\ell}) \log P(x_{\ell k_\ell}|x_{1k_1}, x_{2k_2}, \ldots, x_{(\ell-1)k_{(\ell-1)}}) \quad (1.24)$$

In the special case when the s variables are mutually independent (1.22) reduces to

$$I_1(X_1, X_2, X_3, \ldots, X_s) = \sum_{i=1}^{s} I_1(X_i) , \quad (1.25)$$

for proof see Appendix B.

In words formula (1.22) may be interpreted as follows. The mean amount of information needed to select the s values that characterize any object is equal to the mean amount of information needed to select the first value plus the mean amount needed to select the second value given the first one, plus the mean amount needed to select the third value given the first and second values, and so on. When the variables are independent the mean amount of information needed to select the s values in any vector is equal to the sum of the mean amounts needed to select each value, as indicated by (1.25).

If the n objects being studied form a complete population then the mean amount of information needed to select the vector that characterizes a particular object is given by

$$I_2(X_1, X_2, \ldots, X_s) = I_2(X_1) + I_2(X_2|X_1) + I_2(X_3|X_1, X_2) + \ldots\ldots, + I_2(X_s|X_1, X_2, \ldots, X_{s-1}). \quad (1.26)$$

$$= \frac{1}{n} \log \Big[n! \Big/ \prod_{k_1 k_2 \ldots k_s}^{r_1 r_2 \ldots r_s} \Big(\bigcap_{i=1}^{s} n_{ik_i} \Big)! \Big] \quad (1.27)$$

where for any ℓ, $2 \leq \ell \leq s$,

$$I_2(X_\ell|X_1, X_2, \ldots, X_{\ell-1}) = \frac{1}{n} \Big[\sum_{k_1 k_2 \ldots k_{\ell-1}}^{r_1 r_2 \ldots r_{(\ell-1)}} \log \Big(\bigcap_{i=1}^{\ell-1} n_{ik_i} \Big)! - \sum_{k_1 k_2 \ldots k_\ell}^{r_1 r_2 \ldots r_\ell} \log \Big(\bigcap_{i=1}^{\ell} n_{ik_i} \Big)! \Big], \quad (1.28)$$

n_{ik_i} is the number of times the value x_{ik_i} occurs for the variable X_i, $k_i = 1, 2, 3, \ldots, r_i$; $i = 1, 2, \ldots, s$, and $n_{ik_i} \bigcap n_{jk_j}$ is the number of joint occurrences of the value x_{ik_i} for X_i and $x_j k_j$ for X_j such that

$$\prod_{i=1}^{s} \{ \sum_{k_i=1}^{r_i} (\bigcap_{i=1}^{s} n_{ik_i}) \} = n \ .$$

When the s variables are mutually independent (1.26) reduces to

$$I_2(X_1, X_2, \ldots, X_s) = \sum_{i=1}^{s} I_2(X_i) \ , \tag{1.29}$$

for proof see Appendix B.

When all the n_{ik_i} are large $I_2(X_1, X_2, \ldots, X_s)$ is a good approximation of $I_1(X_1, X_2, \ldots, X_s)$, (Lemma 1, Appendix A).

It was suggested in the univariate case that the sensitivity of Brillouin's measure to the population size could be overcome to a large degree by using a relative information index. In the multivariate case the relative information index is

$$\frac{I_2(X_1, X_2, \ldots, X_s)}{I_2(X_1, X_2, \ldots, X_s)_{Max}} \tag{1.30}$$

The maximum value, $I_2(X_1, X_2, \ldots, X_s)_{Max}$ occurs when the s variables are mutually independent and the distribution of values on each is uniform. Therefore,

$$I_2(X_1, X_2, \ldots, X_s)_{Max} = \sum_{i=1}^{s} \frac{1}{n} \log \frac{n!}{\{[\frac{n}{r_i}]!\}^{r_i - \tilde{r}_i} \cdot \{([\frac{n}{r_i}] + 1)!\}^{\tilde{r}_i}} \quad (1.31.)$$

where r_i is the number of possible values for each variable X_i, i=1,2,...s.

$[\frac{n}{r_i}]$ is the integer part of $\frac{n}{r_i}$.

$$\tilde{r}_i = r_i(\frac{n}{r_i} - [\frac{n}{r_i}])$$

When n is an exact multiple of each r_i , $\tilde{r}_i = 0$, all i, and (1.31) simplifies to $\sum_{i=1}^{s} \log r_i$, for large n .

Information, Entropy, Order

In many applications of information statistics the expression $I_1 = -\sum_i P_i \log P_i$ has been called a measure of the entropy (Semple & Golledge 1970), order (Medvedkov, 1967a, b), or information received from a distribution (Williams and Lambert 1966a, Marchand 1972, 1975). These terms have been used interchangeably, although the relationship between information, entropy and a third concept, negentropy, has been a controversial one, (Rothstein 1951, 1952, Tillman and Russell 1961, Brillouin 1962, 1964, Marchand 1972, Nauta 1972). Part of this controversy arises from the confusion caused by using I_1 as a definition of information (Brillouin 1956, Chp. 1 and p. 265-67) and a failure to maintain a clear distinction between our prior information about the system we are studying and the information we obtain from experimenting with the system.

The statistical measure of the entropy of a closed system is given by $S = - k \sum_{i=1}^{n} P_i \log P_i$ where n is the number of states that the system can be in, P_i is the probability of the system being in the i^{th} state, and k is some constant. When all the P_i are equal S attains its maximum value. These probabiliites, P_i, can be calculated in such a way that they reflect all our knowledge about the system before carrying out any experiments on it, (Jaynes 1957, 1968, Tribus 1969). Then I_1 or S measure the mean amount of information we need to select the outcome of any experiment, or the mean information content of the outcome of any experiment, by (1.3b) and (1.3d) above.[2] When our prior knowledge or information about the system is minimal we can only legitimately assume that all the outcomes to an experiment are equiprobable. Such a system is said to lack order. In this case the entropy, S, of the system is at a maximum and the mean amount of information needed to select the outcome of any experiment on the system is also at a maximum. Conversely, when our prior information is large some of the outcomes to an experiment are more likely than others and therefore the entropy and the mean amount of information needed to select the outcome of any experiment are reduced. Such a system displays some order. Therefore, by increasing the prior (e.g. prior to carrying out some experiment) information about a system the entropy and the mean amount of information needed to select the outcome of any

---

[2] There is still some controversy among physicists and engineers as to whether the entropy of information theory and the entropy of thermodynamics are the same concepts. Ter Haar (1956), Jauch and Baron (1972) and Skagerstam (1975) argue that they are not equivalent, while Tribus (1961, 1966) has argued that information theory provides a basis for thermodynamics.

experiment on it are decreased. In the literature, however, it has not always been made clear that the information we must increase to reduce the entropy of a system is our prior information about the system before experimenting with it. This information can be increased by either imposing order (e.g. by simplifying the system) or by discovering order already existing in the system. Brillouin (1956, 1962, 1964) has erroneously equated increases in information with decreases in entropy by failing to maintain a clear distinction between the information we possess about a system prior to experimenting on it and the information we obtain from each experiment. He suggested that information I be equated with negentropy, N, where $I = N = -S$. A similiar line of reasoning has been followed by Somenzi (1962) and Bell (1967).

However, if information is equated with negentropy, where negentropy is taken to be the negative of entropy, this implies that information is being measureed in negative quantities, clearly an undesirable property. A more appropriate definition of negentropy is $N = S_{Max} - S$ or $N = I_{1\,Max} - I_1$. In this definition negentropy measures the amount of information we possess about a system before performing an experiment to select the next state it will be in, and N is always non-negative. When $N = 0$, the prior information is minimal, then the most unbiased assertion we can make is that each of the states the system can be in are equiprobable, (Jaynes 1957). When $N = I_{Max}$ the prior information is maximal and we can predict with certainty the next state the system will be in. The entropy expression, S, or I_1, measures the mean information content of any experiment on the system,

or the mean amount of information needed to select the outcome of any experiment from the set of possible outcomes. Therefore, in short, the more disorder that exists in a system the larger will be its entropy and the smaller its negentropy or the amount of prior information available about it. Conversely, when the system displays more order its entropy is smaller and its negentropy larger.

If the system under examination is not well defined then care should be exercised when equating entropy with the degree of order or organization in the system (Klein 1953, Landsberg 1961). For instance, in Chapman (1970, 1973) the entropy does not depend explicitly on the interactions between the points. It is in these interactions, however, that the structure of the system is reflected and it is this structure which is associated with the concept of organization. When information measures are employed to measure the degree of order in a system it should always be clearly stated what kind of order is being measured, i.e. is it biological, physical, sociological, etc? For example, Quastler (1964) offers different estimations of the information content of a specific bacterium to measure its physical, chemical and biochemical order (cf Nauta 1972 p. 262). Unfortunately, at times we read unqualified statements to the effect that entropy measures randomness or disorder in a system. Such unqualified statements are meaningless; the important question is what kind of randomness or disorder does the entropy measure in a particular situation. Every attempt should be made to insure that the system under examination in any study should be as well defined as possible, particulary if the entropies of different

systems, or of the same system at different points in time, are to be compared. In any study, perhaps, this is the most difficult problem of all - to learn how to state clearly: What is the specific question we are trying to answer?

Finally, it is necessary to reiterate the viewpoint of Jaynes (1965) that it is incorrect to consider entropy or any information statistic as an objective measure, as Orloci (1968, 1970) does. The entropy or information measure is an anthropomorphic concept not only in the statistical sense that it measures our uncertainty as to the state of a system, but also in the phenomenological sense because it is a property not of the system, but of the particular experiments one chooses to perform on it.

The Value of Information

Some workers have expressed their concern about the value or usefulness of the information we obtain from information theoretic measures, Belis and Guiasu (1968), Marchand (1972, 1975), Behara (1974). The point at issue here is, of course, the non-equivalence of transmissional (Shannon) and semantic information. As Shannon indicated when introducing his mathematical theory of communication the semantic aspects of communication are irrelevant to the engineering aspects, with which he is concerned. For him "the word information in communication theory relates not so much to what you say as to what you could say. That is, information is a measure of one's freedom of choice when one selects a message" (Shannon and Weaver 1949, p. 100). In information theory the information content of a message is a function of the

inprobability of its occurrence. The usefulness or relevance of that information for the receiver is ignored. These qualitative aspects of the information will be related to its utility for the fulfillment of some goal. A modification of the Shannon information measure to incorporate the qualitative aspects as well as the quantitative aspects of the information content of a message has been proposed by Belis and Guiasu (1968).

Let E_1, E_2, E_3, ..., E_n be a set of events, representing the possible outcomes of some experiment X having the probabilities P_1, P_2, ..., P_n and utilities u_1, u_2, ..., u_n respectively. The potential information content of any event of probability P and utility u, being dependent on both variables is written as I(u,P). It is reasonable to expect that the information measure I(u,P) should satisfy the following two criteria - (1) the potential information content of any two independent events must be equal to the sum of the potential information content of each event separately and (2) the potential information content of an event having a certain probability P must be directly proportional to the utility of that event. Belis and Guiasu have shown that an information measure satisfying these two criteria is

$$I(u,P) = - k\, u \log P \tag{1.32}$$

where k is some arbitrary constant.

It can be shown that the mean information content of any experiment X having as possible outcomes E_1, E_2, ..., E_n with the probabilities P_1, P_2,..., P_n and utilities u_1, u_2, ..., u_n respectively is

$$I_4 = - k \sum_i u_i P_i \log P_i \, . \tag{1.33}$$

If all the outcomes have utility equal to unity and k = 1, then $I_4 = 1$. This information measure has recently received further attention from Skala (1974).

While this attempt at measuring the qualitative aspects of information appears attractive at first sight, it is clear that it depends on an individual's ability to set up a numerical distribution of values on different events. For a start this would probably limit the number of possible events or outcomes to six or seven (Miller 1956). Furthermore, if the amounts of potential information received by different individuals from an event were to be compared it is likely that the utilities would have to be measured on a ratio scale. It is very doubtful if we could reasonably make that assumption, Simon (1957), Shepard (1964). For more elaborate discussions of the various types of information, see in particular McKay (1950, 1956, 1965) and Nauta (1972).

Summary and Conclusions

The purpose of this chapter was to outline and clarify some of the basic concepts of information theory that will be necessary for an understanding of the remainder of this paper. In this chapter the formal equivalence of the concepts of information and uncertainty has been demonstrated. Three information measures have been examined in detail and an attempt has been made to identify the particular type of situations to which each should be applied. Most of the controversy in the literature has been over the question of when to use Shannon's or Brillouin's information measure. By identifying the process which

gives rise to Brillouin's measure it is clear that this measure applies strictly to completely sampled populations, where the population size and the relative frequencies of different types of events within it are known. Shannon's measure is defined for conceptually large populations. The prior probabilities about the types of different events are defined in some way external to the events being studied. From the conceptual population we usually have some sample data which is used to define the probabilities of different types of events in the population. The Shannon and Brillouin measures have been generalized to accommodate objects characterized by a number of nonindependent variables. The relationships between the concepts of information, entropy and order have been examined and hopefully clarified. Part of the confusion that has surrounded these problems has been due to a failure to distinguish between two different types of information - the information we possess before we perform an experiment and the information we can get from the experiment. The former type of information corresponds to negentropy while the latter type corresponds to the entropy of the system on which we are experimenting. Caution is urged when applying the concept of entropy to describe social or economic systems which may not be well defined.

CHAPTER 2

CLASSIFICATION : INFORMATION-THEORETIC APPROACHES.

Introduction

The objective of any classification exercise is to impose some order and coherence on the vast inflow of information we obtain from the real world. As Harvey (1969, p. 326) put it "By grouping sense perception data into classes or sets we transform a mass of unwieldy information so that it may be more easily comprehended and manipulated". Our ability to classify things is one of the basic tools we use when dealing with the world around us.

In general, the classification problem may be stated as follows: given a set of N objects (areal units, towns, individuals) and for each object a set of M attributes, form a partition of the M-dimensional vector space within which each of the objects may be represented as a point, such that the objects within each class or region of the vector space are maximally similar to one another, in terms of some given definition of similarity. In the terminology of information theory the problem may be considereed to be the partitioning of the set of objects into classes such that ones uncertainty about the class a given object belongs to is maximal, while ones uncertainty about the attributes of that object, given the class it is in, is minimal.

The different classificatory strategies have been reviewed in some detail by Williams (1971). Here attention will be focussed only

on those classificatory strategies that are used most frequently, namely those labelled "exclusive, intrinsic, hierarchical" by Williams (1971). These classification strategies are exclusive in the sense that a given object occurs in one class and one class only; the population is divided into a set of mutually exclusive subclasses which nowhere overlap in their membership. They are intrinsic in the sense that all the attributes used are regarded as equivalent. These classifications are hierarchical in the sense that they always optimize a route between the entire population and the set of individuals of which it is composed (Williams 1971). Hence a characteristic of these types of classification is that they always optimize some objective function. The algorithms of Johnson (1967), Ward (1963), and Wishart (1968) are of this type.

In hierarchical classifications several strategies are available. On the one hand, the classification strategy may be agglomerative or divisive. An agglomerative strategy is one that proceeds by progressive fusion, beginning with the individuals and ending with the complete population (e.g. Berry 1965). This procedure is simply called "classification" by Grigg (1965, 1967). A divisive strategy progressively splits the population into classes of diminishing size by the process termed "logical division" by Grigg (1965). On the other hand, the classification strategy may be either monothetic or polythetic. In a monothetic classification every class at every stage is definable by the presence or absence of specified attributes. In a polythetic classification the classes are defined by their general overall

similarity of attribute structure.

In practice the choice of method is usually between divisive monothetic or agglomerative polythetic, since agglomerative monothetic methods can only exist in a trivial sense and the divisive polythetic methods that have been devised usually take too long to execute on standard computers. The most rigorously constructed method, perhaps, is that of Edwards and Cavalli-Sforza (1965) which examines all dichotomous choices at each division. However, since for N individuals this involves the examination of $(2^{N-1}-1)$ possibilities at each division, the method is usually computationally impracticable for large values of N. For instance Gower (1967) states that for N = 41 the process would require more than 54,000 years on a computer with 5 μ sec. access time.

An important characteristic of a divisive procedure is that the final classification is totally dependent on the order in which the attributes are selected to effect division at each stage. In developing classifications of this type, therefore, it is necessary to select the attributes in some order of significance. This of necessity assumes that we know a good deal about the objects being classified; or, in other words, that we possess some theory to help us to identify the important attributes of these objects. Alternatively, the attribute that gives the optimal split at each stage in the classification will be used and this attribute will be determined through the objective function that is to be optimized.

Agglomerative methods have the advantage that they are faster to execute. A disadvantage of an agglomerative classification, however, is that it is inherently prone to a small amount of misclassification, the ultimate cause of which is that the fusion process begins at the bottom of the hierarchy, at the interindividual level, where the possibility of error is greatest (Lambert 1972). The distortion of the data progressively increases up the hierarchy and, therefore, the distortion is greatest in the uppermost classes, which are usually the ones that one is interested in. Divisive classifications are less likely to suffer from this disadvantage since they start at the top of the hierarchy and progressively subdivide the population into classes.

In this chapter existing information - theoretic approaches to classification are reviewed and an indication is given of how they can be applied in a geographic context. In terms of the agglomerative /divsive and monothetic/polythetic dichotomies the procedures to be examined may be classified as follows;

DIVISIVE		AGGLOMERATIVE	
MONOTHETIC	POLYTHETIC	MONOTHETIC	POLYTHETIC
(i) INFORMATION ANALYSIS (WILLIAMS 1968d)	(i) RESCIGNO AND MACCACARO (1960)	(i) THEIL (1967) SEMPLE (1972) BATTY (1974)	(i) THEIL (1967)
	(ii) WALLACE AND BOULTON (1968)		(ii) INFORMATION ANALYSIS (WILLIAMS 1966)
			(iii) CONTINGENCY TABLE ANALYSIS (KULLBACK 1959) ORLOCI (1968, 1970)

The methods used by each of these authors will be examined separately. It will be indicated that the method of information analysis due to MacNaughton-Smith (1965) and Williams and his associates (cf. list of references) is the most widely used information based classification algorithm. For this reason it will be compared with standard classification procedures to indicate some of its advantages and disadvantages. Finally, in this chapter, the problem of constructing spatial-temporal classifications will be examined.

Divisive Monothetic Classification

In a divisive classification one seeks to dichotomize the population at each stage in some way that maximizes the dissimilarity of the resulting classes. When the initial collection of objects is divided into subclasses there is a reduction in the amount of information needed to describe each object; or, in other words, there is a reduction in our uncertainty about the attributes of a given object. The more dissimiliar are the subclasses the greater is the reduction in our uncertainty.

Suppose we have a collection of n objects, each characterized by a vector containing s entries, corresponding to the scores of the object on s variables. Assume the variables are statistically independent - if they are not they may be orthogonalized via principal components analysis or factor analysis. Let n_{ij} denote the score of the j^{th} object on the i^{th} variable, and let $n_i(= \sum_j n_{ij})$ be the total score on the i^{th} variable such that $\sum_i n_i = n$, the total sum of the scores on all the variables. For example in a landuse classification

the n objects to be classified could be some basic spatial units (e.g. census tracts or blocks of a city, counties of a country) and the s variables could correspond to s different landuse types. Then n_{ij} would be the area devoted to landuse i in the $j^{\underline{th}}$ spatial unit, n_i would be the total area devoted to landuse i in a whole city or country and n would be the total area of the city or country. The information content, or the amount of information needed to describe, the unclassified objects may be measured by

$$I_S = -\sum_{i=1}^{s} \left[n_i \sum_{j=1}^{n} \frac{n_{ij}}{n_i} \log \frac{n_{ij}}{n_i} \right]$$

$$= \sum_{i=1}^{s} \left[n_i \log n_i - \sum_{j=1}^{n} (n_{ij} \log n_{ij}) \right] \qquad (2.1)$$

$$\text{or} \quad I_B = \sum_{i=1}^{s} \left[\log n_i! - \sum_{j=1}^{n} \log (n_{ij})! \right] \qquad (2.2)$$

depending on whether Shannon or Brillouin's measure is used. If Shannon's measure is used then it is assumed that the objects being classified form a random sample from some conceptual population and hence the classification produced can be considered to be a representation of the way the objects in the conceptual population should be classified. If the collection of objects being classified is considered to be a completely sampled population then by the arguments in the previous chapter an information measure based on Brillouin's should be used. For this reason, throughout the remainder of this chapter Shannon and Brillouin based measures will

be presented together, they can be distinguished by the subscripts s and B respectively. Similary, when we are not sure that our sample collection of objects reflects all the sources of variability in the parent population an information measure based on Good's, equation (1.21) above, should be used. From (2.1) and (2.2) we can determine by how much each variable increases the total amount of information we need to describe the objects. Hence the variables can be ranked in terms of their ability to differentiate the population. The fission strategy in a divisive monothetic classification is to dichotomize the population at each stage by a particular variable. This assumes that we can order the variables in some way according to their significance. As just indicated, we can do this by examining the way in which (2.1) and (2.2) are calculated.

Therefore, for a divisive monothetic classification we calculate the total amount of information needed to describe the unclassified objects and then rank the variables according to their ability to differentiate the population. Using the variable of highest rank we then dichotomize the population of objects into two subgroups of those possessing or lacking a score on the variables or, perhaps, those having a score greater than or less than some critical level on the variable. The groups formed by the first division are first order groups. These first order groups are then dichotomized by the variables of second highest rank, to give second order groups. The classification continues in this way until it reaches a stage where it is deemed to satisfactorily serve the purpose for which it is being constructed. The classification will be unique only so long as no two variables have the same differenti-

ating ability. This classification method is simple, involving a minimum amount of calculation. Hence it should be possible to handle large data sets in a short time. As yet, however, there have not been any applications of this method.

Divisive monothetic systems for binary data have been extant for a number of years and have reached a high level of efficiency (Lance and Williams 1965, 1968d). Given a collection of n individuals, each characterized by s binary attributes, such that n_i individuals possess the i^{th} attribute Lance and Williams define the information content of the collection as

$$I_s = s\, n \log n - \sum_{i=1}^{s} [\, n_i \log n_i + (n-n_i) \log (n-n_i) \,] \quad . \quad (2.3)$$

As observed above, a Brillouin based measure may be more appropriate:

$$I_B = \sum_{i=1}^{s} [\, \log(n!) - \log(n_i!) - \log (n-n_i)! \,] \quad (2.4)$$

Let I(n) denote the amount of information required to describe the n individuals, without regard to whether the measure is Shannon or Brillouin based. Suppose the collection is subdivided into two subgroups of sizes h and k such that h + k = n. Let the corresponding information measures be I(h) and I(k) respectively. The reduction in the information measure due to the division may be written as

$$\Delta I(n, h\ k) = I(n) - I(h) - I(k) \quad . \quad (2.5)$$

The fission strategy is to dichotomize the population on each attribute in turn and calculate the corresponding ΔI. The population will be subdivided into classes of order one on that attribute for which ΔI is maximum. The subgroups that give the maximum ΔI are most dissimilar. The process is repeated on each of the subgroups of order one, and continues until the required number of subgroups is achieved. It is not necessary that each of the classes of order one be subdivided into classes of order two on the same attribute. MacNaughton-Smith (1965) indicated that this process defines a near optimum split. We may remark here that this method, called divisive information analysis by Lance and Williams (1968d), will only yield a unique classification so long as no two attributes have the same ΔI at any level in the hierarchy.

