BNWL-2152 Add 1
NRC-1
Addendum

Similarity Indices I: What Do They Measure?

by
J. W. Johnston

November 1976

Prepared for The Nuclear Regulatory Commission

Battelle
Pacific Northwest Laboratories

BNWL-2152

PACIFIC NORTHWEST LABORATORY
operated by
BATTELLE
for *the*
ENERGY RESEARCH AND DEVELOPMENT ADMINISTRATION
Under Contract EY-76-C-06-7830

Printed in the United States of America
Available **from**
National Technical Information **Service**
U.S. Department of Commerce
5285 Port Royal Road
Springfield, Virginia **22151**
Price: Printed Copy **$____***; **Microfiche** $3.00

'Pages	NTIS Selling Price
001-025	$4.50
026-050	$5.00
051-075	$5.50
076-100	$6.00
101-125	$6.50
126-150	$7.00
151-175	$7.75
176-200	$8.50
201-225	$8.75
226-250	$9.00
251-275	$10.00
276-300	$10.25

ERDA-RL RICHLAND WA

BNWL-2152
NRC-1
Addendum

Similarity Indices I: What Do They Measure?

by
J. W. Johnston

November 1976

Prepared for The Nuclear Regulatory Commission

Battelle
Pacific Northwest Laboratories
Richland, Washington 99352

SUMMARY

The characteristics of 25 similarity indices used in studies of ecological communities were investigated. The type of data structure, to which these indices are frequently applied, was described as consisting of vectors of measurements on attributes (species) observed in a set of samples. A genera1 similarity index was characterized as the result of a two step process defined on a pair of vectors. In the first step an attribute similarity score is obtained for each attribute by comparing the attribute values observed in the pair of vectors. The result is a vector of attribute similarity scores. These are combined in the second step to arrive at the similarity index. The operation in the first step was characterized as a function, g, defined on pairs of attribute values. The second operation was characterized as a function, F, defined on the vector of attribute similarity scores from the first step. Usually, F was a simple sum or weighted sum of the attribute similarity scores.

The functions, g and F, were then specified for 24 of the 25 similarity indices considered (see Table 4). The indices were grouped into 7 classes. The indices in each class have basically the same way of assigning attribute similarities. Each index was defined and calculational formula exemplified by application to 20 sample vectors with 10 species (attributes) each. The data for the example were extracted from a much larger set of actual data consisting of 278 samples and a species list with 203 species. The goal of this part of the paper was to familiarize the reader with the wide range of similarity indices available and to introduce some of the problems involved in their calculation.

Since "similarity" has connotations of closeness in some sense and the data are in the form of vectors, an attempt was made to relate all the indices to a vector space model with N dimensions, N being the total number of attributes (number of species on the species list). In such a

model, "closeness" can be objectified as "distance" and the similarity indices can be characterized by how they distort the vector space.

The basic property of the samples measured by the indices is thus a distance, or complement of a distance, in some vector space defined to fit the operations g and F. All of the indices except Mountford's K_I and Goodall's S_p were characterized in this way. A summary of some of the characteristics of the 25 indices is given in Table 15.

It was pointed out that some of these indices might be useful in descriptive ecology, but not in objectively discriminating between two populations in the statistical sense of discrimination through hypothesis testing. The statistical problems arise from failure to be able to specify the population sampled, and so define meaningful sampling units before the sample is collected, and the lack of proper application of probabilistic models to derive the statistical distributions of the various similarity indices. Consequently only minor reference was made to statistical concepts; the characteristics of the indices were pointed out by merely using algebra.

The major conclusions were as follows.

- Similarity indices should not be used as the test statistic to discriminate between two ecological communities.
- Some of the indices do not use the information contained in the number of individuals per species, and so are insensitive to changes in biomass or total number of individuals.
- Some of the indices ignore negative matches (instances where a species is not observed in either of the two samples being compared). This results in changing the dimensionality of the vector space, making the comparison of indices calculated for different pairs of vectors from the same study questionable.

- Some of the indices make the number of individuals per species relative to the total number of individuals in the sample. Such indices are also insensitive to changes in biomass or total number of individuals.
- Some indices use different divisors of numbers of individuals per species. This destroys the equal interval property required for meaningful comparison of indices from the same study.
- Some indices require statistical standardization, or categorization, of the data to avoid being unduly weighted by species which have very many (say more than 1000) individuals per species.
- Gower's (1971) General Coefficient of Similarity, when negative matches are included, has none of the above defects and can be applied to any level of measurement (discrete, presence absence data through continuous, biomass data). It is recommended as the index of choice for descriptive studies among those considered.

The specific indices which have the above failings are identified (quite cryptically) in