Originally, information analysis, both divisive and agglomerative (as will be seen later) was defined for binary data. This type of classification was of use to ecologists who often are only interested in whether or not certain species of plants occur in given plots of land. There have been some attempts to make this classification method applicable to other types of data. It has been extended to handle-mixed data with promising results, Lance and Williams (1971). A number of attempts have also been made to make this algorithm applicable to purely quantitative data. The important problem, of course, with quantitative data is that of defining a suitable measure of information for a continuous distribution. Provided that the form of the distribution is known then such a measure can be defined, but generally the form of the distribution is unknown. There is only one alternative, to define a series of classes

covering the range of the quantitative variable and record only presence or absence in a class rather than an actual value. This is the approach adopted by Lance and Williams (1967c) when they propose that continuous variables be "chopped" into multistate ordinal attributes, by dividing the range of observed values into m equal sized sets. The choice of m, however, is arbitrary, and this clearly affects the final results. The computed information measures increase with m, but so also does the amount of computation time required. Lance and Williams suggested that m might be taken equal to eight.

The problem of measuring the information content of quantitative data has also received attention from Dale (1971) and Dale et al. (1971, 1972). The first proposal by Dale involves making a conceptual distinction between the units to be classified and the samples recorded. Suppose some basic spatial unit to be classified is composed of n subunits. For each of s independent variables the number of subunits characterized by the $i^{\underline{th}}$ variable is denoted by n_i as usual. The information content of the basic spatial unit is

$$I_S = s\, n \log n - \sum_{i=1}^{s} [(n_i \log n_i) + (n-n_i) \log (n-n_i)] \qquad (2.6)$$

or

$$I_B = \sum_{i=1}^{s} [\log n! - \log n_i! - \log (n-n_i)!\,] \qquad (2.7)$$

This approach insures that each basic spatial unit to be classified has an information content greater than zero at the commencement of the classification. The classification then proceeds in the usual fashion.

The second proposal by Dale was to measure the information content of the collection of objects to be classified in a way which was similiar in spirit to the measures given above by (2.1) and (2.2). Williams (1972) has developed a model which can be used for either qualitative or quantitative data and which partitions the information into its qualitative and quantitative components. This problem was examined earlier by Orloci (1968), though not in a rigorous conceptual manner.

In general to form a classification of n objects, each characterized by s independent multi-state variables we may proceed as follows. For each variable j let there by r_j possible states that it can assume, j=1, 2, ..., s. Let n_{ij} be the number of times the $i^{\underline{th}}$ state occurs for the $j^{\underline{th}}$ variable such that

$$\sum_{i=1}^{r_j} n_{ij} = n_j \qquad \text{and} \qquad \sum_i \sum_j n_{ij} = n \ .$$

Then
$$I_s = s\, n_j \log n_j - \sum_i \sum_j (n_{ij} \log n_{ij}) \tag{2.8}$$

If $n_j = n$, all j, then

$$I_s = s\, n \log n - \sum_i \sum_j (n_{ij} \log n_{ij})$$

The Brillouin based measure is

$$I_B = \sum_{j=1}^{s} \left[\log (n_j!) - \sum_{i=1}^{r_j} \log (n_{ij})! \right] \tag{2.9}$$

Classification proceeds in a way that is similar to William's method for binary data. The population of objects is divided into classes on each variable in turn and an information loss measure, ΔI, is calculated for each division. The population is divided into classes of order one on that variable for which ΔI is maximal. The process is repeated on each of the resulting subgroups. The amount of computation needed for these classifications is not excessively large. Furthermore, these methods have some advantages over standard methods as will be seen below. However, as yet there have not been any applications of this type of divisive classification in the literature.

Divisive Polythetic Classification

Attempts to develop information based algorithms for divisive polythetic classification have been made by Rescigno and Maccacaro (1960) and Wallace and Boulton (1968). Both of these algorithms are mathematically more complicated than the algorithms for divisive monothetic classification. In the following a brief outline is given of the strategy involved in each algorithm. First we deal with the work of Rescigno and Maccacaro.

Assume that one has a set X of objects to be classified. Let m different variables $Y_i (i=1, 2, \ldots, m)$ be defined so that each Y_i is able to divide X into $P_i (P_i \geq 1)$ subsets. To the set X let some variable $Y_a (1 \leq a \leq m)$ be applied so that X is partitioned into subsets of order one: $X_1^{(a)}, X_2^{(a)}, \ldots\ldots, X_{P_a}^{(a)}$. Let n_i be the number of elements in each subset $X_i^{(a)}$. Then the mean information content of each element in X is $H(Y_a) = - \sum_i n_i/N \log n_i/N$, where N is the total

number of elements in X. The authors, in an attempt to establish a new terminology, call $H(Y_a)$ the repartment effected by Y_a. The mean repartment which can be effected by some variable Y_{b_1} in the subsets defined by Y_a is taken to be

$$S(Y_a ; Y_{b_1}) = H(Y_a, Y_{b_1}) - H(Y_a)$$

where $H(Y_a, Y_{b_1})$ is the repartment effected by the successive application of the variables Y_a and Y_{b_1}.

$S(Y_a; Y_{b_1})$ is interpreted as being a measure of the disorder existing inside the subsets defined by Y_a in respect of the variable Y_{b_1}. The authors then show that the disorder existing inside the subsets defined by Y_a in respect of all the other variables taken one by one, two by two and so on up to i by i can be measured by

$$\sum_{b_1} S(Y_a ; Y_{b_1}) = \sum_{b_1} H(Y_a, Y_{b_1}) - (m-1)\, H(Y_a) \quad .$$

$$\sum_{b_1 b_2} S(Y_a ; Y_{b_1}, Y_{b_2}) = \sum_{b_1 b_2} H(Y_a, Y_{b_1}, Y_{b_2}) - \frac{m-2}{2} \sum_{b_1} H(Y_a, Y_{b_1})$$

$$\vdots$$

$$\sum_{b_1 b_2 \ldots b_i} S(Y_a ; Y_{b_1}, Y_{b_2}, \ldots, Y_{b_i}) = \sum_{b_1 b_2 \ldots b_i} H(Y_a, Y_{b_1}, Y_{b_i}) - \frac{m-i}{i} \sum_{b_1 b_2 \ldots b_{(i-1)}} H(Y_a, Y_{b_1}, \ldots, Y_{b_{(i-1)}}) \quad (2.10)$$

The corresponding measures computed on the whole set, instead of on the subsets defined by Y_a are

$$\sum_{b_1} S^* (Y_a; Y_{b_1}) = \sum_{b_1} H(Y_{b_1}) .$$

$$\sum_{b_1 b_2} S^* (Y_a; Y_{b_1}, Y_{b_2}) = \sum_{b_1 b_2} H(Y_{b_1}, Y_{b_2}) - \frac{m-2}{2} \sum_{b_1} H(Y_{b_1}) .$$

.
.
.

$$\sum_{b_1 b_2 \ldots b_i} S^* (Y_a; Y_{b_1}, Y_{b_2}, \ldots, Y_{b_i}) = \sum_{b_1, b_2, \ldots, b_i} H(Y_{b_1}, Y_{b_2}, \ldots, Y_{b_i})$$

$$- \frac{m-i}{i} \sum_{b_1 b_2 \ldots b_{i-1}} H(Y_{b_1}, Y_{b_2}, \ldots Y_{b_{(i-1)}}) \quad (2.11$$

To measure the relative disorder existing inside the subsets defined by Y_a in respect of all the other variables taken one by one, two by two and so on up to i by i they use the following ratios,

$$R_1(Y_a) = \sum_{b_1} S(Y_a; Y_{b_1}) / \sum_{b_1} S^*(Y_a; Y_{b_1})$$

$$R_2(Y_a) = \sum_{b_1 b_2} S(Y_a; Y_{b_1}, Y_{b_2}) / \sum_{b_1 b_2} S^*(Y_a; Y_{b_1}, Y_{b_2})$$

.
.
.

$$R_i(Y_a) = \sum_{b_1 b_2 \ldots b_i} S(Y_a; Y_{b_1}, Y_{b_2}, \ldots, Y_{b_i} / \sum_{b_1 b_2, \ldots b_i}$$

$$S^*(Y_a; Y_{b_1}, Y_{b_2}, \ldots . Y_{b_i}) \quad (2.12)$$

$$\text{Let} \qquad R(Y_a) = \sum_{i=1}^{m-1} R_i(Y_a) \tag{2.13}$$

The variable Y_a for which $R(Y_a)$ is minimal is used to divide the set X into subsets of order one, which may be called classes of order one. Then by applying the same algebraic process to each of the classes of order one a collection of classes of order two can be obtained for each of them. In like manner, for each class of order two a collection of classes of order three can be obtained, and so on to classes of order m. This progressive division of the set X through classes of various order produces the classification.

The final classification arrived at by this procedure is unique only if no pair of variables have the same value of minimum relative disorder at any hierarchical level. There do not appear to have been any applications of this procedure reported in the literature of biology, ecology or geography. The reason for this is probably the excessive amount of computation time that would likely be required for any data set of reasonable size. As indicated in the introduction to this chapter this is a characteristic of most divisive polythetic classification procedures.

A second information based divisive polythetic classification algorithm has been developed by Wallace and Boulton (1968a,b). This is probably the most rigorous probabilistic procedure yet proposed. The authors treat the classification problem as an exercise in optimal coding: the best classification is considered to be that which results in the briefest recording of all the attribute information. The

measurements that are recorded on the objects to be classified are considered as messages about these objects. The information measure used in this algorithm is much more detailed than the one used by Lance and Williams in their information analysis. In this algorithm the information measure takes into account the number of the class to which any individual x belongs, the type of class it is - this can be any one of a predefined dictionary of classes (e.g. normal, uniform), the parameters of the class and the position of the individual x with respect to the class (e.g. the distance of x from the mean of the class values). Shannon and Brillouin based formulae for the information measure are given in the (1968a) and (1969) papers by Wallace and Boulton. This algorithm can be applied to quantitative as well as qualitative data. However, while a program has been developed for the algorithm (Wallace and Boulton 1968b), as yet there do not appear to have been any applications of it in the literature.

Agglomerative Monothetic Classification

As indicated in the introduction to this chapter classifications of this type can only exist in a trivial sense. Suppose we are given a collection of objects, characterized by values on one variable (attribute). The problem is to agglomerate these objects into classes in a way that will optimize some objective function. The usual objective function tries to maximize the between class variance and minimize the within class variability. In information theory terminology this amounts to maximizing the amount of information we need to select the class a particular object belongs to, while

minimizing the information needed to select the score of a particular object on the variable. Various clustering schemes which incorporate this notion have been used in geographical studies (Spence and Taylor 1969). An attractive decomposition procedure has been proposed by Theil (1967, 1972), based on Shannon's third axiom, (Shannon and Weaver 1949). The procedure has been used extensively by Semple et al. (1971, 1972, 1973) to measure inequality in spatial distributions.

Suppose some variable phenomenon X is distributed over m regions and let each region j contain r_j subregions (j = 1, 2, 3, ..., m). Let P_j be the probability of the variable occurring in the $j^{\underline{th}}$ region and let P_{ij} be the probability of the variable being represented in subregion i of region j. Then a measure of the between regional and within regional information content of the distribution is

$$I_s = - \sum_{j=1}^{m} P_j \log P_j - \sum_{j=1}^{m} P_j \left(\sum_{i=1}^{r_j} \frac{P_{ij}}{P_j} \log \frac{P_{ij}}{P_j} \right) \qquad (2.14)$$

where $\sum_{i=1}^{r_j} P_{ij} = P_j$ and $\sum_{j=1}^{m} P_j = 1$.

The corresponding Brillouin based measure can be derived in the following way. Suppose n units of some variable X are distributed over m regions in the proportions $n_1, n_2, \ldots, n_m$ such that $\sum_{j=1}^{m} n_j = n$. For example n could be the total income arising in a country, or the total population of a country. Then n_j would represent the total income, measured in dollars, arising in region j or the population of region j. A measure of the between region variability in the distribution is

$$I_B' = \frac{1}{n} [\log n! - \sum_{j=1}^{m} \log n_j!]$$

Assume each region is composed of r_j subregions and that the variable X is distributed over these subregions in the proportions $n_{1j}, n_{2j}, \ldots, n_{r_j j}$, where $\sum_{i=1}^{r_j} n_{ij} = n_j$.

Then a measure of the information content of the distribution within region j is

$$\frac{1}{n_j} [\log n_j! - \sum_{i=1}^{r_j} \log (n_{ij})!]$$

and the mean within-region measure is

$$I_B'' = \sum_{j=1}^{m} n_j/n \{1/n_j [\log n_j! - \sum_{i=1}^{r_j} \log (n_{ij})!] \} .$$

Combining I_B' and I_B'', the total between-region and within-region information content of the distribution is

$$I_B = \frac{1}{n} [\log n! - \sum_{j=1}^{m} \log n_j!] + \sum_{j=1}^{m} \frac{n_j}{n} \left(\frac{1}{n_j} [\log n_j! - \sum_{i=1}^{r_j} \log (n_{ij})!] \right) \qquad (2.15)$$

Writing $P_j = n_j/n$, $P_{ij} = n_{ij}/n_j$ and applying Stirling's approximation, if n is large, it is seen that (2.15) corresponds to (2.14).

The first term on the right hand side of (2.14) and (2.15) is a between-region information measure, whereas the second term is the within region measure. Nutenko (1970) has shown that on aggregating the subregions

the between region measure decreases monotonically as the size of the region increases, and that the within region measure is monotonically increasing. In classification the aim is to maximize the between region measure and minimize the within region measure. Batty (1972) used the Shannon based measure, (2.14), to aggregate zones in Reading, following the method of aggregation devised by Ward (1963). At each stage in the hierarchy the between region measure is maximised by computing the measure for every possible aggregation of single spatial units to their spatially adjacent regions. A spatial form of (2.14) was presented by Batty (1974). He argued that this procedure was likely to produce a suboptimal solution at any level in the hierarchy, although what the author considers to be an optimal solution is not defined. Therefore, he proposed an alternative algorithm which starts with a basic feasible solution and then proceeds in a trial and error fashion to derive a new solution which minimizes the within region information measure.

Agglomerative Polythetic Classification

In the light of the foregoing discussion the most obvious way to construct an agglomerative polythetic classification is to generalise the measures given above, by (2.14) and (2.15), and to apply the same aggregation strategy. Assume that instead of one there are s independent variables distributed over m regions. The information measures corresponding to (2.14) and (2.15) then are

$$I_s = - \sum_{k=1}^{s} \{ \sum_{j=1}^{m} P_{jk} \log P_{jk} - \sum_{j=1}^{m} P_{jk} (\sum_{i=1}^{r_j} \frac{P_{ijk}}{P_{jk}} \log \frac{P_{ijk}}{P_{jk}}) \} \qquad (2.16)$$

$$\text{and} \quad I_B = \sum_{k=1}^{s} \{ \frac{1}{n_k} [\log n_k! - \sum_{j=1}^{m} \log n_{jk}!] + \sum_{j=1}^{m} \frac{n_{jk}}{n_k} \left(\frac{1}{n_{jk}} [\log n_{jk}! - \sum_{i=1}^{r_j} \log n_{ijk}!] \right) \} \qquad (2.17)$$

where P_{jk} is the probability of variable k being represented in region j.

P_{ijk} is the probability of variable k being represented in subregion i of region j.

n_k is the total number of units of variable k occurring in region j.

n_{ijk} is the number of units of variable k occurring in subregion i of region j.

As before the aggregation method will follow Ward's (1963) procedure.

The method of information analysis in its original formulation by MacNaughton-Smith (1965) and Williams (1966) was an agglomerative polythetic procedure for binary data. The information measures used are the same as those used for divisive classification. The information content of the objects or groups of objects to be fused is measured in the usual manner. If the s variables that characterize a group of objects are not mutually independent then instead of (2.8) and (2.9) the following measures should be used,

$$I_s = n \log n - \sum_{k_1 k_2 \ldots k_s}^{r_1 r_2 \ldots r_s} \left(\bigcap_{j=1}^{s} n_{jk_j} \right) \log \left(\bigcap_{j=1}^{s} n_{jk_j} \right) \qquad (2.18)$$

$$\text{or} \quad I_B = \log n! - \sum_{k_1 k_2 \ldots k_s}^{r_1 r_2 \ldots r_s} \log \left(\bigcap_{j=1}^{s} n_{jk_j} \right)! \qquad (2.19)$$

where n_{jk_j} is the number of times the state k_j occurs for variable j, cf. equations (1.23) and (1.27) above, and Appendix B. When two objects or two subgroups h and k are fused the information measure I(n) for the composite group is greater than the sum of the information measures I(h) and I(k), for h and k respectively. The more dissimilar are h and k the greater is the increase in the information measure due to the fusion. Thus at any stage in the hierarchy those two objects are to be fused so that their union produces the smallest increase, ΔI, in the information measure, where

$$\Delta I = I(n) - [\, I(h) + I(k) \,] \quad .$$

This method of classification has been extensively used by Williams and his associates (see list of references), with considerable success. As far as this author is aware there have been only two applications of this algorithm in the geographical literature. Alexander (1972) used this method to classify districts within central Perth according to their land-use patterns and Bryant (1974) used this technique to classify areal units on the basis of the types of agricultural and urban changes that were taking place in the Paris region. These authors used information analysis in its simplest form, i.e. when the recorded data are treated as binary.

A further approach to agglomerative polythetic classification is through the use of contingency tables. The data for an individual or group of n individuals characterized by s attributes (variables) may be arranged in a variable/size class matrix, where the observed states on each variable may be partitioned into certain classes. Let r denote the number of size classes for each variable. Assume the variables are the

rows and the size classes are the columns of the matrix. Let n_{ij} be the number of individuals having values within the j^{th} size class of variable i. Let the row sums be $r_i (= \sum_j n_{ij})$, and let the grand sum be $N = \sum_i \sum_j n_{ij}$. For example the individual to be classified might be a subregion where s represents the number of different landuse types that occur and r represents the number of different intensity levels possible for each landuse. Then n_{ij} would be the number of landunits devoted to landuse i at the j^{th} level of intensity in the subregion.

The analysis of contingency tables by information statistics has been extensively studied by Kullback (1959), Kuppermann (1959), Kullback et.al. (1962), Good (1963, 1965) and Tribus (1969). The total information content of the matrix described above may be written as

$$N \log N - \sum_{i=1}^{s} \sum_{j=1}^{r} (n_{ij} \log n_{ij})$$

The information content of the variables ignoring the size classes is

$$N \log N - \sum_{i=1}^{s} (r_i \log r_i)$$

and for each variable the information measure associated with the division into size classes is

$$r_i \log r_i - \sum_{j=1}^{r} (n_{ij} \log n_{ij}) \quad .$$

The last measure can be summed over all the variables and then the total information content of the matrix may be partitioned as follows

$$N \log N - \sum_i \sum_j (n_{ij} \log n_{ij}) = \{ N \log N - \sum_i (r_i \log r_i) \}$$

$$+ \{ \sum_i [r_i \log r_i - \sum_j n_{ij} \log n_{ij}] \}$$

$$\text{Total} = \left(\begin{array}{c} \text{variables ignoring} \\ \text{size classes} \end{array} \right) + \left(\begin{array}{c} \text{size classes} \\ \text{within variables} \end{array} \right).$$

Alternatively, the partition may be given as

$$N \log N - \sum_i \sum_j (n_{ij} \log n_{ij}) = \{ N \log N - \sum_j (c_j \log c_j) \}$$

$$+ \{ \sum_j [c_j \log c_j - \sum_i (n_{ij} \log n_{ij})] \}$$

$$\text{Total} = \left(\begin{array}{c} \text{sizes ignoring} \\ \text{variables} \end{array} \right) + \left(\begin{array}{c} \text{variables within} \\ \text{sizes} \end{array} \right)$$

Then the row-column interaction may be measured by

$$[N \log N - \sum_i (r_i \log r_i)] - [\sum_j (c_j \log c_j) - \sum_i \sum_j (n_{ij} \log n_{ij})] \qquad (2.20)$$

The corresponding Brillouin based measure would be

$$[\log N! - \sum_{i=1}^{s} \log r_i!] - \sum_j [\log c_j! - \sum_i \log (n_{ij})!)] \qquad (2.21)$$

This measure is known as the mutual or transinformation (Kotz 1966). If the variable classification of the data is denoted by X and the size class classification by Y then (2.20) and (2.21) may be denoted by I(X; Y) where

$$I(X;Y) = I(X) - I(X|Y) \tag{2.22}$$

and $I(X|Y)$ is the information received from the X classification, given the Y classification. In terms of the example given earlier the Y classification relates to how intensively the different landunits in the subregion are utilized and the X classification relates to the actual landuses occurring in the subregion. $I(X;Y)$ provides a measure of how these classifications are interrelated.

There are a number of measures available for assessing the relatedness of two classifications, X and Y. One such measure is Rajski's (1961b) coherence coefficient which is given by

$$R(X,Y) = \sqrt{[1 - d^2(X,Y)]} \tag{2.23}$$

where

$$d(X,Y) = 1 - \frac{I(X;Y)}{I(X,Y)}$$

$$= \frac{I(X|Y) + I(Y|X)}{I(X,Y)} \quad , \text{ by (2.22)} \tag{2.24}$$

$$\text{and } I(X,Y) = N \log N - \sum_i \sum_j (n_{ij} \log n_{ij}) \quad .$$

The coefficient $d(X,Y)$ has been shown to be a metric measure in the set of all discrete probability distributions, (Rajski 1961a). The value of $R(X,Y)$ varies between zero and unity. Like the correlation coefficient when $R(X,Y) = 0$, X and Y are independent and $R(X,Y) = 1$ implies X and Y are functions of each other.

An alternative measure of the relatedness of the X and Y classifications, for large samples, is given by χ^2 with (s-1) (r-1) degrees of freedom because of the asymptotic equality of 2I(X;Y) and χ^2, (Wilks 1935, Kullback 1959). Rapid calculation of 2I(X;Y) can be performed from the table of 2n log n by Woolf (1957) or the table of n log n provided by Kullback (1959). Some other information based measures of the degree of correlation between variables were given by Linfoot (1957).

In the classification of subregions whose mutual information content has been measured by I(X;Y) the fusion strategy is the same as usual. For any pair of subregions h and k a coefficient

$$\Delta I_{hk} = I(X;Y)_{h+k} - I(X;Y)_h - I(X;Y)_k$$

is calculated. That pair of subregions for which ΔI_{hk} is minimal are fused first. The whole procedure is reiterated until all the subregions have been classified.

Information measures based on contingency tables have been used by Orloci (1968, 1969, 1970a, b, 1971) in his cluster analyses where the fusion strategy is the same as used here. The equivocation information measure, which corresponds to the total information content of a matrix minus the information content of the row-column interaction (the mutual information), was used in a classification by Williams et.al. (1973). However, the measure failed to give satisfactory results.

Some Properties of Information Analysis

The technique of information analysis developed by MacNaughton-Smith and Williams is the one that has been used most extensively by

ecologists and biologists. There have been a number of studies where the results provided by this technique have been compared with results obtained from applying other algorithms to the same data sets, e.g. Williams et.al. (1966), Lambert and Williams (1966) and Lance and Williams (1966). The algorithms against which information analysis has been compared are generally those using a centroid sorting fusion strategy and similarity coefficients such as the correlation coefficient or a variant of the Euclidean distance measure. Comprehensive reviews of the large number of similarity measures available have been given by Sokal and Sneath (1963) and Williams and Dale (1965).

In comparison with other algorithms information analysis appears to have many advantages.

(i) The similarity measure, ΔI, in information analysis increases monotonically with successive fusions. Therefore, there is no ambiguity when determining the stage at which different fusions occur in the classification. Euclidean distance measures and the correlation coefficient, however, are liable to occasional failure of monotonicity, (Lance and Williams 1966).

(ii) Information analysis is the only algorithm for which there is a known theoretical reason why the population should be fused into clearly separated groups and why the "chaining" effect should rarely occur. By the "chaining" effect is meant the tendency for a given group to grow in size by the addition of single individuals or groups much smaller than itself rather than by fusion with other groups of comparable size. Since the information content of a group increases with group size the algorithm will tend to delay the fusion of large groups, or the addition of outlying

individuals to existing groups until relatively late in the analysis. The effect will be demonstrated more formally below.

(iii) The information statistic provides a measure of the heterogeneity or diversity of a class at each stage in the hierarchy. If two regions h and k are combined to form a new region n then the most commonly used algorithms, using correlation coefficients or variants of Euclidean distance as similarity measure, provide a measure of how similar the regions h and k are, but they do not provide a measure of the heterogeneity of n. In this sense, it is reasonable to assume that the more dissimilar h and k are, the more heterogeneous will be n, so far as a single fusion or fission is concerned. If these measures could be accumulated over the hierarchy then a genuine measure of the heterogeneity of n could be obtained. However, neither the correlation coefficient nor the variants of Euclidean distance are additive in this sense. The convention in the past has been to take the (h, k) coefficient, technically a measure of a single fusion, as the best available measure of the heterogeneity of n, taken over the whole hierarchy up to the point at which n occurs.

However, the situation is changed completely when the information statistic is used, for this may be considered as an (n, h k) coefficient, defining the difference between n on the one hand and h and k jointly on the other hand. The information measure is completely additive for if we let I(n) be the information content of the composite group n then by definition

$$I(n) = I(h) + I(k) + \Delta I(n, h\ k) \quad .$$

Hence the information measure can be accumulated successively and a measure of the diversity or heterogeneity of a group at each level in the hierarchy can be determined.

(iv) A further advantage of information analysis is that it is completely insensitive to skewed distributions. In Euclidean models such distributions are usually standardized by variance. Then the rare events assume disproportionate importance and, therefore, distort the final results.

(v) The treatment of ordinal data by information statistics is completely rigorous. Hence their potential should receive more attention from those especially concerned with psychological data, where it is safest not to assume a metric structure in the data one is dealing with.

(vi) In the literature attention has been given to the fact that the information statistic has properties that may be used to construct a stopping rule on the hierarchy. Kullback (1959) showed that for large samples $2\Delta I$ is approximately distributed as χ^2 with as many degrees of freedom as there are attributes. Lambert and Williams (1966) have advocated using $2\Delta I$ with its χ^2 approximation to test for homogeneity in the two final groups. The initial null hypothesis must be that the two final groups are samples from a single population. If the null hypothesis is accepted then no further tests will be required. If, however, the null hypothesis is rejected then this implies the acceptance of the hypothesis of two or more underlying populations. In order to continue testing, a new null hypothesis must be postulated - the only logical one is that the two final groups are each samples from two distinct populations. Then the subgroups within each of these populations are compared, by the χ^2 test. The procedure continues until the null hypothesis is accepted. This procedure,

however, has been objected to by Bottomley (1971), who has indicated that it is incorrect to consider the subgroups as random samples from the population under consideration, since the individuals in the subgroups have been fused in a specific way.

Field (1969) suggested starting at the base of the dendrogram and testing each fusion in turn. Again he is testing whether two subgroups may be regarded as coming from one initial population. This procedure is also open to Bottomley's objection. Furthermore, a serious disadvantage in testing a dendrogram from the base is that the initial fusions necessarily involve very small sample sizes. In this type of situation the χ^2 approximation is far too crude to give useful results.

Information analysis, however, is not without some disadvantages. The technique is subject to what is, perhaps, a unique form of misclassification. If the population of individuals to be classified contains several substantially homogeneous subgroups with each of which is associated one or more outlying members, information analysis is liable to classify all such outlying members into a single "nonconformist" group, irrespective of their individual affinities. For example, in a study by Edye, Williams and Pritchard (1970) where there were only fifty one individuals in the population information analysis produced two nonconformist groups.

The nonconformist groups arise because of the sensitivity of the information statistic to group size. This effect can be demonstrated in the following manner.

Let ΔI denote the information measure between two individuals.

ΔI_i denote the information measure between an individual and a group.

ΔI_g denote the information measure between groups of equal size.

$\Delta I'$ denote the information measure between two entities (i.e. either or both may be individuals or groups).

Consider two groups, one of m and the other of n individuals such that the members of each group are identical; every member of one group possesses and every member of the other group lacks a single binary attribute. Then, by (2.6)

$$\Delta I' = (m+n) \log (m+n) - m \log m - n \log n \ .$$

Assume $m \geq n$, then $m = a\,n$, $a \geq 1$ and

$$\Delta I' = n \{ (a+1) \log (a+1) + (a+1) \log n - a \log a - a \log n - \log n \}$$

$$= n \{ (a+1) \log (a+1) - a \log a \} \ .$$

To obtain ΔI let $n = a = 1$.

Then $\Delta I = 2 \log 2$.

For ΔI_g let $a = 1$, $n \neq 1$, so that $\Delta I_g = 2\,n \log 2 = n\,\Delta I$.

Therefore, in comparisons between groups of equal size the information measure increases directly with group size. For ΔI_i, the information measure between an individual and a group, let $a \neq 1$, $n = 1$.

Then $\Delta I_i = (a+1) \log (a+1) - a \log a$

$$= \log a + (a+1) \log (1 + 1/a)$$

$$= (\log a) + 1 + 1/2a - 1/6a^2 + \ldots\ldots$$

$$\simeq (\log a) + 1$$ (Williams, Clifford, Lance 1971).

Therefore, for an individual waiting to be assigned to a group the likelihood of fusion diminishes as the group grows, according to the logarithm of the current group size and this effect continues indefinitely. It

follows that an outlying member of a large and otherwise fairly homogeneous group may have its fusion delayed so long that the information gain, ΔI, for fusion with another substantially unrelated single element may be less than that for fusion with the group whose members it most closely resembles. It is this situation that gives rise to the nonconformist groups. This is also the reason why the "chaining" effect does not occur.

The seriousness of the tendency to produce nonconformist groups depends ultimately on the purpose for which the classification is being constructed. In taxonomy where every individual must be accounted for in the best possible way the tendency for nonconformist groups to occur cannot be tolerated. In truly taxonomic situations Lance and Willians (1966) suggest that the information analysis results should be checked against the results of another strategy - they suggested centroid sorting with the nonmetric similarity coefficient used by ecologists because of the relatively good grouping and monotonicity properties of this strategy. If one is interested in identifying relatively homogeneous groups and keeping the anomalous individuals separate, as ecologists appear to be (Lance and Williams 1967b), then information analysis gives satisfactory results. It is easy to identify the nonconformist group(s) by calculating the diversity of each group. The nonconformist group(s) have a disproportionately high diversity measure in comparison with the other groups.

Information analysis as it was initially formulated can only be applied to binary data. While there have been some attempts to extend the procedure to quantitative data these have not been very successful. For continous variables it has been suggested that the sum of squares

method of Burr (1968, 1970) is more appropriate since it is less likely to produce nonconformist groups. However, this method suffers from being sensitive to skewed distributions.

A further point of concern for ecologists about information analysis is the way in which it gives equal weight to the joint presence or joint absence of attributes, (Webb et. al. (1967), Field (1969), Austin (1972)). This property may affect the fusion sequence and result in non-homogeneous groups. For example, the fusion of a pair of individuals with a large number of attributes in common may be delayed until after the fusion of a pair possessing few or maybe no attributes in common. Field (1969) suggested that for heterogeneous systems it is more appropriate to use only the joint presence of attributes when calculating the information measures. The appropriate information measures then are

$$I_S = \left(\sum_j n_j\right) \log \left(\sum_j n_j\right) - \sum_j (n_j \log n_j) \tag{2.25}$$

and

$$I_B = \log \left(\sum_j n_j\right)! - \sum_j \log (n_j)! \tag{2.26}$$

where n_j is the number of individuals possessing the $j^{\underline{th}}$ attribute.

From the foregoing discussion it should be clear that information analysis has many advantages, and for this reason it should re receive more attention from geographers than has hitherto been the case.

Spatial-Temporal Classification

The problem of devising statistical techniques for spatial-temporal analysis has not yet received much attention in the literature. There are a host of techniques available for analyzing spatial and temporal patterns separately. However, the number of techniques avalable

for the simultaneous analysis of spatial and temporal patterns is very limited. The following is an outline of a simple method, largely modified from Dale et. al. (1970) and Williams et. al. (1969). If it assumed that a temporal sequence possesses the Markov property that the future probability behaviour of some system under study is uniquely determined once the state of the system at the present stage is known, then it is the transitions between the states, rather than the states themselves, that carry the important information. The first necessity then is to set up a matrix of the observed transitions. Here we may note that other probability processes (e.g. Bernoulli) are classifiable in this manner also.

Consider a region R composed of the subregions R_1, R_2, ..., R_n. Let the time interval T be subdivided into equal-length subintervals, T_1, T_2, ..., T_t , where $T_{i-1} < T_i$, $i = 1, 2, ..., t$. Assume the simplest case where the subregions are characterized by only one variable (attribute) which can assume any one of r_1 different states. For any one of the subregions, say R_k, consider the sequence of states that occur for the variable. This sequence may be represented by a vector

$$(a_{k0}, a_{k1}, ..., a_{kt})$$

where $a_{kt'}$ is the state assumed by subregion k on the variable at time t' , $1 \leq k \leq n$, $0 \leq t' \leq t$. There are n of these vectors, one for each subregion. From these vectors a matrix of transition frequencies may be constructed. This is an $r_1 \times r_1$ matrix where the (i, j)$^{\underline{th}}$ entry represents the frequency with which the $j^{\underline{th}}$ state succeeds the $i^{\underline{th}}$ state. Denote the frequencies by n_{ij}; $i,j = 1, 2, 3, ..., r_1$ such that

$$\sum_j n_{ij} = n_i \quad , \text{ the } i^{\underline{th}} \text{ row total.}$$

$$\sum_i n_{ij} = n_j \quad , \text{ the } j^{\underline{th}} \text{ column total.}$$

$$\sum_i \sum_j n_{ij} = N \quad , \text{ the grand total.}$$

Denote the transition matrix for subregion R_k by M_k. The maximum likelihood estimate of the probability that given the $i^{\underline{th}}$ prior state, the $j^{\underline{th}}$ state will follow is n_{ij}/n_i and the mean information content of each transition from the $i^{\underline{th}}$ prior state is

$$I_s^i = - \sum_{j=1}^{r_1} (n_{ij}/n_i \log n_{ij}/n_i) \quad .$$

The information content of all the transitions from the $i^{\underline{th}}$ prior state is $n_i(I_s^i)$. Then summing over all prior states the information content of the entire matrix M_k is

$$I_{k_S} = \sum_{i=1}^{r_1} [\, n_i \log n_i - \sum_{j=1}^{r_1} (n_{ij} \log n_{ij}) \,] \tag{2.27}$$

The corresponding Brillouin based measure is

$$I_{k_B} = \sum_{i=1}^{r_1} [\, \log (n_i)! - \sum_{j=1}^{r_1} \log (n_{ij})! \,] \tag{2.28}$$

These information measures can now be used as a basis for the usual information based classification procedures. Consider two subregions R_h and R_k with transition matrices M_h and M_k respectively, and denote the information contents of these matrices by I_h and I_k. Combine R_h and R_k to produce a larger region R_ℓ where the transition matrix of R_ℓ is M_ℓ and

M_ℓ is the element by element sum of M_h and M_k. Denote the information content of M_ℓ by I_ℓ . The information gain associated with the fusion of R_h and R_k is

$$\Delta I = I_\ell - I_h - I_k$$

As usual agglomerative classification proceeds by fusing at each stage that pair for which ΔI is minimum.

If the subregions being classified are characterized by a number of independent variables, say s, then the information measures are

$$I^1_{k_s} = \sum_{x=1}^{s} \{ \sum_{i=1}^{r_x} [(n_i^x \log n_i^x) - \sum_{j=1}^{r_x} (n_{ij}^x \log n_{ij}^x)] \} \tag{2.29}$$

or

$$I^1_{k_B} = \sum_{x=1}^{s} \{ \sum_{i=1}^{r_x} [\log n_i^x! - \sum_{j=1}^{r_x} \log (n_{ij}^x)!] \} \tag{2.30}$$

where n_{ij}^x = the number of times the $j^{\underline{th}}$ state succeeds the $i^{\underline{th}}$ state on variable x.

$n_i^x = \sum_{j=1}^{r_x} n_{ij}$, the $i^{\underline{th}}$ row total in the transition matrix of variable x.

r_x = the number of different states that variable x can assume.

The statistics presented above to measure the information content of the transition matrices have one undesirable property. For instance if some of the subregions do not pass through all the possible states for a variable then the rows in the transition matrix corresponding to these states will contain all zero elements. If there are more than two sub-

regions like this then there is a possibility that a pair of them will combine at a very early stage in the analysis even though they may have none, or very few states in common.

This difficulty can be overcome if in addition we include in the analysis a measure of the information content associated with the partition of the grand total into row totals. For subregion R_k characterized by a single variable this information measure may be written as I_k^* where

$$I_{k_S}^* = N \log N - \sum_{i=1}^{r_1} (n_i \log n_i) \ . \tag{2.31}$$

or $$I_{k_B}^* = \log N! - \sum_{i} \log (n_i)! \ . \tag{2.32}$$

Then an alternative measure of the information content of the transition matrix for R_k is

$$\tilde{I}_k = I_k + I_k^* \ .$$

Then combining equations (2.27) and (2.31),

$$\tilde{I}_{k_S} = N \log N - \sum_{i=1}^{r_1} \sum_{j=1}^{r_1} (n_{ij} \log n_{ij}) \ . \tag{2.33}$$

Similarly, combining (2.28) and (2.32),

$$\tilde{I}_{k_B} = \log N! - \sum_{i=1}^{r_1} \sum_{j=1}^{r_1} \log (n_{ij})! \tag{2.34}$$

If the subregions are characterized by more than one variable then it follows immediately that for s independent variables

$$\tilde{I}^{1}_{k_s} = \sum_{x=1}^{s} \left[N^x \log N^x - \sum_{i=1}^{r_x} \sum_{j=1}^{r_x} n^x_{ij} \log n^x_{ij} \right] \tag{2.35}$$

and

$$\tilde{I}^{1}_{k_B} = \sum_{x=1}^{s} \left[\log N^x! - \sum_{i=1}^{r_x} \sum_{j=1}^{r_x} \log (n^x_{ij}!) \right] \tag{2.36}$$

where $N^x = \sum_{i=1}^{r_x} \sum_{j=1}^{r_x} n^x_{ij}$, the grand total of the entries in the transition matrix of the states for variable x in R_k.

The row totals do not appear in (2.33) - (2.36). In fact the expressions (2.33) and (2.34) are those that would be obtained by regarding the transition matrix as a single nominal multistate attribute with r_1^2 states.

This type of classification should receive some consideration from agricultural geographers who have hitherto been content to construct static classifications of agricultural land use, while it is clear that agricultural landuse patterns vary from year to year. In general this classification method can be applied to any sequential data.

Summary

In this chapter the classification problem was presented and some of the most commonly used strategies to deal with it were outlined. The classification procedures based on information theory that have been developed by ecologists and biologists have been reviewed and it has been indicated how these procedures may be applied to spatial classification.

The method of information analysis developed by the William's school has been the one most extensively used. Some of the properties of this method have been outlined. Finally, a simple information based algorithm for spatial-temporal classification was discussed.

From the literature reviewed in this chapter it is clear that information theory provides us with a number of statistics that can be used to construct classification algorithms. Hitherto most of the applications of these statistics to classification methods have been by ecologists, though generally under the restrictive assumption of independent variables. Here that assumption has been relaxed in some cases. Furthermore, almost all the measures that have been given in the literature are Shannon based. This measure is appropriate in ecologic and biologic situations where one is usually constructing species classifications from sample data and these classifications are designed to have universal applicability. However, in geographic situations the collection of objects to be classified, at times forms a completely sampled population. Then, by the arguments in the first chapter the information measures should be based on Brillouin's. From the literature reviewed it is also clear that as yet geographers have failed to take advantage of these potentialities of information statistics for classification. In the next chapter an algorithm for classifying individuals, characterized by scores on ordinal multistate attributes, will be developed. It is felt that this algorithm should be of relevance to behavioral geographers, who often have to manipulate psychological data.

CHAPTER 3

A CLASSIFICATION ALGORITHM FOR ORDINAL DATA

Introduction

In recent years many disciplines, including geography (Johnston 1968, 1970), have displayed an increasing interest in taxonomic procedures. An extensive review of some of the recent literature was provided by Spence and Taylor (1970). An integral part of all taxonomic procedures is a similarity or proximity measure. In the natural sciences the most commonly used similarity indices are based on some Euclidean distance measure. These similarity measures are almost always assumed to be metric and based on interval or ratio scale data. This is not a serious assumption when the objects to be classified are inanimate. If, however, one wishes to classify a collection of individuals into groups on the basis of some psychological attributes, then the requirement that the similarity measures have the properties of a metric may be too stringent. If one assumes that an individual's psychological system can be represented as an abstract metric system then one must also assume that the psychological system has the properties of a metric space. These properties are

$$\begin{aligned}
&\text{(i)} && d(x,y) \geq 0 && \text{, Positivity.}\\
&\text{(ii)} && d(x,x) = 0 && \text{, Reflexivity.}\\
&\text{(iii)} && d(x,y) = d(y,x) && \text{, Symmetry}\\
&\text{(iv)} && d(x,y) \leq d(x,z) + d(z,y), && \text{, Triangle Inequality.}
\end{aligned}$$

where $d(x,y)$ is a distance measure between x and y, (Cullen, 1968).

Psychological data, however, are most reliable in an ordinal form. Psychological attributes have no known unique zero, precluding ratio data, and it is not often easy for respondents to supply even reliable interval data, (Simon 1957). Data in rank order or ordinal form such as preference data (Gould 1965, Gould and White 1974) may not satisfy the reflexivity or symmetry properties of a metric. Shepard (1964) has indicated that interstimulus similarities are likely to violate the triangle inequality. In choice theory the assumption corresponding to the triangle inequality is that of transitivity of preferences. In reality this assumption is again rather restrictive, (Simmons, 1974).

Nevertheless, in one form or another metric approaches initially pervaded almost the entire field of psychological scaling. Among the proponents of this approach were Stevens (1951) and Torgerson (1958). Nonmetric approaches were advanced in the early sixties by Shepard (1962) and later developed by successive workers, notably Kruskal (1964a, 1964b). Coombs (1964) quite correctly argued that simple qualitative judgements can usually be made, not only with greater ease and assurance, but also with greater reliability and validity than numerical or quantitative judgements.

In recent years psychologists have developed multidimensional scaling, MDS, procedures that take nonmetric data as imput and impose a metric structure on them. The potential of these techniques for research in behavioral geography has been indicated recently in monographs by Golledge and Rushton (1972) and Brummell and Harman (1974). Since MDS techniques are designed to locate n stimuli and/or individuals

as points in an r-dimensional space, they may be regarded as special types of classificatory procedures, because they impose a certain amount of order on the data. Also, since the output from MDS techniques has a metric structure, the commonly used clustering algorithms, such as those of Berry (1961), and Johnson (1967) may be applied.

In this chapter a classification algorithm which takes the raw data in their non-metric form as input is presented. The algorithm calculates an index of the diversity of the individuals to be classified. The most typical individual is identified and the first cluster is formed around him. All individuals for which their similarity with the most typical one is greater than a critical value, ALPHA, go into the first cluster. Then all of these individuals are removed from the data set and the same overall procedure is applied to the remaining set. The process continues until every individual is classified. A simple example to illustrate the steps involved is given and some possible extensions of the algorithm are outlined. Then the algorithm is applied to a sample of 198 individuals, each characterized by scores on twelve independent factors. Finally a wide range of situations where this algorithm could be applied are indicated.

The Algorithm

(1) *The Data;* Assume that one has a sample of M individuals and that for each individual one has N pieces of information on N attributes. The N pieces of information about the $i^{\underline{th}}$ individual may be arranged in a vector

$$\underset{\sim}{X}_i = (x_{i1}, x_{i2}, \ldots\ldots, x_{iN}) \quad , \quad i = 1, 2, 3, \ldots\ldots, M.$$

For each variable j let there be r_j different values that it can assume where

$$2 \leq r_j < \infty \quad , \quad j = 1, 2, \ldots, N.$$

For example, if the sixth variable measures an individuals educational attainments and five different levels are allowed, then $r_6 = 5$. The data for the M individuals are recorded in an M x N matrix, $X = \{x_{ij}\}$, where the x_{ij} entry represents the value (score) assumed by individual i on variable j , i = 1, 2, ... M, j = 1, 2, ...N.

It is assumed that the data are ordinal. All that is known about any two values concerns only which one is larger. Nothing is explicitly stated about how much larger. The actual numerical values have no quantitative meaning, but serve only as coding identifiers.

(2) *Information Content Of The Data;* Consider the following statement, "individual i assumes the value k on variable j". The information content of that statement is by definition,

$$\text{Information Content} = \log \left[\frac{\text{the receiver's probability that i assumes the value k on variable j after the statement.}}{\text{the receiver's probability that i assumes the value k on variable j before the statement.}} \right] \quad (3.1)$$

Clearly, in the absence of noise, the numerator in (3.1) is unity. Writing the denominator as $P_{ij}(k)$, then (3.1) may be rewritten as

$$\text{Information Content} = - \log P_{ij}(k) .$$

In general, the information content of any entry $x_{ij} = k$ in the data matrix X is

$$I(x_{ij} = k) = - \log P_{ij}(k) \quad , \tag{3.2}$$

where

$$P_{ij}(k) = \text{Prob } (x_{ij} = k) \quad .$$

Assume that all the information one has about the sample of individuals is contained in X. The probabilities, P_{ij}, are calculated from the observed frequencies in the matrix. $P_{ij}(k)$ is the receiver's prior probability that any individual i assumes the value k on variable j. If the sample of individuals being studied are representative of a larger group then $P_{ij}(k)$ is the receiver's prior probability that any individual i in the larger group assumes the value k on variable j. Then, by the arguments of chapter one, $P_{ij}(k)$ does not vary from one individual to the next. $P_{ij}(k)$ is then taken as

$$P_{ij}(k) = P_j(k) = n_{jk}/M \ ,$$

where n_{jk} is the number of times the event $x_{ij} = k$ $(i = 1, 2, \ldots\ldots, M)$ is observed. If, however, the collection of individuals being studied form a complete population then, again by the argument in chapter one, $P_{ij}(k)$ will vary from one individual to the next. Specifically, for the first individual $P_{1j}(k) = n_{jk}/M.$ and for the $L + 1^{\underline{st}}$, $(L < M)$, individual

$$P_{(L+1)j}(k) = \frac{n_{jk} - \ell_{jk}}{M-L}$$

where ℓ_{jk} is the number of times the event $x_{ij} = k$ occurs among the first L events observed, and $\sum_{k=1}^{r_j} \ell_{jk} = L$.

The total information content of X depends on whether the data are from a sample or a population. If the former is the case, then one may denote by n_{jk} ($k = 1, 2, \ldots\ldots, r_j$) the number of times the score k occurs as an entry among the M entries in column j, such that

$$\sum_{k=1}^{r_j} n_{jk} = M \qquad \text{and} \qquad n_{jk} = M\, P_j(k) .$$

The total information content of these n_{jk} entries is

$$n_{jk}\,(-\log P_j(k)) = -M\, P_j(k) \log P_j(k) .$$

Then the total information content of the M entries in column j of X is

$$-M \sum_{k=1}^{r_j} P_j(k) \log P_j(k)$$

$$= MI_1(j) \quad ,$$

where $I_1(j)$ is the mean information content of each entry in column j. If the variables are independent the column informations are additive and the total information content of X is

$$I_1(X) = M \sum_{j=1}^{N} I_1(j) . \tag{3.3}$$

It is known that for each j, $I_1(j)$ attains its maximum when each of the r_j different scores are equally likely, i.e., $P_j(k) = 1/r_j$,

$k = 1, 2, \ldots\ldots, r_j$. Therefore, the maximum information content of column j of X is

$$M\, I_{1\ Max}(j) = -M \sum_{k=1}^{r_j} (1/r_j) \log (1/r_j)$$

$$= M \log r_j \ .$$

Then the maximum information content of X is

$$I_{1\ Max}(X) = M \sum_{j=1}^{N} \log r_j \qquad (3.4)$$

If the sample of individuals being studied is a population then the information content of column j of X is calculated from Brillouin's formula as

$$M\, I_2(j) = [\ \log M! - \sum_{k=1}^{r_j} \log n_{jk}!\ ]$$

Assuming the variables are independent the total information content of X is

$$I_2(X) = M \sum_{j=1}^{N} I_2(j) = \sum_{j=1}^{N} [\ \log M! - \sum_{k=1}^{r_j} \log n_{jk}!\ ] \qquad (3.5)$$

The maximum information content of the matrix, X, by Brillouin's formula is

$$I_{2\ Max}\ (X) = \sum_{j=1}^{N} [\ \log M! - (r_j - r) \{\ \log [\ M/r_j\ ]\ !\ \}$$

$$- r \{\ \log (\ [\ M/r_j\ ] + 1\ )\ !\ \}\ ] \qquad (3.6)$$

where

$[\ M/r_j\ ]$ is the integer part of M/r_j

and

$$r = r_j\ (\ M/r_j - [\ M/r_j\ ]\ ).$$

One may observe here that (3.3) and (3.5) permit one to rank the variables according to the amount of information received from each. The variable with the largest M I(j) conveys more information about the sample than any other. Conversely, variables with relatively small I(j) values convey little information about the individuals, and, therefore, may be regarded as poor differentiating characteristics of the sample.

(3) *Diversity and Typicality Indices;* An index of the diversity, D, of the sample of individuals is provided by the ratio of the actual total information content of X over the theoretical maximum information content. When Shannon's formula is used to calculate the information content of X one may write the diversity as

$$D_1 = I_1(X)/I_{1\ Max}\ (X) \qquad (3.7)$$

$$= \sum_{j=1}^{N} I_1(j) / \sum_{j=1}^{N} \log r_j$$

Similarly, when the sample is a population

$$D_2 = I_2(X)/I_{2\ Max}\ (X) \qquad (3.8)$$

The diversity index varies between the limits

$$0 \leq D_k \leq 1 \quad , \quad k = 1, 2. \qquad (3.9)$$

$D_k = 0$ implies $\sum_{j=1}^{N} I_k(j) = 0$, and in fact $I_k(j) = 0$, $j=1, 2, \ldots N$,

since $I_k(j) \geq 0$, all j, $k = 1, 2$. Therefore $D_k = 0$ implies that all the individuals have the same values on all the variables, i.e., the individuals are identical and there is not any diversity. Conversely, when $D_k = 1$, $k = 1, 2$, any individual is equally likely to choose any one of the values that are possible for each variable.

Given the diversity measure for the sample it is necessary to first identify the most typical individual in the sample before cluster formation can be initiated. The most typical individual is the one that conveys more information about the sample than any other.

If the $i^{\underline{th}}$ individual is omitted from the sample of size M one may rewrite the total diversity of the remaining M-1 individuals as D_k^i. Each individual i, $i = 1, 2, \ldots, M$, is omitted in turn and the corresponding D_k^i are calculated. The typicality of individual i is defined to be

$$T_k^i = D_k^i - D_k \quad , k = 1, 2. \tag{3.10}$$

Then

$$T_1^i = \frac{\left[\sum_{j=1}^{N} \sum_{k=1}^{r_j} \tilde{P}_j(k) \log \tilde{P}_j(k) - \sum_{j=1}^{N} \sum_{k=1}^{r_j} P_j(k) \log P_j(k)\right]}{\sum_{j=1}^{N} \log r_j} \tag{3.11}$$

where $\tilde{P}_j(k)$ is the probability of any individual assuming the value k on variable j, when the probabilities are calculated from the observed frequencies over M-1 individuals.

Similarly,

$$T_2^i = \frac{I_2(X)}{I_{2\ Max}(X)} - \frac{\sum_{j=1}^{N} [\log\ (M-1)! - \sum_{k=1}^{r_j} \log\ (n_{jk}^*)!]}{\sum_{j=1}^{N} [\log\ (M-1)! - (r_j - r')\ \log([\frac{M-1}{r_j}]!) - r' \log\{([\frac{M-1}{r_j}]+1)!\}]} \qquad (3.12)$$

where n_{jk}^* is the number of times the score k is observed on variable j such that $\sum_k n_{jk}^* = M-1$.

If individual i has a common value with a large number of other individuals on a number of variables its removal will have an equalizing effect on the frequencies n_{jk}^* and on the $\tilde{P}_j$'s. In this case

$$D_k^i > D_k \quad \text{and} \quad T_k^i > 0 \ , \quad k = 1, 2.$$

Conversely, if individual i has values on a number of variables which are not shared by many others, then its removal results in a more peaked frequency distribution. Then

$$D_k^i < D_k \quad \text{and} \quad T_k^i < 0 \ , \quad k = 1, 2.$$

The last point is illustrated by the following simple example. The scores of ten individuals on a certain attribute are recorded below. Each individual can assume anyone of five possible scores.

INDIVIDUAL	1	2	3	4	5	6	7	8	9	10.
SCORE	4	3	4	5	2	4	1	3	4	2.

Let n_k denote the number of occurrences of score k, k = 1, 2, 3, 4, 5. Then $n_1 = 1$; $n_2 = 2$; $n_3 = 2$; $n_4 = 4$; $n_5 = 1$.

The mean information content per score is given by

$$I_1 = 0.6388 \qquad I_2 = 0.4577 .$$

Omit the sixth individual who has the same score as three others.

Then $$I_1 = 0.6592 \qquad I_2 = 0.4643 .$$

Omit the seventh individual who is unlike the remainder, then

$$I_1 = 0.5374 \qquad I_2 = 0.3976 .$$

The corresponding typicalities, by (3.11) and (3.12) are

$$T_1^6 = 0.0276 \qquad T_2^6 = 0.0439$$

$$T_1^7 = -0.1306 \qquad T_2^7 = -0.0841 .$$

The individual, t, with the highest typicality is regarded as the most typical one. In the example above the sixth individual is said to be more typical than the seventh. This individual by his removal creates the maximum increase in the diversity of the reduced (M-1) x N data matrix. This individual t is designated to be the nucleus of the first group of individuals.

(4) *Similarity Measure and Group Formation;* For group formation a similarity measure between individuals is needed. One cannot use an Euclidean distance measure because of the nonmetric nature of the data. Instead a similarity measure is defined in terms of the number of

identical scores by a pair. The similarity between two individuals, a and b, is defined as

$$S_{a,b} = \sum_{j=1}^{N} \begin{cases} r_j & \text{when } x_{aj} = x_{bj} \\ 0 & \text{when } (x_{aj}).(x_{bj}) = 0 \end{cases} \tag{3.13}$$

This similarity measure is modified from Hyvarinen (1970). The similarity, $S_{a,b}$, is obtained by summing up r_j for all j that have the same value in the two rows a and b for each column j. Unequal values do not contribute to the sum. Clearly, $0 \leq S_{a,b} \leq \sum_{j=1}^{N} r_j = R$ and $R \geq 2N$, since $r_j \geq 2$, all j.

The similarity measure defined in (3.13) is justified by the following probabilistic consideration. Suppose the r_j values for any variable j are randomly distributed then the conditional probability of individual b having the same value on variable j as individual a has, given that a's value is known, is $1/r_j$. The larger r_j is, the less likely it is that a coincidence will occur. Weighting each coincidence by the appropriate r_j gives an average of one per variable independent of the number of possible choices r_j. Thus by definition the variables contribute equally to the similarity.

If the variables are statistically independent then the expectation of the similarity of a and b, given the values that a assumes, is

$$E(S_{b|a}) = \sum_{j=1}^{N} r_j \, P(r_j) = N ,$$

where $P(r_j)$ is the conditional probability that a and b have the same

values on variable j, given that the values of a are known. The degree of similarity between a pair, a and b, can be partitioned as follows.

(i) $S_{a,b} = R$, a and b are identical.

(ii) $N < S_{a,b} < R$, there is varying similarity.

(iii) $N = S_{a,b}$, neutral point.

(iv) $0 < S_{a,b} < N$, there is varying dissimilarity.

(v) $S_{a,b} = 0$, a and b are completely dissimilar.

To form clusters first identify the most typical individual, t, in the sample. Then all other individuals i, (i=1, 2, ..., M, i≠t), satisfying the inequality

$$S_{t,i} \geq \alpha S_{t,t} = \alpha R = \text{ALPHA} \tag{3.14}$$

are allocated to the first cluster. The coefficeint α defines the class limiting factor ALPHA and lies between the limits $N/R \leq \alpha \leq 1$.

If a cluster is formed when $\alpha = 1$ then all the individuals in that cluster are identical to the most typical individual in the sample. As α approaches N/R the individuals in the cluster become more diverse. After the first cluster is formed those individuals are removed from the data set. Then the same procedure, from the beginning, is applied to the remaining set of individuals. That is, the most typical individual is identified in the remaining set and a new cluster is formed. The process continues until all the individuals have been assigned to a cluster.

When the range of values, r_j, is the same for each variable the cut off point or class limiting factor, ALPHA, can be selected to represent the number of attributes an individual has in common with the most typical individual. For example, suppose the individuals are measured on twelve attributes and that for each attribute there are six possible values. Then R = 72, and for $\alpha = 0.25$ it is required that $S_{t,i} \geq 18 = \text{ALPHA}$, i.e., any individual i must be similiar to the most typical one on at least three attributes. Hence in this case any cluster of individuals satisfying the inequality $S_{t,i} \geq 18$ are each similar to the nucleus of the cluster on at least three attributes.

A computer program, CLASSINF, has been written for the algorithm. It has been programmed for the CDC 6400 system at McMaster University. A listing of the program is given in Appendix C.

Example : The following simple example is given to illustrate the steps involved in this classification procedure. It is emphasized that this example is for purely illustrative purposes and no attempt is made to interpret the groups that result from the classification. Data are available on six attributes of each of ten individuals. For simplicity it has been assumed that there are only five possible values for each attribute.

The first four attributes relate to individual shopping behavior. The data are the coded responses of the individuals to the following statements,

(1) I enjoy buying expensive clothes.
(2) I regard shopping as a necessity rather than a pleasure.
(3) I buy at stores which undersell their competitors.
(4) I am willing to sacrifice quality for low prices.

The responses were coded as 1, 2, 3, 4, or 5 depending on whether an individual strongly disagreed, (1); disagreed, (2); was indifferent, (3); agreed, (4); or strongly agreed, (5), with the statements.

The fifth attribute measured each individuals level of educational attainment, ranging from (1) for elementary school level to (5) for University degree. The sixth attribute indicates the age group an individual belongs to, ranging from (1) for 0-15 years to (5) for over sixty years old. The scores recorded for the ten individuals on the six attributes are contained in Table 1.

		VARIABLE					
		1.	2.	3.	4.	5.	6.
I	1.	2	1	5	4	4	2
N	2.	2	1	5	2	2	4
D	3.	1	2	4	2	5	4
I	4.	2	2	5	2	5	3
V	5.	2	2	5	3	3	2
I	6.	2	1	3	3	3	3
D	7.	2	1	5	5	4	4
U	8.	2	2	4	1	4	5
A	9	2	2	5	2	5	3
L	10.	1	2	4	3	3	4

TABLE I .

From Table I a matrix of probabilities, $P_j(k)$, is calculated for Shannon's information measure and a matrix of frequencies is calculated for Brillouin's measure. These matrices are

VARIABLE	SCORE				
	0.2	0.8	0.0	0.0	0.0
	0.4	0.6	0.0	0.0	0.0
	0.0	0.0	0.1	0.3	0.6
	0.1	0.4	0.3	0.1	0.1
	0.0	0.1	0.3	0.3	0.3
	0.0	0.2	0.3	0.4	0.1

(a)

2	8	0	0	0
4	6	0	0	0
0	0	1	3	6
1	4	3	1	1
0	1	3	3	3
0	2	3	4	1

(b)

The entry (j,k) corresponds to $P_j(k)$ in (a) and to n_{jk} in (b), j=1, 2, ...,6, k = 1, 2, ..., 5.

The information content of column j of Table I is calculated from row j of the matrix (a) or (b) depending on which information measure is being used. Then by Shannon's formula the mean information content of each entry in column 1 of Table I is

$$I_1(1) = -[\,0.2 \log 0.2 + 0.8 \log 0.8 + 0 \log 0 + 0 \log + 0 \log 0]$$
$$= 0.2173 \text{ decits, where } 0 \log 0 \text{ is taken as zero.}$$

By Brillouin's formula the measure is

$$I_2(1) = \frac{1}{10}[\log 10! - (\log 2! + \log 8! + \log 0! + \log 0! + \log 0!)]$$
$$= 0.1653 \text{ decits.}$$

The total information content of any column j of Table I is $10[I_k(j)]$ k = 1, 2. The results are, in summary

j	$10[I_1(j)]$	j	$10[I_2(j)]$
1	2.173	1	1.653
2	2.922	2	2.322
3	3.899	3	2.924
4	6.160	4	4.401
5	5.706	5	4.225
6	5.558	6	4.100

By (3.3) and (3.5), if the variables are independent, the total information content of Table I is

$$I_1 = 26.420 \qquad I_2 = 19.627$$

The maximum information content of Table I by (3.4) is

$$I_{1\ \mathrm{Max}} = 60 \log 5 , \quad \text{since } r_j = 5 , \text{ all } j$$
$$= 41.938 .$$

Similarly, by (3.6), $I_{2\ \mathrm{Max}} = 30.327$.

Then the diversity of the ten individuals is, by (3.7) and (3.8),

$$D_1 = 0.6299 , \quad D_2 = 0.6472 .$$

Since $0 \leq D_1, D_2 \leq 1$ one may conclude that these ten individuals are reasonably diverse.

The next step is to calculate the typicalities, T_k^i , k = 1, 2, i = 1, 2, ..., 10.

If the first individual is omitted from the sample then the new probability and frequency matrices are

VARIABLE	SCORE									
	0.222	0.778	0.000	0.000	0.000	2	7	0	0	0
	0.333	0.667	0.000	0.000	0.000	3	6	0	0	0
	0.111	0.444	0.333	0.000	0.111	1	4	3	0	1
	0.000	0.000	0.111	0.333	0.556	0	1	1	3	5
	0.000	0.111	0.333	0.222	0.333	0	1	3	2	3
	0.000	0.111	0.333	0.444	0.111	0	1	3	4	1
	(a)					(b)				

The total information content , maximum information content and diversity measures of these matrices are

$$I_1^1 = 25.378 \qquad I_{1\ \mathrm{Max}}^1 = 41.938 \qquad D_1^1 = 0.6051$$

$$I_2^1 = 16.688 \qquad I_{2\ \mathrm{Max}}^1 = 26.134 \qquad D_2^1 = 0.6386$$

The superscript 1 is used to indicate that these measures are obtained when the first individual is omitted.

The typicality of the first individual is, by (3.10)

$$T_1^1 = D_1^1 - D_1 = -0.0248 \quad ; \quad T_2^1 = D_2^1 - D_2 = -0.0086 .$$

In a similar way the typicalities of the remaining nine individuals are calculated. The results are

j	T_1^j	j	T_2^j
1	-0.0248	1	-0.0086
2	-0.0110	2	0.0077
3	-0.0116	3	-0.0018
4	0.0122	4	0.0279
5	0.0024	5	0.0164
6	-0.0279	6	-0.0134
7	-0.0149	7	0.0029
8	-0.0447	8	-0.0249
9	0.0122	9	0.0279
10	-0.0156	10	-0.0066

The most typical individual is by definition the one with the highest typicality. Here the highest typicality is shared, in both cases, by the fourth and ninth individuals. Since both of these individuals are identical the first one, i.e. the fourth, is chosen to be the nucleus of the first cluster.

To determine which individuals go into the first cluster the similarity between each individual and the most typical one is calculated by (3.13). For example, the similarity between the first and fourth individuals is

$$S_{4,1} = 5 + 0 + 5 + 0 + 0 + 0 = 10 .$$

Similarly, $S_{4,2} = 15$; $S_{4,3} = 15$; $S_{4,4} = 30$; $S_{4,5} = 15$; $S_{4,6} = 10$;

$S_{4,7} = 10$; $S_{4,8} = 10$; $S_{4,9} = 30$; $S_{4,10} = 5$.

The size of the first cluster is determined by the magnitude of the cut-off point, ALPHA. For ALPHA = 10 all the individuals, except the tenth go into the first cluster. When ALPHA = 15 the first cluster contains the following individuals, 2, 3, 4, 5, 9. If ALPHA is increased to 20 there are only two members, 4 and 9, in the first cluster. After the first cluster is formed those individuals are removed from the sample. The algorithm is then reapplied to the remaining set of individuals, until everyone is classified. The following classifications have been constructed from different values of ALPHA.

	ALPHA = 10	ALPHA = 15	ALPHA = 20
GROUP		GROUP MEMBERS	
1	1,2,3,4,5,6,7,8,9	2,3,4,5,9	4,9
2	10	1,7	5
3	-	10	1,2,7
4	-	6	3,10
5	-	8	6
6	-	-	8

Clearly, the results of the first classification may be ignored, since it forces almost the whole sample into one group. The choice of which one of the remaining two to accept may depend on the purpose for which the classification is being constructed. If the second classification (ALPHA = 15) is accepted then the members of each

cluster are similar to the nucleus of the cluster on at least three attributes. In the second classification the tenth individual forms a group by himself. However, he is similar to the fifth individual on three attributes. The fifth individual is the second most typical one in the sample. Hence, perhaps, the tenth individual may be incorporated into the first group. The two remaining individuals, 6 and 8, are virtually unclassifiable.

Some Extensions Of The Algorithm

I Suppose one is unable to obtain data on some attributes for one or more of the individuals. This situation may arise if an interviewee is unco-operative and does not supply all the pieces of information that the interviewer requires. Suppose, in general, one does not have data for the $x_{\ell m}$ entry of the data matrix. Then the most unbiased value that one can assign to $x_{\ell m}$ is that value which minimizes the increase in the total information content of the matrix. Since the column informations are additive, minimising the total information content of the data matrix with respect to $x_{\ell m}$ is equivalent to minimizing the information content of column m with respect to $x_{\ell m}$.

Let the entries in column m be $x_{1m}, x_{2m}, \ldots, x_{(\ell-1)m}, x_{(\ell+1)m} \ldots$ x_{Mm} and let the number of possible values for each individual be r_m, $2 \leq r_m < \infty$. Let n_{mk} denote the number of times the value k occurs among the M-1 entries in column m. Before assigning a value to $x_{\ell m}$ the column information is

$$- (M-1) \sum_{k=1}^{r_m} \left(\frac{n_{mk}}{M-1} \right) \log \left(\frac{n_{mk}}{M-1} \right) \tag{3.15}$$

and after it is

$$- M \sum_{k=1}^{r_m} \left(\frac{\overset{*}{n}_{mk}}{M} \right) \log \left(\frac{\overset{*}{n}_{mk}}{M} \right) \tag{3.16}$$

where $\overset{*}{n}_{mk} = n_{mk} + Z_k$ (3.17)

and $Z_k = 0$ or 1, such that $\sum_{k=1}^{r_m} Z_k = 1$.

The value assigned to $x_{\ell m}$ is the one that minimizes (3.16). This value is obtained by setting $Z_k = 1$ for each $k = 1, 2, \ldots, r_m$, in turn, and then by comparing the results to obtain the minimum one.

If Brillouin's measure is being used then the information content of column m before a value is given to $x_{\ell m}$ is

$$\log (M-1)! - \sum_{k=1}^{r_m} \log (n_{mk})! \tag{3.18}$$

and after $x_{\ell m}$ has been assigned a value it is

$$\log M! - \sum_{k=1}^{r_m} \log (\overset{*}{n}_{mk})! \tag{3.19}$$

where $\overset{*}{n}_{mk}$ is again defined by (3.17).

As above, $x_{\ell m}$ is assigned that value which minimizes (3.19).

II. When the variables are statistically independent the total information content of the data matrix is measured by

$$I_k(X) = M \sum_{j=1}^{N} I_k(j) \quad , \quad k = 1, 2. \tag{3.20}$$

If however, one is classifying individuals on the basis of data obtained from interviews it is highly unlikely that the attributes or variables, on which scores have been recorded, are statistically independent. In interviews it is more often the case that a number of questions are asked to measure the same underlying dimension of variability. Denoting each of the variables by X_i, i=1, 2, 3, ...,N expression (3.20) may be rewritten as

$$I_k(X_1,X_2,\ldots,X_N) = M \sum_{j=1}^{N} I_k(j) \ .$$

When the variables are not mutually independent the total information measures then are

$$I_k(X_1,X_2\ldots,X_N) = I_k(X_1) + I_k(X_2|X_1) + I_k(X_3|X_1,X_2) + \ldots\ldots\ldots + I_k(X_N|X_1,X_2,\ldots,X_{N-1}) \quad , \ k = 1, 2. \tag{3.21}$$

where

$$I_1(X_1,X_2,\ldots,X_N) = - M \left[\sum_{k_1k_2\ldots k_N}^{r_1r_2\ldots r_N} P(x_{1k_1},x_{2k_2},\ldots,x_{Nk_N}) \right.$$

$$\left. \log P(x_{1k_1},x_{2k_2},\ldots,x_{Nk_N}) \right] \tag{3.22}$$

and

$$I_2(X_1,X_2,\ldots,X_N) = \log M! - \sum_{k_1k_2\ldots k_N}^{r_1r_2\ldots r_N} \log \left(\bigcap_{i=1}^{N} n_{ik_i} \right)! \tag{3.23}$$

$P(x_{1k_1},x_{2k_2},\ldots,x_{Nk_N})$ is the joint probability of an individual having the scores k_1 on variable X_1, k_2 on variable X_2 and so on up to k_N on X_N. $\bigcap_{i=1}^{N} n_{ik_i}$ is the number of joint occurrences of the scores k_1 on X_1 k_2 on X_2 up to k_N on X_N.

The denominator, $I_{k\ Max}$, (k = 1, 2), in the diversity measure remains unchanged since the maximum amount of information can be obtained from the data when the variables are independent . This follows from the inequality for the bivariate case, $I_k(X_1)+I_k(X_2|X_1) \leq I_k(X_1)+I_k(X_2)$, k = 1,2. (Ash 1965). The diversity measure in the case of nonindependent variables then is

$$D_k = [I_k(X_1) + I_k(X_2|X_1) + \ldots + I_k(X_N|X_1,X_2,\ldots,X_{N-1})]/\sum_{j=1}^{N} I_{k\ Max}(j) \qquad (3.24)$$

Thus the classification algorithm outlined here can be modified to handle data on nonindependent variables, without having to transform the data.

An Application Of The Algorithm

The algorithm has been used to discover groups within a sample of 198 individuals. Each individual was presented with seventy five statements representing five sets of fifteen statements on five underlying scales. The scales were related to (1) convenience orientation, (2) price orientation, (3) fashion orientation, (4) quality orientation and (5) status orientation of the individuals. These were considered to be five fundamental psychological orientations underlying the evaluation of shopping alternatives.

The objective of the study was to see if it was possible to discriminate between retail patronage groups in terms of their dispositional characteristics and further whether the dispositional measures provided more accurate discrimination than variables related to locational and socio-demographic characteristics of the consumers. Factor scores were derived from a factor analysis of the responses to the 75 statements by 198 individuals. The sample was composed of 114 psychology students and 84 shoppers. Of the 84 shoppers 51 were patrons of boutique stores, while the remaining 33 were patrons of department stores. From a varimax orthogonal rotation twelve significant factors (with eigenvalues > 1.0) emerged. They represent various combinations of the 75 statements and hence to varying degrees reproduce the original scales. The highest intercorrelation among the factors was 0.078, between factors 1 and 8.

The majority of the factor scores lay in the interval [-2,2]. Since the scores vary continuously it was necessary to convert them to ordinal form to make them suitable for the algorithm. The distribution of scores on each factor were partitioned into eight intervals. The intervals chosen and the ordinal score assigned to each are as follows.

SCORE	INTERVAL		
1	$(-\infty, -1.5]$	5	[0.0, 0.5)
2	(-1.5, -1.0]	6	[0.5, 1.0)
3	(-1.0, -0.5]	7	[1.0, 1.5)
4	(-0.5, 0.0)	8	$[1.5, \infty)$.

where

$$[0.5, 1.0) = \{ x \mid 0.5 \leq x < 1.0 \}$$

$$(-1.0, -0.5] = \{ x \mid -1.0 < x \leq -0.5 \}$$

Hence the factor scores were converted into 8-state ordinal attributes (variables). These data, listed in Appendix D, were used as input for the classification program. Classifications have been constructed for ALPHA at three levels, 16, 24 and 32, i.e. when the members of each group are similar to the nucleus of the group on at least 2, 3 or 4 attributes respectively. The classifications were performed using both Shannon's and Brillouin's information measures. In all of the cases the results from both information measures were identical.

Results

When Shannon's information measure is used the average information per individual on each of the twelve variables is as follows.

VARIABLE		VARIABLE	
1	0.829*	9	0.822
2	0.811	8	0.827
3	0.801	9	0.810
4	0.823	10	0.803
5	0.790	11	0.822
6	0.799	12	0.835

* The units of measurement are decits.

From these measures one may note the very limited variability among the variables. The maximum value that could occur in each case is 0.903 decits. One may conclude from these figures that each of the

12 variables are almost equally good as a differentiating characteristic (in the sense of Grigg 1965) of the sample. The diversity index for the sample is D = 0.902. Remembering that D ranges between zero and unity it is clear that this sample of individuals are very heterogeneous in the way they evaluate shopping alternatives.

In the first classification ALPHA was set at 16, i.e. it was required that the members of each group be similar to the nucleus of the group on at least two attributes. Three groups and two unclassifiable individuals emerge. The first group has 152 members. Its nucleus is individual number 29. The group members are

1	3	4	5	7	8	9	10	12	13	14
15	16	17	18	19	20	22	23	24	25	26
27	29	30	31	32	33	35	36	38	39	41
42	43	44	45	46	47	48	49	50	51	52
53	54	55	56	58	59	60	61	62	63	64
65	66	68	69	70	71	72	73	74	75	77
78	79	80	81	82	83	87	88	89	90	91
92	93	95	98	99	101	103	104	105	106	107
108	110	111	112	113	114	115	116	117	121	122
123	125	127	130	131	132	135	136	137	138	139
140	142	143	144	148	149	150	151	152	153	155
157	158	159	160	161	162	163	164	165	166	168
169	170	171	173	174	175	178	179	182	187	189
190	191	192	193	194	195	196	197	198.		

This group contains 94 students, 21 department store patrons and 37 boutique store patrons. Hence the first group takes 88.33% of all the students, 63.63% of the department store patrons and 72.55% of the boutique store patrons.

The second group has 33 members, of whom 16 are students, 8 department store patrons and 9 boutique store patrons. The group

members are as follows.

2	6	11	21	28	34	37	40	57	84	85
94	96	97	100	102	119	126	133	134	141	145
146	147	154	167	172	180	181	183	184	185	188.

The nucleus of this group is individual number 40, a student.

The third group has eleven members, the nucleus of the group being individual 124, a department store patron. The group members are

67 76 86 109 118 120 124 128 129 176 177.

Two individuals remain unclassified, they are individuals 156 and 186. Their pairwise similarity is less than 16, hence they remain as two single member groups. The main disadvantage of this classification is the fact that 76.77% of the individuals go into the first group.

A second classification was performed with ALPHA set at 24. The result is 11 groups, varying according to number of members from 90 to 1. However, the first six groups contain 191 of the individuals. The most typical individual in the sample is number 29, a student. The first group has 90 members, of whom 59 are students, 7 department store patrons and 24 boutique store patrons. Hence the first group contains 51.75% of the students, 21.21% of the department store patrons and 47.06% of the boutique store patrons. The members of the first group are as follows.

3	4	5	7	8	10	12	14	15	17	18	19	22
24	26	29	30	31	32	33	35	38	41	43	44	45
46	49	51	52	53	55	60	61	62	66	68	69	71
72	75	77	78	79	82	83	87	88	92	98	99	103
104	105	106	107	108	110	111	116	121	125	132	135	137
144	149	150	152	153	157	158	160	161	162	163	165	166
169	170	173	174	178	179	182	193	195	196	197	198.	

The second group has 61 members, of whom 31 are students, 16 department store patrons and 14 boutique store patrons. The group members are

2	11	13	25	34	37	39	40	42	47	48	54	56
57	58	63	67	70	73	80	81	84	89	90	91	93
94	96	101	109	113	115	118	119	123	124	126	127	128
130	131	133	134	138	141	142	147	148	151	154	155	159
171	172	175	183	185	187	189	190	194.				

The nucleus of this group is individual number 119, a department store patron. The group contains almost 50% of the department store patrons in the sample.

The third group has 17 members. Eleven are students, one is a department store patron and the other five are boutique store patrons. The nucleus of the group is individual number 180, a boutique store patron. The group members are as follows.

1 9 16 20 36 59 64 65 97 102 112 117 167 168 180 181 186.

These first three groups account for 168 of the individuals. The next three groups take 23 of the remaining 30 individuals. The last seven individuals are unclassifiable. The members of the fourth, fifth and sixth groups are as follows.

Group 4; 21 27 50 86 114 122 129 139 140 145 146 156

Group 5; 74 76 136 143 176

Group 6; 23 85 95 120 184 192

The interesting property of this classification is that the nuclei of the first three groups represent the three groups that are known a priori to be contained in the sample, i.e., the nucleus of the first group is a student, for the second group it is a department store patron and for the third group the nucleus is a boutique store patron. This classification results,in a manageable number of groups. Six groups take 191 of the individuals, and the remaining seven may be lumped into a seventh group of "nonconformists".

A third classificaiton was performed with ALPHA set at 32, i.e. it required that the members of each group be similar to the nucleus of the group on at least four attributes. The results were rather unsatisfactory. Altogether 27 groups were required, 13 of them being single member groups. The first four groups contained respectively 48, 36, 30 and 22 members. However, the fact that 22 groups are required for the remaining 62 individuals renders this classification unacceptable.

Therefore, of the three classifications the second one appears to be the most acceptable. This classification gives a manageable number of groups and furthermore the nuclei of the first three groups represent the three different groups of individuals that are known a priori to exist in the sample. Of course, different classifications would be obtained if the factor scores were partitioned

into different intervals. If instead of eight, only four intervals were used then it is very likely that an acceptable classification would be obtained at a much higher level of ALPHA. There is a trade-off involved here. By increasing the number of scores possible for each attribute one is increasing the amount of detail in the input data. However, the more detailed is the input data the less is the number of attributes on which one can expect a pair to be similar. Conversely, if the number of scores possible on each attribute is small then much of the detail has been dispensed with and it is reasonable to expect groups of individuals to emerge that are similar to the group nuclei on a relatively large number of attributes.

Some Further Applications Of The Algorithm

It was noted in the introduction that one of the main advantages of the algorithm outlined in this chapter is that it can treat ordinal data, without having to first transform it. Hence it is felt that the algorithm may be of special use when dealing with psychological data where it is safest to assume only ordinal properties. This algorithm may be used to classify individuals, given some data on their preferences (e.g. Residential preferences as in Gould and White, (1974)) or on their perceptions and attitudes as in hazard studies, such as Saarinen (1966). The data may consist of a collection of statements of opinion with a numerical value assigned to each statement, (e.g. Likert scale data), as was the case in the example given, or the data may be obtained from semantic-differential scales presented to the individuals. The data may relate to biographical

characteristics such as an individual's education, religion, age, family status, socio-economic status, zone of work or residence etc. A further possible application of the algorithm would be in relation to profile analysis of individuals, as an alternative to the techniques outlined by Nunnally (1967), which require data with a metric structure.

Summary

In this chapter a classification algorithm based on some concepts from information theory has been outlined. An advantage of the algorithm is that it does not require data with metric properties. The total information content of the data available on a sample of individuals is calculated and a diversity index for the sample is computed. The most typical individual in the sample is identified. This individual becomes the nucleus of the first group which contains all the individuals who have a similarity with the most typical one greater than some critical level. When the first group has been formed it is removed from the sample, and the algorithm is applied to the remaining individuals. The process continues until every individual is classified. A computer program has been written for the algorithm, a listing of it is given in Appendix C. The computational steps involved were illustrated via a simple example and then the algorithm was applied to a more realistic data set. Finally, a wide range of possible applications of interest to the behavioral geographer were indicated for the algorithm.

CHAPTER 4

SUMMARY AND CONCLUSIONS

The purpose of this paper was to examine the potentialities of information theory for the construction of classifications in geography. With this purpose in mind the first chapter was devoted to a discussion of some pertinent concepts and measures that are prevalent in the literature of information theory. The concepts of information and uncertainty were shown to be equivalent. Three different information measures were discussed and the different situations where each measure is applicable were identified. In particular, it was argued that Shannon's measure is defined for very large populations, while Brillouin's measure applies strictly to completely sampled populations. The argument was based on a new derivation of Brillouin's measure. The Shannon and Brillouin measures were generalized to measure the amount of information needed to describe objects characterized by a number of nonindependent attributes.

In the second chapter the classifciation problem was posed. The information based classification procedures that are being used by ecologists and biologists were reviewed and it was indicated how these procedures could be used for spatial classification. A further information based classification algorithm was given in chapter three. The algorithm classifies individuals characterized by scores on multistate ordinal attributes. It is felt that this algorithm could be of

relevance to behavioral geographers, who oftentimes have to manipulate large quantities of data obtained from questionnaires, data which may lack metric properties.

From this paper there can be little doubt that information theory provides us with a host of statistics applicable to classification problems in geography. It is also abundantly clear that as yet geographers have failed to take advantage of these potentialities of information statistics. The basic concept underlying the measures given in this paper is that they measure the uncertainty present in a distribution. In this context uncertainty refers to the probability of events in a series on the basis of already observed occurrences. Throughout the paper it has been emphasized that different information measures are appropriate in different situations.

In comparison with analysis of variance techniques information statistics have a wider generality. In analysis of variance type applications uncertainties are calculated from the variance, if the form of the distribution is known. However, there are some cases where even given the form of the distribution it may not be possible to calculate its variance, or the variance may be infinite as in the case of the Cauchy distribution (Meyer 1975, p. 252). Information theory provides us with a method of measuring the uncertainty in such distributions, Novitskii (1966). Information statistics are more general in that they can be equally easily calculated for nominal and ordinal data as well as for interval or ratio scale data. While the analysis of variance technique may be less general it retains information about the metric properties of the data, if these properties exist. On the other hand,

information statistics do not provide any information about the metric properties of the data, even if they do exist.

The information statistics presented throughout the paper are purely descriptive measures. While they may give some indication as to the nature of the process that is at work in a particular situation, they do not describe the process. Some indication as to the nature of the process at work may be obtained if the observed information measures for a distribution are related to the measures that would be obtained from a distribution predicted by some theory. If there is a close correspondence between the measures from the observed and the theoretical distributions, then one may infer that the process which the theory explains is also the process that is actually occurring. Statements to the effect that information measures, or entropy measure the randomness or disorder in a distribution should be made with caution. The important question is what kind of randomness or disorder is being measured. This requires a clear specification of the system being studied.

Further work along this line could deal with some of the following problems. Some of the statistics given in this paper, particularly those in chapter two, may need to be modified along the lines suggested by Curry (1972) and Sheppard (1975). Alternate information metrics, such as those by Renyi (1967) or Behara and Nath (1973) may warrant examination. Consideration should also be given to the application of information theory to other statistical problems in geography.

APPENDIX A

An example to illustrate the process that leads to Brillouin's information measure.

Consider a message composed of the following sequence of symbols,

A A B C E D B D A C E D B A C B E A .

Denote the number of A's, B's, C's, D's and E's by n_1, n_2, n_3, n_4 and n_5 respectively. Then $n_1 = 5$; $n_2 = 4$; $n_3 = 3$; $n_4 = 3$; and $n_5 = 3$ such that $\sum_{i=1}^{5} n_i = N = 18$.

Let K denote the number of symbols preceding any given symbol. Denote the information content of the symbol A by I(A). I(A) is a function of the positon of A in the sequence. To calculate the information content of the sequence of symbols above proceed as follows.

When

K = 0	I(A) = - log 5/18	= log (N-0) - log (n_1-0)
K = 1	I(A) = - log 4/17	= log (N-1) - log (n_1-1)
K = 2	I(B) = - log 4/16	= log (N-2) - log (n_2-0)
K = 3	I(C) = - log 3/15	= log (N-3) - log (n_3-0)
K = 4	I(E) = - log 4/14	= log (N-4) - log (n_5-0)
K = 5	I(D) = - log 3/13	= log (N-5) - log (n_4-0)
K = 6	I(B) = - log 3/12	= log (N-6) - log (n_2-1)
K = 7	I(D) = - log 2/11	= log (N-7) - log (n_4-1)
K = 8	I(A) = - log 3/10	= log (N-8) - log (n_1-2)
K = 9	I(C) = - log 2/9	= log (N-9) - log (n_3-1)
K = 10	I(E) = - log 2/8	= log (N-10)- log (n_5-1)
K = 11	I(D) = - log 1/7	= log (N-11)- log (n_4-2)
K = 12	I(B) = - log 2/6	= log (N-12)- log (n_2-2)
K = 13	I(A) = - log 2/5	= log (N-13)- log (n_1-3)
K = 14	I(C) = - log 1/4	= log (N-14)- log (n_3-2)
K = 15	I(B) = - log 1/3	= log (N-15)- log (n_2-3)
K = 16	I(E) = --log 1/2	= log (N-16)- log (n_5-2)
K = 17	I(A) = - log 1/1	= log (N-17)- log (n_1-4)

Summing over the whole sequence, the total information content of the 18 symbols is

$$B = \sum_{k=0}^{N-1} \log (N-K) - \sum_{i=1}^{5} \sum_{k_i=0}^{n_i-1} \log (n_i-k_i)$$

$$= \log N! - \sum_{i} \log n_i!$$

The mean information content per symbol is

$$I_2 = \frac{B}{N} = \frac{1}{N} [\log N! - \sum_{i} \log n_i!] \quad .$$

APPENDIX B

First a multivariate measure corresponding to Shannon's univariate measure, $I_1 = - \sum_i P_i \log P_i$ is given.

Definition (1) : (S, A, P) is a probability measure space when S is a set of sample points, A is an algebra of events and P is a probability measure.

Definition (2) : $A = \{ X = x_{ik_i} \mid i=1, 2, \ldots, s \; ; \; k_i=1, 2, \ldots r_i \; ; \; s, r_i \in \mathbf{Z}^+ \}$.

Definition (3) : On the probability measure space (S, A, P) the information content of any event $X = x_{ik_i}$ is equal to $- \log P(x_{ik_i})$, where $P(x_{ik_i}) = \text{Prob } (X = x_{ik_i})$.

Theorem (I,a) : Let $I_1(X_1, X_2, \ldots, X_s)$ be a function defined for any integer s and for all values $P(x_{ik_i})$, such that $P(x_{ik_i}) \geq 0$ $(i = 1, 2, \ldots, s \; ; \; k_i = 1, 2, \ldots, r_i)$ and

$$\sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \cdots \sum_{k_s=1}^{r_s} P(x_{1k_1}, x_{2k_2}, \ldots, x_{sk_s}) = 1 \quad .$$

If for any s this function is continuous with respect to all of its arguments then for any collection of s-dimensional vectors the mean information content per vector is

$$I_1(X_1, X_2, \ldots, X_s) = I_1(X_1) + I_1(X_2|X_1) + I_1(X_3|X_1, X_2) + \ldots$$

$$+ I_1(X_s|X_1, X_2, \ldots, X_{s-1}) \quad .$$

where

$$I_1(X_1,X_2,\ldots,X_s) = -\sum_{k_1k_2,\ldots k_s}^{r_1r_2\ldots r_s} P(x_{1k_1},x_{2k_2},\ldots,x_{sk_s}) \cdot \log P(x_{1k_1},x_{2k_2},\ldots,x_{sk_s}) .$$

When the s distributions represented by the s-dimensional vectors are mutually independent $I_1(X_1,X_2,\ldots,X_s) = \sum_{i=1}^{s} I_1(X_i)$.

Proof: The proof is by induction.

Let s = 1, then

$$I_1(X_1) = -\sum_{k_1=1}^{r_1} P(x_{1k_1}) \log P(x_{1k_1})$$

$$= I_1(X_1) \text{ , by definition (3) .}$$

Let s = 2,

$$I_1(X_1,X_2) = -\sum_{k_1=1}^{r_1}\sum_{k_2=1}^{r_2} P(x_{1k_1},x_{2k_2}) \log P(x_{1k_1},x_{2k_2})$$

$$= -\sum_{k_1}\sum_{k_2} P(x_{1k_1},x_{2k_2}) \log P(x_{1k_1}) P(x_{2k_2}|x_{1k_1})$$

$$= -\sum_{k_1k_2} P(x_{1k_1},x_{2k_2}) \log P(x_{1k_1}) - \sum_{k_1}\sum_{k_2} P(x_{1k_1},x_{2k_2}) \cdot \log P(x_{2k_2}|x_{1k_1})$$

$$= -\sum_{k_1} P(x_{1k_1}) \log P(x_{1k_1}) - \sum_{k_1} P(x_{1k_1}) \sum_{k_2} P(x_{2k_2}|x_{1k_1}) \cdot \log P(x_{2k_2}|x_{1k_1})$$

$$= I_1(X_1) + I_1(X_2|X_1) .$$

Assume the theorem is true for s = n-1

$$I_1(X_1,X_2,\ldots,X_{n-1}) = I_1(X_1) + I_1(X_2|X_1) + \ldots. + I_1(X_{n-1}|X_1,X_2,\ldots X_{n-2}) .$$

Let s = n, then

$$I_1(X_1,X_2,..,X_n) = - \sum_{\substack{r_1,r_2,..,r_n \\ k_1k_2,..,k_n}} P(x_{1k_1},x_{2k_2},..,x_{nk_n}) \log P(x_{1k_1},x_{2k_2},..,x_{nk_n})$$

$$= - \sum_{k_1k_2,..,k_n} P(x_{1k_1},x_{2k_2},\ldots,x_{nk_n})$$

$$\log P(x_{1k_1},x_{2k_2},\ldots,x_{(n-1)k_{(n-1)}})$$

$$P(x_{nk_n}|x_{1k_1},x_{2k_2},\ldots,x_{(n-1)k_{(n-1)}})$$

$$= - \sum_{k_1k_2,..,k_{(n-1)}} P(x_{1k_1},x_{2k_2},\ldots,x_{(n-1)})$$

$$\log P(x_{1k_1},x_{2k_2},\ldots,x_{(n-1)k_{(n-1)}})$$

$$- \sum_{k_1k_2,..,k_{(n-1)}} P(x_{1k_1},x_{2k_2},\ldots,x_{(n-1)})$$

$$\sum_{k_n} P(x_{nk_n}|x_{1k_1},x_{2k_2},\ldots,x_{(n-1)k_{(n-1)}})$$

$$\log P(x_{nk_n}|x_{1k_1},\ldots x_{(n-1)k_{(n-1)}})$$

$$= I_1(X_1,X_2,\ldots,X_{(n-1)}) + I_1(X_n|X_1,X_2,\ldots X_{(n-1)}) .$$

Hence for any integer s

$$I_1(X_1,X_2,\ldots,X_s) = I_1(X_1) + I_1(X_2|X_1) + \ldots + I_1(X_s|X_1,X_2,\ldots,X_{(s-1)}).$$

where, for any ℓ, $2 \leq \ell \leq s$

$$I_1(X_\ell|X_1,X_2,\ldots,X_{(\ell-1)}) = - \sum_{k_1k_2\ldots k_\ell} P(x_{1k_1},x_{2k_2},\ldots,x_{\ell k_\ell}) \cdot$$

$$\log P(x_{\ell k_\ell}|x_{1k_1},\ldots,x_{(\ell-1)k_{(\ell-1)}}) .$$

In the special case when the s distributions, represented by the s-dimensional vectors, are mutually independent

$$I_1(X_1,X_2,\ldots,X_s) = - \sum_{k_1k_2\ldots k_s} P(x_{1k_1})\, P(x_{2k_2})\ldots P(x_{sk_s}) \log P(x_{1k_1}) \cdot$$

$$P(x_{2k_2})\ldots P(x_{sk_s})$$

$$= - \sum_{i=1}^{s} I_1(X_i) .$$

Theorem (1.a) may be interpreted as follows, the mean amount of information needed to select the s entries in a particular vector is equal to the mean amount of information needed to select the first entry plus the mean amount needed to select the second entry given the first one, plus the mean amount needed to select the third entry given the first and second entries, and so on. When the distributions are mutually independent the mean amount of information needed to select the n entries will be equal to the sum of the mean amounts needed to select each entry.

Now a multivariate information measure corresponding to Brillouin's univariate measure is given.

Definition (4) : (Brillouin's Information Measure). For any collection of N events such that n_1 are of type 1, n_2 are of type 2 up to n_{r_i} of type r_i such that $\sum_{i=1}^{r_i} n_i = N$, the mean information content per event is $I_2 = \frac{1}{N} \log (N! / \prod_{i=1}^{r_i} n_i!)$.

Some Notation.

(1) $\bigcap_{i=1}^{s} n_{ik_i} = n_{1k_1} \cap n_{2k_2} \cap \ldots \cap n_{sk_s}$.

(2) $\prod_{i=1}^{s} \{ \sum_{k_i=1}^{r_i} (\bigcap_{i=1}^{s} n_{ik_i}) \} = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \cdots\cdots \sum_{k_s=1}^{r_s} (\bigcap_{i=1}^{s} n_{ik_i})$.

(3) $\prod_{i=1}^{s} \{ \prod_{k_i=1}^{r_i} (\bigcap_{i=1}^{s} n_{ik_i}) \} = \prod_{k_1=1}^{r_1} \prod_{k_2=1}^{r_2} \cdots\cdots \prod_{k_s=1}^{r_s} (\bigcap_{i=1}^{s} n_{ik_i})$.

Theorem (1,b) : In any collection of n s-dimensional vectors let n_{ik_i} be the number of times the integer x_{ik_i}, $k_i = 1, 2, .., r_i$, occurs as the $i^{\underline{th}}$ entry in each of the n vectors, $n_{ik_i} \cap n_{jk_j}$ be the number of joint occurrences of the integer x_{ik_i} in the $i^{\underline{th}}$ place and the integer x_{jk_j} in the $j^{\underline{th}}$ place of each vector, such that

$$\prod_{\substack{i=1 \\ i \neq \ell}}^{s} \{ \sum_{k_i=1}^{r_i} (\bigcap_{i=1}^{s} n_{ik_i}) \} = n_{\ell k_\ell} \quad , \quad \ell = 1, 2, \ldots, s$$

and $\prod_{i=1}^{s} \{ \sum_{k_{i=1}}^{r_i} (\bigcap_{i=1}^{s} n_{ik_i}) \} = n$. Then the mean information

content per vector is $I_2(X_1,X_2,.....,X_s) =$

$I_2(X_1) + I_2(X_2|X_1) + ..., + I_2(X_s|X_1,X_2,..,X_{(s-1)})$.

where

$$I_2(X_1,X_2,...,X_s) = \frac{1}{n} \log[\, n!/ \prod_{i=1}^{s} \{ \prod_{k_i=1}^{r_i} (\bigcap_{i=1}^{s} n_{ik_i})!\}].$$

When the s distributions represented by the s-dimensional vectors are mutually independent,

$$I_2(X_1,X_2,...,X_s) = \sum_{i=1}^{s} I_2(X_i) .$$

Proof: The proof is by induction.

Let s = 1, then

$$I_2(X_1) = \frac{1}{n} \log \frac{n!}{\prod_{k_1=1}^{r_1} n_{1k_1}!}$$

$$= I_2(X_1) \text{ , by definition (4).}$$

Let s = 2,

$$I_2(X_1,X_2) = \frac{1}{n} \log[\frac{n!}{\prod_{k_1=1}^{r_1} \prod_{k_2=1}^{r_2} (n_{1k_1} \cap n_{2k_2})!}]$$

$$= \frac{1}{n} \log[\frac{n!}{\prod_{k_1} (n_{1k_1})!} \cdot \frac{\prod_{k_1} (n_{1k_1})!}{\prod_{k_1} \prod_{k_2} (n_{1k_1} \cap n_{2k_2})!}]$$

$$= \frac{1}{n} \log \left[\frac{n!}{\prod_{k_1} (n_{1k_1})!} \right] + \frac{1}{n} \log \left[\prod_{k_1} \frac{(n_{1k_1})!}{\prod_{k_2} (n_{1k_1} \cap n_{2k_2})!} \right]$$

$$= \frac{1}{n} \log \left[\frac{n!}{\prod_{k_1} (n_{1k_1})!} \right] + \sum_{k_1} \frac{1}{n_{1k_1}} \log \left[\frac{(n_{1k_1})!}{\prod_{k_2} (n_{1k_1} \cap n_{2k_2})!} \right]$$

$$= I_2(X_1) + \sum_{k_1} \frac{n_{1k_1}}{N} \left(\frac{1}{n_{1k_1}} \log \left[\frac{(n_{1k_1})!}{\prod_{k_2} (n_{1k_1} \cap n_{2k_2})!} \right] \right)$$

$$= I_2(X_1) + I_2(X_2|X_1) \ .$$

Assume the theorem is true for s = n-1,

$$I_2(X_1,X_2,\ldots,X_{n-1}) = \frac{1}{n} \log \left[\frac{n!}{\prod_{i=1}^{n-1} \{ \prod_{k_i=1}^{r_i} (\bigcap_{i=1}^{n-1} n_{ik_i})! \}} \right]$$

$$= I_2(X_1) + I_2(X_2|X_1) + \ldots + I_2(X_{n-1}|X_1,X_2,\ldots,X_{n-2})$$

Let s = n, then

$$I_2(X_1,X_2,\ldots,X_n) = \frac{1}{n} \log \left[\frac{n!}{\prod_{i=1}^{n} \{ \prod_{k_i=1}^{r_i} (\bigcap_{i=1}^{n} n_{ik_i})! \}} \right]$$

$$= \frac{1}{n} \log \left[\frac{n!}{\prod_{i=1}^{n-1} \{ \prod_{k_i=1}^{r_i} (\bigcap_{i=1}^{n-1} n_{ik_i})! \}} \cdot \right.$$

$$\left. \prod_{i=1}^{n-1} \{ \prod_{k_i=1}^{r_i} \frac{(\bigcap_{i=1}^{n-1} n_{ik_i})!}{\prod_{k_n=1}^{r_n} (\bigcap_{i=1}^{n} n_{ik_i})!} \} \right]$$

$$= \frac{1}{n} \log \left[\frac{n!}{\prod_{i=1}^{n-1} \{ \prod_{k_i=1}^{r_i} (\bigcap_{i=1}^{n-1} n_{ik_i})! \}} \right] +$$

$$\sum_{k_1 k_2 \ldots k_{n-1}} \frac{(\bigcap_{i=1}^{n-1} n_{ik_i})}{n} \left[\frac{1}{\bigcap_{i=1}^{n-1} n_{ik_i}} \log \frac{(\bigcap_{i=1}^{n-1} n_{ik_i})!}{\prod_{k_n} (\bigcap_{i=1}^{n} n_{ik_i})!} \right]$$

$$= I_2(X_1, X_2 \ldots\ldots X_{n-1}) + I_2(X_n | X_1, X_2, \ldots\ldots X_{n-1}) \ .$$

Hence for any integer s

$$I_2(X_1, X_2, \ldots\ldots, X_s) = I_2(X_1) + I_2(X_2 | X_1) + I_2(X_3 | X_1, X_2)$$

$$+ \ldots\ldots + I_2(X_s | X_1, X_2, \ldots X_{s-1}) \ .$$

where for any ℓ, $2 \leq \ell \leq s$

$$I_2(X_\ell | X_1, X_2, \ldots\ldots, X_{\ell-1}) = \frac{1}{n} \left[\sum_{k_1 k_2 \ldots k_{\ell-1}} \log (\bigcap_{i=1}^{\ell-1} n_{ik_i})! \right.$$

$$\left. - \sum_{k_1 k_2 \ldots k_\ell} \log (\bigcap_{i=1}^{\ell} n_{ik_i})! \right]$$

When the s distributions are mutually independent the mean information content per vector is

$$I_2(X_1, X_2, \ldots, X_s) = \frac{1}{n} \log \left[\prod_{i=1}^{s} \left\{ \frac{n_i!}{\prod_{k_i=1}^{r_i} (n_{ik_i})!} \right\} \right]$$

$$= \sum_{i=1}^{s} I_2(X_i) \quad \text{where} \quad n = n_i = \sum_{k_i=1}^{r_i} n_{ik_i} \ .$$

The result of theorem (1.b) is interpreted in exactly the same way as the result of theorem (1.a).

Lemma 1: For large values of n_{ik_i}, $I_2(X_1,X_2,\ldots,X_s)$ is a good approximation of $I_1(X_1,X_2,\ldots,X_s)$.

Proof: $I_2(X_1,X_2,\ldots,X_s) = \frac{1}{n} \log \left[n!/\prod_{k_1} \prod_{k_2} \ldots\ldots \prod_{k_s} (\bigcap_{i=1}^{s} n_{ik_i})! \right]$.

If, and only if, all the n_{ik_i} are large Stirling's approximation, $\log x! = x \log x - x$, can be applied.

Then

$$I_2(X_1,X_2,\ldots,X_s) = \frac{1}{n} \{ \log n! - \sum_{k_1k_2\ldots k_s} \log (\bigcap_{i=1}^{s} n_{ik_i})! \}$$

$$= \frac{1}{n} \{ n \log n - n - \sum_{k_1k_2\ldots k_s} [(\bigcap_{i=1}^{s} n_{ik_i}) \log(\bigcap_{i=1}^{s} n_{ik_i}) - \bigcap_{i=1}^{s} n_{ik_i}] \}$$

$$= \log n - \sum_{k_1k_2\ldots k_s} \frac{(\bigcap_{i=1}^{s} n_{ik_i})}{n} \log(\bigcap_{i=1}^{s} n_{ik_i})$$

$$= - \sum_{k_1k_2\ldots k_s} \frac{\bigcap_{i=1}^{s} n_{ik_i}}{n} \log \frac{\bigcap_{i=1}^{s} n_{ik_i}}{n}$$

$$= - \sum_{k_1k_2\ldots k_s} P(x_{1k_1},x_{2k_2},\ldots,x_{sk_s}) \log P(x_{1k_1},x_{2k_2},\ldots,x_{sk_s})$$

$= I_1(X_1,X_2,\ldots,X_s)$, when the probabilities are defined in a relative frequency manner.

```
C     ****************************************************************
C     PROGRAM CLASSINF  WRITTEN BY J A WALSH AND M J WEBBER,MCMASTER
C     UNIVERSITY,HAMILTON,ONTARIO.
C
C     INFORMATION THEORETIC AGGLOMERATIVE CLASSIFICATION ALGORITHM
C
C     THIS ALGORITHM FIRST IDENTIFIES THE MOST TYPICAL INDIDIDUAL IN THE
C     SET TO BE PARTITIONED. ALL THE OTHER INDIVIDUALS ARE COMPARED WITH
C     THE MOST TYPICAL . THOSE WITH A SIMILARITY INDEX GREATER THAN ALPHA
C     FORM THE FIRST CLUSTER. THE ALGORITHM IS THEN REAPPLIED TO THE
C     REMAINING SET OF INDIVIDUALS UNTIL EVERYONE IS ASSIGNED TO A CLASS
C     ALPHA IS AN ARBITRARY CUT-OFF POINT.
C
C     FOR EACH INDIVIDUAL AN IDENTIFICATION NUMBER IS PUNCHED IN
C     COLUMNS 1 TO 5.
C     NO IS THE NUMBER OF INDIVIDUALS TO BE CLASSIFIED.
C     NV IS THE NUMBER OF VARIABLES.
C     NVV IS THE NUMBER OF VARIABLES PLUS THE IDENTIFICATION NUMBER
C     FOR EACH INDIVIDUAL.
C
C     RJ(I) IS THE NUMBER OF DIFFERENT VALUES THAT CAN OCCUR FOR
C     VARIABLE I. THE VALUES MUST BE POSITIVE INTEGERS GREATER THAN
C     ZERO AND LESS THAN TEN.
C     MAXK IS THE MAXIMUM OF THE RJ,S.
C
C     INDATA IS THE MATRIX OF INPUT DATA.
C     ENTRY(I,J) IS THE SCORE OF INDIVIDUAL I ON VARIABLE J.
C     SOUT IS A MATRIX OF PROBABILITIES BASED ON RELATIVE FREQUENCIES
C     FOR OPT1 AT 0
C     SOUT IS A MATRIX OF FREQUENCIES FOR OPT1 AT 1.
C     SOUT HAS DIMENSIONS (NV,MAXK)
C
C     S(I) IS THE MEAN INFORMATION CONTENT OF EACH VALUE RECORDED FOR
C     VARIABLE I.
C     TOTS IS THE TOTAL INFORMATION CONTENT OF INDATA .
C     TOTSM IS THE MAXIMUM INFORMATION CONTENT OF INDATA.
C     DIV IS THE DIVERSITY OF THE SET  OF INDIVIDUALS. IT IS THE RATIO
C     OF  TOTS OVER TOTSM.
C     X IS THE DIVERSITY WHEN ONE INDIVIDUAL IS OMITTED FROM INDATA.
C     TYP(I) IS THE TYPICALITY OF INDIVIDUAL I.
C
C     OPT1 IS SET AT 0 FOR SHANNONS MEASURE OF INFORMATION.
C     OPT1 IS SET AT 1 FOR BRILLOUINS MEASURE OF INFORMATION
C
C     GINDEX DENOTES A GROUP
C     INDEX DENOTES THE NUCLEUS OF THE GROUP
C     JINDEX DENOTES THE NUMBER OF INDIVIDUALS IN THE GROUP
C
C     WHEN THERE ARE ONLY TWO INDIVIDUALS REMAINING TO BE CLASSIFIED
C     THE INFORMATION AND DIVERSITY MEASURES ARE NOT CALCULATED.THE
C     SIMILARITY OF THE PAIR IS CALCULATED. IF IT IS GREATER THAN ALPHA
C     THEY FORM THE LAST GROUP.OTHERWISE TWO SEPARATE GROUPS ARE FORMED.
```

```
C
C     OUTPUT.
C
C     INDATA
C     THE VALUE OF ALPHA
C     SOUT
C     S(I),I=1,NV
C     TOTS,TOTSM,DIV
C     THE MOST TYPICAL INDIVIDUAL
C     GROUP NUMBER AND GROUP MEMBERS
C     **************************************************************
C
      DIMENSION TYP(200),NGROUP(200),INTEMP(200,20),LIST(200)
      DIMENSION LOGFAC(20),LGFACN(20),LGFACM(20),A(20),B(20)
      COMMON INDATA(200,20),FREQ(20,10),S(20)
      COMMON RJ(20),SIM(200)
      INTEGER GINDEX,O,OPT1,OPT2
C
C     READ THE DATA VALUES
C
      READ(5,1)NV,NO,MAXK
      READ(5,2) OPT1,ALPHA
      DO 100 I=1,NO
  100 INDATA(I,1)=I
      NW =NV
      NVV=NV+1
      READ(5,3) (RJ(I),I=2,NVV)
      DO 110 I=1,NO
  110 READ(5,4) (INDATA(I,J),J=2,NVV)
      WRITE(6,5) NW,NO
      IF(OPT1.EQ.1) GO TO 111
      WRITE(6,6)
      GO TO 112
  111 WRITE(6,7)
  112 WRITE(6,8)
      WRITE(6,10)
      DO 115 I=1,NO
      K=I
  115 WRITE(6,11) K,(INDATA(I,J),J=2,NVV)
      WRITE(6,9) ALPHA
      GINDEX=0
  205 CONTINUE
      DO 117 I=1,NO
  117 NGROUP(I)=0
C
C     CALCULATE FREQUENCY AND RELATIVE FREQUENCY MATRICES
C
      CALL SFREQ(0,NO,NVV,OPT1,MAXK)
      GINDEX=GINDEX+1
      IF(NO.EQ.2) GO TO 204
C
C     COMPUTE DIVERSITY INDEX
C
```

```
      CALL DIVERS(0,NO,NVV,OPT1,MAXK,TOTS,TOTSM,DIV)
      WRITE(6,12)
      WRITE(6,13) (S(I),I=2,NVV)
      WRITE(6,14)
      WRITE(6,15) TOTS, TOTSM,DIV
C
CCC   CALCULATE THE TYPICALITIES
C
      DO 200 I=1,NO
      CALL SFREQ(I,NO,NVV,OPT1,MAXK)
      CALL DIVERS(1,NO,NVV,OPT1,MAXK,TOTS,TOTSM,X)
  200 TYP(I)=X-DIV
      TMAX=-2
      DO 201 I=1,NO
      IF(TYP(I).LE.TMAX) GO TO 201
      TMAX=TYP(I)
      INDEX=I
  201 CONTINUE
  204 CONTINUE
C
CCC   CALCULATE SIMILARITIES
C
      CALL SIMIL(INDEX,NO,NVV,SIM)
      JINDEX=1
      NGROUP(JINDEX)=INDEX
      DO 202 I=1,NO
      IF(I.EQ.INDEX) GO TO 202
      IF(SIM(I).LT.ALPHA) GO TO 202
      JINDEX=JINDEX+1
      NGROUP(JINDEX)=I
  202 CONTINUE
      L=1
      WRITE(6,16)GINDEX,INDEX,JINDEX
      DO 203 I=1,JINDEX
      N=NGROUP(I)
  203 LIST(I)=INDATA(N,1)
      WRITE (6,18) (LIST(I),I=1,JINDEX)
C
CCC   REDEFINE INDATA LEAVING OUT GROUPED INDIVIDUALS AND REPEAT
C
      K=1
      DO 210 I=1,NO
      IMP=0
      DO 211 J=1,JINDEX
      IF(NGROUP(J).NE.I) GO TO 211
      IMP=IMP+1
  211 CONTINUE
      IF(IMP.EQ.1) GO TO 210
      DO 215 L=1,NVV
  215 INTEMP(K,L)=INDATA(I,L)
      K=K+1
  210 CONTINUE
      NO=NO-JINDEX
      DO 216 I=1,NO
      DO 216 J=1,NVV
```

```
      1    1    2    2    3    3    4    4    5    5    6    6    7    7    8
....5....0....5....0....5....0....5....0....5....0....5....0....5....0....5....0
  216 INDATA(I,J)=INTEMP(I,J)
      IF(NO.GT.2) GO TO 205
      NO=NO+1
      DO 207 I=2,NO
      N=NGROUP(I)
  207 LIST(I)=INDATA(N,1)
      WRITE (6,17)
      WRITE(6,19) (LIST(I),I=2,NO)
C
CC    WRITE FORMAT STATEMENTS
C
    1 FORMAT(5X,3I3)
    2 FORMAT(6X,I1,F3.1)
    3 FORMAT(5X,30F2.0)
    4 FORMAT(5X,30I2)
    5 FORMAT(1H1,5X,50HINFORMATION THEORETIC AGGLOMERATIVE CLASSIFICATIO
     1N/1H0,5X,13HUSING DATA ON,I3,1X,9HVARIABLES/1H ,5X,3HAND,I3,1X,11H
     2INDIVIDUALS)
    6 FORMAT(1H0,5X,50HTHE PROCEDURE USES SHANNONS MEASURE OF INFORMATIO
     1N)
    7 FORMAT(1H0,5X,52HTHE PROCEDURE USES BRILLOUINS MEASURE OF INFORMAT
     1ION)
    8 FORMAT(1H0,5X,68HTHE PROCEDURE USES A MODIFIED VERSION OF HYVARINE
     1N'S SIMILARITY INDEX)
    9 FORMAT(1H0,5X,21HTHE VALUE OF ALPHA IS,F20.3)
   10 FORMAT(1H0,5X,29HTHE INPUT DATA ARE AS FOLLOWS)
   11 FORMAT(1H0,2X,I3,5X,20I3)
   12 FORMAT(1H0,5X,69HTHE AVERAGE INFORMATION PER INDIVIDUAL ON EACH VA
     1RIABLE IS AS FOLLOWS)
   13 FORMAT(1H0,5X,F20.6)
   14 FORMAT(1H0,5X,70HTHE TOTAL INFORMATION,MAXIMUM INFORMATION AND DIV
     1ERSITY ARE AS FOLLOWS)
   15 FORMAT(1H0,3F20.6)
   16 FORMAT(1H0,5X,16HFOR GROUP NUMBER,I3,1X,30HTHE MOST TYPICAL INDIVI
     1DUAL IS,I3,1X,13HTHE GROUP HAS I3,1X,8HMEMBERS.)
   17 FORMAT(1H0,5X,33HTHE MEMBERS OF THE LAST GROUP ARE)
   18 FORMAT(1H0,5X,17HGROUP MEMBERS ARE /(1H ,5X,20I4))
   19 FORMAT(1H0,5X,20I4)
      STOP
      END
C
      SUBROUTINE SFREQ(JO,N1,N2,N3,MAXK)
C     DO SUBROUTINE SFREQ TO CALCULATE FREQUENCIES AND RELATIVE FREQUE
CC    NCIES. FOR OPT1=0 THE OUTPUT WILL BE A MATRIX OF PROBABILITIES.
CC    ENTRY(J,K) IS THE PRIOR PROBABILITY OF ANY INDIVIDUAL HAVING THE
CC    SCORE K ON VARIABLE J. FOR OPT1=1 THE OUTPUT WILL BE A MATRIX OF
CC    FREQUENCIES. ENTRY(J,K) IS THE NUMBER OF TIMES THE SCORE K OCCURS
CC    ON VARIABLE J.
C
      COMMON IN(200,20),SOUT(20,10)
      IF(N3.EQ.0) GO TO 20
      DO 1 I=2,N2
      DO 1 J=1,MAXK
    1 SOUT(I,J)=0.
      IF(J.EQ.0) GO TO 40
      1    1    2    2    3    3    4    4    5    5    6    6    7    7    8
....5....0....5....0....5....0....5....0....5....0....5....0....5....0....5....0
```

```
      F=1.0
      DO 3 J=1,N1
      IF(J.EQ.JO) GO TO 3
      DO 4 I=2,N2
      JSCORE=IN(J,I)                                                    225
 4    SOUT(I,JSCORE)=SOUT(I,JSCORE)+F
 3    CONTINUE
      IF(JO.NE.0) GO TO 40
      WRITE (6,6)
      DO 7 I=2,N2                                                       230
      K=I-1
 7    WRITE(6,8) K, (SOUT(I,J),J=1,MAXK)
      GO TO 40
C
 20   DO 30 I=2,N2                                                      235
      DO 30 J=1,MAXK
 30   SOUT(I,J)=0.
      Z=FLOAT(N1)
      IF(JO.EQ.0) GO TO 31
      Z=Z-1                                                             240
      IF(Z.LE.0) GO TO 32
 31   RF=1./Z
 32   CONTINUE
      DO 33 J=1,N1
      IF(J.EQ.JO) GO TO 33                                              245
      DO 34 I=2,N2
      JSCORE=IN(J,I)
 34   SOUT(I,JSCORE)=SOUT(I,JSCORE)+RF
 33   CONTINUE
      IF(JO.NE.0) GO TO 40                                              250
      WRITE(6,35)
      DO 36 I=2,N2
      K=I-1
 36   WRITE(6,37) K,(SOUT(I,J),J=1,MAXK)
 40   CONTINUE                                                          255
C
CC    WRITE FORMAT STATEMENTS
C
 6    FORMAT(1H1,5X,63HTHE MATRIX OF FREQUENCIES HAS VARIABLES ACROSS AN
     1D SCORES DOWN /1H0,8HVARIABLE)                                    260
 8    FORMAT(1H0,2X,I3,2X,(10F10.3))
 35   FORMAT(1H1,5X,71HTHE MATRIX OF RELATIVE FREQUENCIES HAS VARIABLES
     1ACROSS AND SCORES DOWN /1H0,8HVARIABLE)
 37   FORMAT(1H0,2X,I3,2X,(10F10.3))
      END                                                               265
C
      SUBROUTINE DIVERS(JO,N1,N2,N3,MAXK,TOT,TOTM,DIV)
C     DO SUBROUTINE DIVERS TO CALCULATE THE DIVERSITY OF THE SET OF
CC    INDIVIDUALS
C                                                                       270
      DIMENSION LOGFAC(20),LGFACN(20),LGFACM(20),A(20),B(20),LFAC(20,10)
      COMMON X(4000),SOUT(20,10),S(20),RJ(20)
      REAL LOGFAC,LGFACN,LGFACM,LFAC
      IF(N3.EQ.0) GO TO 10
      SMAX=0.                                                           275
```

```
      DO 1 I=2,N2
      S(I)=0.
      LOGFAC(I)=0.
      Z=FLOAT(N1)
      N4=N1                                                          280
      IF(JO.EQ.0) GO TO 2
      Z=Z-1
      N4=N4-1
      IF(Z.LE.0) GO TO 3
    2 RF=1./Z                                                        285
    3 CONTINUE
      DO 4 K=1,N4
    4 LOGFAC(I)=LOGFAC(I)+ALOG10(FLOAT(K))
      XN=FLOAT(N4)/RJ(I)
      N=XN                                                           290
      R=(XN-FLOAT(N))*RJ(I)
      LGFACN(I)=0.
      DO 5 K=1,N
      XK=FLOAT(K)
      AXK=ALOG10(XK)                                                 295
    5 LGFACN(I)=LGFACN(I)+AXK
      A(I)=(RJ(I)-R)*LGFACN(I)
      M=N
      M=M+1
      LGFACM(I)=0.                                                   300
      DO 6 L=1,M
      XL=FLOAT(L)
      AXL=ALOG10(XL)
    6 LGFACM(I)=LGFACM(I)+AXL
      B(I)=R*LGFACM(I)                                               305
      SMAX=SMAX+(LOGFAC(I)-A(I)-B(I))/Z
      TOTSM=SMAX*FLOAT(N4)
      TOTM=TOTSM
C
      SUM=0.                                                         310
      DO 7 J=1,MAXK
      LFAC(I,J)=0.
      K=SOUT(I,J)
      DO 8 L=1,K
      XL=FLOAT(L)                                                    315
    8 LFAC(I,J)=LFAC(I,J)+ALOG10(XL)
    7 SUM=SUM+LFAC(I,J)
    1 S(I)=(S(I)+LOGFAC(I)-SUM)*RF
      TOT=0.
      DO 9 I=2,N2                                                    320
    9 TOT=TOT+S(I)
      TOT=TOT*FLOAT(N4)
      DIV=TOT/TOTSM
      GO TO 25
C                                                                    325
   10 DO 20 I=2,N2
      S(I)=0.
      DO 20 J=1,MAXK
      IF(SOUT(I,J).LE.0.) GO TO 20
      S(I)=S(I)-SOUT(I,J)*ALOG10(SOUT(I,J))                          330
```

```
      1    1    2    2    3    3    4    4    5    5    6    6    7    7    8
....5....0....5....0....5....0....5....0....5....0....5....0....5....0....5....0

 20   CONTINUE
      SMAX=0.
      DO 21 I=2,N2
 21   SMAX=SMAX+ALOG10(RJ(I))
      TOTSM=SMAX*FLOAT(N1)                                                  335
      TOTM=TOTSM
C
      TOT=0.
      DO 22 I=2,N2
 22   TOT=TOT+S(I)                                                          340
      TOT=TOT*FLOAT(N1)
      DIV=TOT/TOTM
 25   CONTINUE
      END
C                                                                           345
      SUBROUTINE SIMIL(I,N1,N2,S)
C     DO SUBROUTINE SIMIL TO CALCULATE THE SIMILARITY BETWEEN EACH
C     INDIVIDUAL AND THE MOST TYPICAL ONE
C
      COMMON IN(200,20),X(220),R(20),SIM(200)                               350
      DO 1 J=1,N1
 1    SIM(J)=0.
      DO 2 K=1,N1
      DO 2 J=2,N2
      IA=IN(I,J)                                                            355
      IB=IN(K,J)
      IF(IA.NE.IB)GO TO 2
      SIM(K)=SIM(K)+R(J)
 2    CONTINUE
      END                                                                   360

      1    1    2    2    3    3    4    4    5    5    6    6    7    7    8
....5....0....5....0....5....0....5....0....5....0....5....0....5....0....5....0
```

Appendix D. DATA on 198 INDIVIDUALS

1	2	3	6	5	8	4	1	8	2	7	7	6
2	7	7	1	2	5	7	4	7	5	6	5	3
3	3	4	4	2	6	4	3	6	5	4	2	6
4	3	5	5	5	6	5	3	5	4	3	4	5
5	3	5	5	5	4	6	2	7	4	7	3	4
6	2	6	6	4	4	5	5	3	7	4	2	7
7	2	3	5	5	4	5	5	3	6	5	3	6
8	4	5	5	4	6	4	4	5	3	4	3	6
9	4	3	8	6	3	4	4	4	5	5	6	6
10	4	4	4	4	5	4	3	2	8	5	6	6
11	7	5	3	8	4	3	5	2	4	4	4	5
12	5	4	4	5	4	5	8	2	5	5	4	5
13	8	6	6	5	5	3	5	6	4	6	5	7
14	4	7	5	6	4	6	3	5	5	4	4	6
15	4	4	5	4	4	3	1	4	6	6	6	4
16	5	6	4	6	4	6	8	4	3	3	2	6
17	4	4	2	3	6	5	6	6	6	6	5	3
18	6	4	5	6	3	3	6	3	6	7	7	5
19	4	3	6	7	4	5	5	2	5	3	5	6
20	2	1	6	3	4	5	6	4	6	7	6	4
21	1	5	3	4	5	3	6	4	4	6	6	5
22	5	5	5	5	7	1	3	7	6	5	5	2
23	7	3	3	8	5	3	3	3	8	6	4	3
24	4	4	5	5	4	4	6	2	4	5	4	2
25	3	5	3	4	4	4	5	5	3	3	2	5
26	4	5	2	6	3	4	4	5	5	5	4	6
27	3	6	5	1	3	3	6	5	2	5	5	5
28	1	3	8	3	7	4	4	1	3	5	1	5
29	4	4	5	5	5	5	3	5	6	3	3	6
30	6	4	3	5	5	4	4	5	6	3	4	6
31	5	3	5	5	4	1	1	5	3	4	8	8
32	4	6	7	3	5	5	4	6	4	6	4	6
33	7	5	2	5	5	5	5	7	1	4	7	5

34	3	6	3	6	6	2	1	6	4	4	5	5
35	6	4	4	5	4	3	3	6	2	6	7	2
36	2	3	5	4	3	4	4	4	6	6	6	5
37	6	5	5	4	6	4	5	7	4	2	5	5
38	4	5	3	7	4	4	3	5	5	5	6	6
39	2	4	4	5	3	6	5	4	4	6	5	7
40	3	3	4	6	4	3	4	6	4	4	5	5
41	4	6	4	5	5	2	3	5	5	6	7	6
42	4	6	4	6	3	4	4	3	4	5	4	6
43	5	4	5	3	5	4	5	2	5	4	4	5
44	2	4	5	4	5	5	5	4	5	4	3	5
45	4	3	4	5	4	6	3	2	3	3	1	8
46	5	3	7	5	7	6	5	5	6	6	7	5
47	2	5	5	4	5	3	4	2	4	4	7	4
48	5	5	5	4	5	4	5	6	3	7	6	5
49	5	6	5	5	4	5	6	1	6	4	6	7
50	3	8	3	5	3	3	4	5	5	7	7	2
51	6	5	3	6	5	5	4	4	5	5	4	6
52	4	5	6	6	6	1	2	5	6	2	4	1
53	2	5	5	5	3	4	3	3	6	5	4	6
54	5	4	2	5	4	4	5	6	4	5	5	4
55	4	3	2	6	3	5	6	4	5	5	7	6
56	5	4	4	6	4	3	5	3	6	6	6	4
57	3	3	4	6	4	3	3	6	5	6	5	3
58	3	6	5	4	4	4	5	6	5	6	3	5
59	7	4	4	8	6	5	6	6	5	5	6	5
60	6	7	2	5	4	5	6	5	3	6	5	6
61	6	2	5	3	8	7	3	4	4	4	4	6
62	5	4	1	4	4	5	2	7	6	5	4	4
63	6	6	6	6	3	5	5	3	5	4	6	6
64	5	3	6	3	8	5	6	4	8	3	1	1
65	8	1	6	7	4	4	4	5	5	3	5	4
66	4	5	5	5	6	2	3	7	7	4	4	6

67	5	5	8	5	2	8	1	4	1	4	4	4
68	2	4	5	4	5	4	3	4	6	4	1	5
69	5	4	5	3	7	4	6	5	3	6	4	6
70	3	6	3	6	5	4	5	6	5	6	3	5
71	3	6	6	7	5	2	3	2	5	3	4	8
72	5	6	3	4	5	5	2	7	8	4	3	7
73	8	8	4	8	3	4	3	5	5	7	8	5
74	5	2	5	4	4	7	3	8	3	6	2	8
75	6	4	4	6	5	4	3	5	3	6	3	5
76	2	8	7	4	7	4	5	4	3	6	8	5
77	4	3	5	3	6	4	3	2	8	6	1	5
78	6	4	3	5	3	7	3	3	6	6	5	3
79	5	6	3	5	5	6	5	4	1	6	3	6
80	5	4	8	2	8	3	5	5	1	8	4	7
81	1	4	8	4	4	4	5	2	6	5	5	5
82	4	3	5	3	6	6	3	5	6	4	5	2
83	6	4	4	5	3	4	3	5	1	6	8	3
84	3	7	4	5	8	4	5	7	5	7	2	5
85	6	5	5	7	4	8	1	6	8	2	2	1
86	7	2	6	8	5	2	5	7	2	5	5	2
87	5	6	4	5	1	8	4	5	3	5	1	6
88	6	4	5	5	3	6	5	6	6	4	6	8
89	5	5	4	6	6	6	1	4	5	3	3	8
90	3	5	7	5	4	3	7	2	4	4	5	6
91	5	5	5	4	6	5	5	3	5	4	7	4
92	4	5	4	5	5	6	4	4	5	4	3	3
93	4	5	4	7	4	4	2	4	5	5	1	6
94	1	3	5	7	2	4	4	4	4	7	5	8
95	2	5	3	3	4	3	3	8	8	3	5	4
96	3	6	4	5	4	2	4	1	7	4	5	3
97	3	8	6	5	2	7	4	4	5	5	4	4
98	3	3	5	5	6	4	3	5	7	3	3	5
99	3	5	3	5	2	6	2	5	6	8	3	1

100	7	3	7	5	1	2	4	8	5	6	7	8
101	5	5	3	5	5	6	4	6	5	6	5	3
102	6	3	5	2	8	4	8	1	2	8	6	5
103	4	6	5	2	1	3	5	4	5	3	3	8
104	4	4	7	3	6	4	2	4	3	6	3	5
105	4	5	2	3	5	4	2	3	5	4	3	7
106	6	1	3	6	3	8	3	4	6	8	5	6
107	2	4	3	2	5	5	3	2	6	4	3	5
108	5	2	4	5	3	3	3	4	5	5	5	6
109	5	6	4	4	3	6	2	1	2	6	6	2
110	2	5	3	4	8	8	8	5	6	3	4	6
111	6	5	5	5	7	4	4	4	3	3	7	8
112	7	1	6	5	5	8	6	2	5	6	5	4
113	6	6	4	6	5	3	5	3	3	4	6	6
114	5	6	6	4	5	6	6	5	3	5	4	7
115	7	5	5	1	3	5	5	4	7	4	6	4
116	7	4	5	4	5	5	4	5	3	4	5	5
117	6	5	4	5	4	3	8	4	6	6	2	4
118	6	5	4	3	3	5	2	1	1	7	4	1
119	5	5	4	2	3	4	5	6	4	4	5	3
120	2	2	5	3	3	7	5	3	3	6	4	4
121	6	4	5	3	4	5	5	6	3	2	7	4
122	3	6	4	3	6	7	6	5	4	3	4	5
123	8	6	1	2	6	5	4	1	6	4	5	5
124	5	6	2	1	3	2	5	3	3	2	4	6
125	7	6	1	3	5	2	4	1	6	3	6	5
126	4	3	7	2	4	4	6	3	5	4	7	5
127	4	6	4	2	4	4	4	2	6	6	5	5
128	5	6	6	3	3	6	4	3	5	5	4	3
129	6	6	7	1	1	2	3	7	5	5	5	3
130	5	6	5	2	5	7	5	2	5	4	4	4
131	7	7	5	2	5	6	4	4	3	4	5	3
132	5	6	7	3	3	5	3	5	4	1	2	5

133	3	5	3	2	4	2	5	6	5	6	4	4
134	6	5	4	1	8	3	5	3	6	6	4	2
135	6	5	4	3	5	2	3	5	3	4	4	3
136	6	8	5	1	2	4	3	7	3	1	6	3
137	4	7	5	7	6	6	6	5	4	4	4	2
138	6	4	4	6	3	6	6	6	6	5	4	2
139	4	8	3	4	4	5	6	3	2	5	5	4
140	6	2	1	2	2	3	7	5	2	3	1	5
141	4	5	4	3	6	4	5	4	3	5	5	5
142	5	6	4	4	4	3	3	4	4	5	3	5
143	6	8	6	1	1	5	3	7	4	1	6	3
144	4	5	4	1	2	5	7	6	6	4	5	6
145	6	6	6	1	4	6	6	6	5	5	4	7
146	6	2	6	1	4	3	8	7	8	6	5	6
147	5	6	6	3	4	5	4	4	4	4	4	5
148	5	6	4	3	3	5	6	6	2	5	3	2
149	4	3	5	4	4	5	5	6	2	3	5	4
150	4	5	5	5	3	5	7	5	4	5	5	5
151	4	5	6	4	5	4	5	6	4	6	4	4
152	4	6	4	4	3	5	6	5	5	3	4	3
153	5	3	5	5	4	5	4	6	1	5	4	4
154	6	5	4	5	7	6	4	6	4	5	5	1
155	5	5	4	8	7	7	8	5	5	4	4	6
156	4	2	6	3	7	3	1	4	2	5	1	1
157	4	4	3	5	3	7	5	5	6	4	7	6
158	5	3	2	5	4	5	4	6	6	5	5	4
159	3	4	5	6	4	4	6	3	4	6	5	2
160	4	4	5	6	5	4	4	5	6	5	6	2
161	3	3	5	4	4	5	4	6	4	7	5	6
162	2	4	6	6	1	4	4	2	6	4	3	1
163	3	4	6	5	4	5	5	6	4	3	6	3
164	2	8	6	7	3	5	7	6	3	5	3	2
165	4	4	6	5	6	5	6	6	4	1	6	4

166	4	5	6	6	5	3	7	6	5	3	6	5
167	3	7	6	7	4	6	8	7	3	2	3	7
168	4	1	7	6	2	2	8	4	5	4	3	5
169	7	8	8	4	6	5	5	3	6	3	8	4
170	5	4	3	6	4	5	6	5	4	3	7	3
171	3	5	7	5	3	4	6	3	4	3	2	4
172	6	7	4	4	3	7	6	3	6	6	5	4
173	5	2	7	6	4	5	5	4	3	3	3	3
174	3	3	5	4	3	3	4	5	6	2	6	3
175	5	4	5	7	3	4	5	4	4	4	4	3
176	1	7	1	4	8	2	8	3	3	1	7	3
177	1	1	1	2	7	8	1	4	1	2	7	6
178	4	4	2	6	3	3	4	5	5	6	4	3
179	6	6	4	7	5	2	6	5	5	3	4	4
180	5	3	6	6	6	4	8	4	5	2	6	4
181	3	3	8	6	5	8	6	4	5	1	6	2
182	7	1	2	8	4	3	3	1	2	3	3	3
183	5	3	7	5	3	2	2	3	2	4	2	4
184	6	2	2	3	1	3	8	8	4	8	2	3
185	3	4	4	6	4	4	8	4	2	5	4	3
186	8	5	2	5	6	4	6	4	7	6	2	2
187	8	1	4	5	5	4	7	6	8	1	6	3
188	1	1	4	5	2	1	4	4	1	1	4	5
189	5	4	3	4	3	4	7	5	3	6	4	4
190	2	7	3	7	5	6	5	6	4	3	7	1
191	8	5	4	7	6	5	4	7	2	3	1	1
192	7	5	3	6	4	4	7	5	2	1	3	7
193	6	2	4	3	5	5	5	5	4	5	6	3
194	4	5	6	4	3	8	5	5	5	2	5	3
195	3	4	4	7	5	6	6	5	5	3	5	4
196	6	4	5	6	4	5	3	5	3	3	6	5
197	4	4	4	4	4	1	6	4	3	7	3	2
198	7	3	6	5	6	5	4	5	4	3	6	6

BIBLIOGRAPHY.

Abramson, N. (1963) *Information Theory and Coding*. New York : McGraw Hill.

Alexander, I.C. (1972) Multivariate Techniques in Landuse Studies; The case of information analysis, *Regional Studies*, (6), p. 93-103.

Ash, R. (1965) *Information Theory*. New York : Wiley.

Austin, M.P. (1972) Models and Analysis of Descriptive Vegetation Data, in Jeffers, J.N.R. (Ed.), *Mathematical Models in Ecology*, p. 61-87, Blackwell Scientific Publications, Oxford.

Baer, R.M. (1953) Some General Remarks on Information Theory and Entropy, in Quastler, H.,(Ed.), *Information Theory and Biology*, Urbanna : University of Illinois Press.

Basharin, G.P. (1959) On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables. *Theory Probab. Applic.*, (4) p. 333-336.

Barnard, G.A.(Ed.) (1962) *The Foundations of Statistical Inference*. London : Methuen.

Batty, M. (1972) Entropy and Spatial Geometry, *Area*, (4), p. 230-36.

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

Behara, M. (1974) Entropy and Utility, in Menges, G., (Ed.), *Information, Inference and Decision*. p. 145-154, Holland; D.Reidel Publishing Company.

Behara, M. and Nath, P. (1973) Additive and Non-additive Entropies of Finite Measurable Partitions, in Behara, M., Krickelberg, K., Wolfowttz, J., *Probability and Information Theory II*, p.102-138, Berlin : Springer-Verlag.

Belis, M. and Guiasu, S. (1968) A Quantitative-Qualitative Measure of Information in Cybernetic Systems, *I.E.E.E. Transactions on Information Theory*, (14), p. 593-94.

Bell, D.A. (1967) *Information Theory and its Engineering Applications*. Fourth Edition, London : Sir Isaac Pitnam and Son Ltd.

Berry, B.J.L. (1961) A Method For Deriving Multi-Factor Uniform Regions, *Przegled Geograficzny*, (33), p.263-282.

Berry, B.J.L. (1965) The Mathematics of Economic Regionalization - *Proceedings of the 4th General Meeting of the Commission on Methods of Economic Regionalization of the International Geographical Union*, Prague, 1967.

Bottomley, J. (1971) Some Statistical Problems Arising from the Use of the Information Statistic in Numerical Classification, *Journal of Ecology*, (59), p. 339-42.

Boulton, D.M. and Wallace, C.S. (1969) The Information Content of a Multistate Distribution, *Journal of Theoretical Biology*, (23), p. 269-78.

Boulton, D.M. and Wallace, C.S. (1970) A Program for Numerical Classifcation, *Computer Journal*, (13), p. 63-69.

Bowman, K.L, Hutcheson, K., Odum, E.P., Shenton, L.R. (1971) Comments on the Distribution of Indices of Diversity, in Patil, G.P., Pielou, E.C,, and Waters, W.E., (Eds.) *Proceedings of the Symposium on Statistical Ecology*, vol. III, p. 315-336, Pennsylvania.

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

Brillouin, L. (1962) Observation, Information and Imagination, in Dockx, S., Bernays, P., (Eds.), *Information and Prediction in Science*, p. 1-14, New York : Academic Press.

Brillouin, L. (1964) *Scientific Uncertainty and Information*. New York: Academic Press.

Brummell, A., Harman, R.J. (1974) Behavioral Geography and Multidimensional scaling. Dept. of Geography, McMaster University, Hamilton, Discussion Paper No. 1.

Bryant, C.R. (1974) An Approach to the Problem of Urbanization and Structural change in Agriculture : A case study from the Paris Region, *Geografiska Annaler*, (56)B, p. 1-27.

Burr, E.J. (1968) Cluster Sorting with Mixed Character Types, I. Standardization of Character Values, *Australian Computer Journal* , (1), p. 97-99.

Burr, E.J. (1970) Cluster Sorting with Mixed Character Types II, Fusion Strategies , *Australian Computer Journal*, (2), p. 98-103.

Chapman, G.P. (1970) The Application of Information Theory to the Analysis of Population Distributions in Space, *Economic Geography* , (46) , p. 317-331.

Chapman, G.P. (1973) The Spatial Organization of the Population of the United States and England and Wales, *Economic Geography*, (49), p. 325-343.

Coombs, C.H. (1964) *A Theory of Data*. New York : Wiley.

Cullen, H.F. (1968) *Introduction To General Topology*. Boston : Heath and Company.

Curry, L. (1972) A Spatial Analysis of Gravity Flows, *Regional Studies*, (6), p. 131-147.

Dale, M.B. (1971) Information Analysis of Quantitative Data, in Patil, G.P., Pielou, E.C., Waters, W.E., (Eds.), *Statistical Ecology*, vol. 3, p. 133-148, Pennsylvania State University.

Dale, M.G., MacNaughton-Smith, P., Williams, W.T, Lance, G.N. (1970) Numerical Classification of Sequences, *Australian Computer Journal*, (2), p. 9-13.

Dale, M.G., Lance, G.N., Albrecht, L. (1971). Extensions of Information Analysis, *Australian Computer Journal*, (3), p. 29-34.

Dale, M.B., Anderson, J. (1972) Qualitative Information Analysis, *Journal of Ecology*, (60), p. 639-653.

DeGroot, M.H. (1962) Uncertainty, Information, and Sequential Experiments. *Annals of Mathematical Statistics*, (33), p.404-419.

Dutta, H. (1968) A Hundred Years of Entropy, *Physics Today*, (21), p.75-79.

Edwards, A.W.F. and Cavalli-Sforza, L.L. (1965) A Method For Cluster Analysis, *Biometrics*, (21), p. 362-375.

Edye, L.A., Williams, W.T., Pritchard, A.J. (1970) Numerical Analyses of Variation Pattern in Australian Introductions of Glycine Wightii - *Australian Journal of Agricultural Research*, (21), p. 57-69.

Feller, W. (1968) *An Introduction To Probability Theory and its Applications*. vol. (1), 3rd edition. New York : Wiley.

Field, J.G. (1969) The Use of the Information Statistic In the Numerical Classification of Heterogenous Systems. *Journal of Ecology*, (57), p. 565-569.

Garner, W.R. (1962) *Uncertainty and Structure as Psychological Concepts*. New York.

Garner, W.R., McGill, W.J. (1956) The Relation between Information and Variance Analyses, *Psychometrika*, (21), p. 219-228.

Garrison, C.B. and Paulson, A.S. (1973) An Entropy Measure of the Geographic Concentration of Economic Activity, *Economic Geography* (49), p. 319-324.

Georgescu-Roegen, N. (1971) *The Entropy Law and the Economic Process.* Harvard University Press.

Goldman, S. (1953) *Information Theory.* New York : Prentice Hall.

Golledge, R.G. and Rushton G. (1972) *Multidimensional Scaling : Review and Geographical Applications*, Washington D.C. Association of American Geographers, Commision on Colledge Geography, Technical Report, No. 10.

Good, I.J. (1950) *Probability and the Weighting of Evidence.* London : Charles Griffin and Co. Ltd.

Good, I.J. (1953) The Population Frequencies of Species and the Estimation of Population Parameters, *Biometrika*, (40), p. 237-264.

Good, I.J. (1956) Some Terminology and Notation In Information Theory, *Proc. I.E.E., Part C*, vol. (103) p. 200-204.

Good, I.J. (1959) Kinds of Probability, *Science*, (129), p. 443-447.

Good, I,J. (1963) Maximum Entropy For Hypothesis Formulation, especially for Multidimensional Contingency Tables, *Annals of Mathematical Statistics*, (34) p. 911-934.

Good, I.J. (1965) *The Estimation of Probabilities : An Essay on Modern Bayesian Methods.* Cambridge, Mass : The M.I.T. Press.

Gould, P. (1965) On Mental Maps, Ann Arbor, Michigan Inter-University Community of Mathematical Geographers.

Gould, P. and White, R. (1974) *Mental Maps.* Harmondsworth, England : Penguin.

Gower, J.C. (1967) A Comparison of Some Methods of Cluster Analysis, *Biometrics*, (23), p. 623-637.

Grigg, D.B. (1965) The Logic of Regional Systems, *Annals of the Association of American Geographers*, (55), p. 465-491.

Grigg, D,B. (1967) Regions, Models and Classes in Chorley, R.J., Haggett, P. (Eds.), *Models In Geography*, p. 461-507, London : Methuen.

Gurevich, B.L. (1969)a. Measure of Feature Based and Areal Differentiation and Their Use in City Services, *Soviet Geography: Review and Translations*, (10), p. 380-386.

Gurevich, B.L. (1969)b. Geographical Differentiation and Its Measures in a Discrete System, *Soviet Geography : Reviews and Translations*, (40), p. 387-413.

Harvey, D. (1969) *Explanation In Geography*. London : Arnold.

Hutcheson, K. (1970) A Test For Comparing Diversities based on the Shannon Formula, *Journal of Theoretical Biology*, (29), p. 151-154.

Hyvarinen, L.P. (1970) *Information Theory For Systems Engineers*. Berlin : Springer-Verlag.

Ingarden, R.S. and Urbanik, R. (1962) Information without Probability, *Colloq. Math*, (9), p. 131-150.

Ingels, F.M. (1971) *Information and Coding Theory*. London : Intext Educational Publisher.

Jauch, J.M., and Baron, J.G. (1972) Entropy, Information and Szilard's Paradox. *Helvetica Physica Acta*, (45), p. 220-232.

Jaynes, E.T. (1957) Information Theory and Statistical Mechanics *Physical Review*, (106), p. 620-630, and (108), p. 171-190.

Jaynes, E.T. (1963)a Information Theory and Statistical Mechanics in Ford, K.W., (Eds.), *Statistical Physics*, vol. 3, p. 181-218.

Jaynes, E.T. (1963)b. New Engineering Applications of Information Theory, in Bogdonoff, J.L., Kozin, F., (Eds.), *Proceedings of First Symposium on Engineering Applications of Random Function Theory and Probability*, p. 163-203, New York : Wiley.

Jaynes, E.T. (1965) Gibbs vs Boltzmann Entropies, *American Journal of Physics*, (33), p. 391-398.

Jaynes, E.T. (1968) Prior Probabilities. *I.E.E.E. Transactions on Systems Science and Cybernetics*, SSC-4, p. 227-241.

Jeffreys, H. (1939) *Theory of Probability*. First edition. Oxford : Clarendon Press.

Johnson, S.C. (1967) Hierarchical Clustering Scheme, *Psychometrika*, (32), p. 241-255.

Johnson, H.A. (1970) Information Theory in Biology after 18 years. *Science*, (1968), p. 1545-50.

Johnston, R.J. (1968) Choice In Classification - The subjectivity of objective methods - *Annals of the Association American Geographers*, (58), p. 575-589.

Johnston, R.J. (1970) Grouping and Regionalization : Some Methodological and Technical Observations, *Economic Geography*, (46), p. 292-305.

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

Klein, M.J. (1953) Order, Organization and Entropy, *British Journal of the Philosophy of Science*, (4), p. 158-160.

Kotz, S. (1966) *Recent results in Information Theory*. London : Methuen.

Kruskal, J.B. (1964)a. Multidimensional Scaling by Optimizing Goodness of Fit, *Psychometrika*, (29), p. 1-27.

Krushal, J.B. (1964)b. Non-metric Multidimensional Scaling, A Numerical Method, *Psychometrika*, (29), p. 115-129.

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

Kullback, S., Kuppermann, M., Ku, H.H. (1962) Tests For Contingency Tables and Markov Chains, *Technometrics*, (4), p. 573-608.

Kuppermann, M. (1959) A Rapid Significance Test For Contingency Tables, *Biometrics*, (15), p. 625-628.

Lambert, J.M. and Williams, W.T. (1966) Comparison of Information Analysis and Association Analysis, *Journal of Ecology*, (54), p. 635-665.

Lambert, J.M. (1972) Theoretical Models For Large-scale vegetation Survey, in Jeffers, J.N. R., (Ed.), *Mathematical Models in Ecology*, p. 87-111, Oxford; Blackwell Scientific Publications.

Lance, G.N. and Williams, W.T. (1966) Computer Programs For Hierarchical Polythetic Classification, *Computer Journal*, (9), p. 60-64.

Lance, G.N. and Williams, W.T. (1967)a. Note On the Classification of Multilevel Data, *Computer Journal*, (9), p. 381-82.

Lance, G.N. and Williams, W.T. (1967)b. A General Theory of Classificatory Sorting Strategies, I, Hierarchical Systems, *Computer Journal*, (9), p. 373-380.

Lance, G.N. and Williams, W.T. (1967)c. Mixed Data Classificatory Programs, I, Agglomerative Systems, *Australian Computer Journal*, (1) p. 15-21.

Lance, G N. and Williams, W.T. (1968)a. Mixed Data Classificatory Programs, II, Divisive Systems, *Australian Computer Journal*, (1) p. 82-86.

Lance, G.N., Williams, W.T., Milne, P.W. (1968)b. Mixed Data Classificatory Programs III, Diagnostic Systems, *Australian Computer Journal*, (1), p. 178-182.

Lance, G.N., and Williams, W.T. (1968)c. A General Theory of Classificatory Sorting Strategies, II, Clustering Systems, *Computer Journal*, (10), p. 271-277.

Lance, G.N. and Williams, W.T. (1968)d. Note on a New Information-Statistic Classificatory Program, *Computer Journal*, (11), p. 195.

Lance, G.N. and Williams, W.T. (1971) A Note On a New Divisive Classification, Program For Mixed Data, *Computer Journal*, (14), p. 154-155.

Landsberg, P.T. (1961) *Entropy and the Unity of Knowledge*. Cardiff : University of Wales Press.

Lee, R. (1974) Entropy Models In Spatial Analysis. University of Toronto, Department of Geography, Discussion Paper No. 15.

Linfoot, E.H. (1957) An Informational Measure of Correlation, *Information and Control*, (1), p. 85-89.

Lloyd, M., Zar, J.G., J.R. (1968) On the Calculation of Information Theoretical Measures of Diversity. *The American Midland Naturalist*, (79), p. 257-272.

MacDonald, D.K.C. (1952) Information Theory and Its Application To Taxonomy, *Journal of Applied Physics*, (23), p. 529-31.

MacKay, D.M. (1950) The Nomenclature of Information Theory, in Forester, H. Von., (Ed.), *Proceedings of the 8th Conference on Cybernetics*, p. 222-235, New York.

MacKay, D.M. (1956) The Place of "Meaning" in the Theory of Information in Cherry, C., (Ed.), *Information Theory*, p. 215-225, London.

MacKay, D.M. (1965) Information and Prediction In Human Sciences, in Dockx, S., Bernays, P., (Eds.), *Information and Prediction in Science*, p. 255-269, New York : Academic Press.

MacNaughton-Smith, P. (1965) *Some Statistical and other Numerical Techniques For Classifying Individuals*. London : H.M.S.O.

Marschak, J. (1974) Entropy, Economics, Physics, Western Management Science Institute, University of California, Los Angeles. Working Paper 221.

Marchand, B. (1972) Information Theory and Geography, *Geographical Analysis*, (4), p. 234-258.

Marchand, B. (1975) On the Information Content of Regional Maps: The Concept of Geographical Redundancy, *Economic Geography*, (51), p. 117-128.

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

Medvedkov, Y. (1967)a. The Regular Component In Settlement Patterns as shown on a map, *Soviet Geography : Review and Translation*, (8), p. 150-168.

Medvedkov, Y. (1967)b. The Concept of Entropy In Settlement Pattern Analysis, *Papers of the Regional Science Association*, (18), p. 165-168.

Medvedkov, Y. (1970) Entropy, An Assesment of Potentialities in Geography, *Economic Geography*, (46), p. 307-316.

Meyer, S.L. (1975) *Data Analysis For Scientists and Engineers*. New York : Wiley.

Miller, G.A. (1956) The Magical Number Seven, Plus or Minus Two, *Psychological Review*, (63), p. 81-97.

Mogridge, M.J.H. (1972) The Use and Misuse of Entropy In Urban and Regional Modelling of Economic and Spatial Systems. London : Centre for Environmental Studies, Working Paper No. 80.

Nauta, D. (1972) *The Meaning of Infomation*. The Hague : Mouton.

Novitskii, P.V. (1966) The Concept of an Error Entropy Value, *Measurement Techniques*, p.872-875.

Nunnally, J.C. (1967) *Psychometric Theory*. New York : McGraw Hill.

Nutenko, L.Y. (1970) An Information Theory Approach to the Partitioning of an area, *Soviet Geography : Review and Translations*, (11), p. 540-544.

Orloci, L. (1968) Information Analysis in Phyto-sociology : Partition, Classification and Prediction, *Journal of Theoretical Biology*, (20), p. 271-284.

Orloci, L. (1969) Information Analysis of Structure In Biological Collections, *Nature*, (223), p. 483-484.

Orloci, L. (1970)a. Automatic Classification of Plants based on Information Content, *Canadian Journal of Botany*, (48), p. 793-802.

Orloci, L. (1970)b. Analysis of Vegetation Samples based on the use of Information, *Journal of Theoretical Biology*, (29), p. 173-189.

Orloci, L. (1971)a. Information Theory Techniques For Classifying Plant Communities, in Patil, G.P., Pielou, E.C., Waters, W.E., (Eds.), *Statistical Ecology*, vol. III, p. 259-271, Pennsylvania State University.

Orloci, L. (1971)b. An Information Theory Model For Pattern Analysis, *Journal of Ecology*, (59), p. 343-349.

Pearson, E.S. and Hartley, H.O. (1954) *Biometrika Tables For Statisticians*, Cambridge University Press.

Peet, R.K. (1974) The Measurement of Species Diversity, *Annual Review of Ecology and Systematics*, (5) p. 285-307.

Pielou, E.C. (1966)a. Shannon's Formula as a Measure of Specific Diversity - its use and misuse, *American Naturalist*, (100), p. 463-465.

Pielou, E.C. (1966)b. Species Diversity and Pattern Diversity in the Study of Ecological Succession, *Journal of Theoretical Biology*, (10), p. 370-383.

Pielou, E.C. (1966)c. The Measurement of Diversity In Different types of Biological Collections, *Journal of Theoretical Biology*, (13), p. 131-144.

Pielou, E.C. (1967) The use of Information Theory in the study of the diversity of biological populations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4, p. 163-177.

Pielou, E.C. (1969) *An Introduction to Mathematical Ecology*. New York : Wiley.

Quastler, H. (1953) *On the Use of Information Theory in Biology*. Urbana.

Quastler, H. (1964) *The Emergence of Biological Organization*. New Haven : Yale University Press.

Rajski, C. (1961)a. A Metric Space of Discrete Probability Distributions, *Information and Control*, (4), p. 371-377.

Rajski, C. (1961)b. Entropy and Metric Spaces, in Cherry, C., (Ed.), *Information Theory, 4th London Symposium*, p.41-46, London.

Rapoport, A. (1954) What is Information?, *Synthese*, (9), p. 157-173.

Rapoport, A. (1956) The Promise and Pitfalls of Information Theory, *Behavioural Science*, (1), p. 303-309.

Renyi, A. (1967) Statistics and Information Theory, *Studia Scientifiarum Mathematicarum Hungarica*, (2), p. 249-256.

Rescigno, A. and Maccacaro, G.A. (1960) The Information Content of Biological Classifications, in Cherry, C., (Ed.), *Information Theory*, p. 437-447, London.

Rothstein, J. (1951) Information, Measurement and Quantum Mechanics, *Science*, (114), p. 171-175.

Rothstein, J. (1952) Information and Thermodynamics, *Physical Review*, (85), p. 135.

Saarinen,T.F. (1966) The Perception of Drought Hazard on the Great Plains, Chicago : University of Chicago, Dept. of Geography, Research Paper No. 106.

Semple, R.K., and Golledge, R.G. (1970) An Analysis of Entropy Changes in a Settlement Pattern over time. *Economic Geography*, (46), p. 157-160.

Semple, R.K., and Wang, L.H. (1970) A Geographical Analysis of Redundancy in Interurban Transportation Links. University of Toronto, Department of Geography, Discussion Paper, No.5.

Semple, R.K. and Griffin, J.M. (1971) An Information Analysis of Trends In Urban Growth Inequality in Canada, Ohio State University, Department of Geography, Discussion Paper No. 19.

Semple, R.K., Youngmann, C.E., Zeller, R.E. (1972) Economic Regionalization and Information Theory : An Ohio Example. Ohio State University. Department of Geography, Discussion Paper No. 28.

Semple, R.K. and Gauthier, H.L. (1972) Spatial-Temporal Trends In Income Inequalities in Brazil. *Geographical Analysis*, (4), p. 169-180.

Semple, R.K. (1973) Recent Trends In the Spatial Concentration of Corporate Headquarters. *Economic Geography*, (49), p. 309-18.

Shannon, C. and Weaver, W. (1949) *The Mathematical Theory of Communication*, University of Illinois Press.

Shepard, R.N. (1962) The Analysis of Proximities; Multidimensional Scaling with an unknown distance function.
I *Psychometrika*, (27), p. 125-140.
II *Psychometrika*, (27), p. 219-246.

Shepard, R.N. (1964) Attention and Metric Structure of the Stimulus Space. *Journal of Mathematical Psychology*, (1), p. 54-87.

Sheppard, E.S. (1975) Entropy In Geography ; An Information Theoretic Approach to Bayesian Inference and Spatial Analysis, University of Toronto, Department of Geography, Discussion Paper 18.

Simmons, P.J. (1974) *Choice and Demand*, London : Macmillan.

Simon, H.A. (1957) *Models of Man*. New York : Wiley.

Skagerstam, B.S. (1975) On the Notions of Entropy and Information, *Journal of Statistical Physics*, (12), p. 449-462.

Skala, H.J. (1974) Remarks on Semantic Information, in Menges, G., (Ed.), *Information, Inference and Decision*, p. 181-188, Holland; D. Reidel Publishing Company.

Sokal, R.R. and Sneath, P.H.A. (1963) *Numerical Taxonomy*, San Francisco : Freeman.

Somenzi, V. (1962) Entropy, Information and Mind Body Problem in Dockx, S., Bernays, P. (Eds.), *Information and Prediction in Science*, p. 229-235, New York : Academic Press. (1965).

Stevens, S.S. (1951) Mathematics, Measurement and Psychophysics, in Stevens, S.S., (Ed.), *Handbook of Experimental Psychology*, vol. 1, p. 1-76, New York : Wiley.

Ter Haar, D. (1954) *Elements of Statistical Mechanics*. New York : Rinehart and Company.

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

Theil, H. (1972) *Statistical Decomposition Analysis*. Amsterdam : North Holland.

Tillman, F. and Russel B.R. (1961) Information and Entropy, *Synthese*, (13), p. 233-241.

Torgerson, W.S. (1958) *Theory and Methods of Scaling*, New York : Wiley.

Tribus, M. (1961) Information Theory as the basis for Thermostatics and Thermodynamics, *Journal of Applied Machanics*, (28), p. 1-8.

Tribus, M., Shannon, P. and Evans, R. (1966) Why Thermodynamics is a Logical Consequence of Information Theory, *Journal of the American Institute of Chemical Engineers*, (12), p. 244-248.

Trubus, M. (1969) *Rational Descriptions, Decisions and Designs*. Oxford : Pergamon.

Urbanik, K. (1972) On the Concept of Information, in Gani, J., Sarkedi, K., Vincze, I., (Eds.), *Progress in Statistics*, vol. II, p. 863-868, Amsterdam : North Holland.

Van Soest, J.L. (1955) A Contribution of Information Theory to Sociology, *Synthese*, (9), p. 265-273.

Von Mises, R. (1957) *Probability, Statistics and Truth*. New York : Macmillan.

Wallace, W.C. and Boulton, D.M. (1968) An Information Measure For Classification, *Computer Journal*, (11), p. 185-194.

Ward, J.H. (1963) Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, (58), p. 236-244.

Webb, L.J., Tracey, J.G., Williams, W.T., Lance, G.N. (1967)a. Studies in Numerical Analysis of Complex Rain-Forest Communities *Journal of Ecology*, (55), p. 171-191, and p. 525-538.

Webber, M.J. (1975) The "Meaning" of Entropy Maximising Models, Paper read at the Symposium on Mathematical Land Use Theory, McMaster University, Hamilton, April 19, 1975.

Whitla, D.K., (Ed.),(1968) *Handbook of Measurement and Assessment in Behavioral Sciences*, Addison-Wesley.

Wilks S.S. (1935) The Likelihood Test of Independence in Contingency Tables, *Annals of Mathematics and Statistics*, (6), p. 190-196.

Williams, W.T. and Dale, M.B. (1965) Fundamental Problems in Numerical Taxonomy, *Advances in Botanical Research*, (2), p. 35-68.

Williams, W.T. and Lance, G.N. (1965) Logic of Computer Based Intrinsic Classifications, *Nature*, (207), p. 159-161.

Williams, W.T., Lambert, J.M. (1966)a. Multivariate Methods in Plant Ecology : Similarity Analyses and Information Analysis, *Journal of Ecology*, (54), p.427-445.

Williams, W.T., Lance, G,N. (1969) Choice of Strategy in the Analysis of Complex Data, *The Statistician*, (18), p. 31-43.

Williams, W.T., Lance, G.N., Webb, L.J., Tracey, J.G., Dale, M.B. (1969). Studies in the Numerical Analysis of Complex Rain-Forest Communities III; The Analysis of Successional Data, *Journal of Ecology*, (57), p. 515-535.

Williams, W.T., Lance, G.N., Webb, L.J., Tracey, J.G., Connell, J.H. (1969). Studies in the Numerical Analysis of Complex Rain-Forest Communities, IV; A Method for the Elucidation of Small-Scale Forest Pattern, *Journal of Ecology*, (57), p. 635-654.

Williams, W.T. (1971) Principles of Clustering, *Annual Review of Ecology and Systematics*, (2), p. 303-326.

Williams, W.T., Clifford, H.T., Lance, G.N. (1971) Group Size Dependence: a Rationale for Choice between Numerical Classifications, *Computer Journal*, (14), p. 157-162.

Williams, W.T. (1973) Partition of Information, *Australian Journal of Botany*. (20), p. 235-240.

Williams, W.T., Lance, G.N., Webb, L.J., Tracey, J.G. (1973). Studies in the Numerical Analysis of Complex Rain-Forest Communities, VI, Models for the Classification of Quantitative Data, *Journal of Ecology*, (61), p. 47-71.

Wilson, A.G. (1970) *Entropy in Urban and Regional Modelling*. London : Pion.

Wishart, D. (1968) An Algorithm for Hierarchical Classifications, *Biometrics*, (25), p. 165-170.

Woolf, B. (1957) The Log Likelihood Ratio Test. Methods and Tables for Tests of Heterogeneity in Contingency Tables, *Annals of Human Genetics*, (21), p. 397-409.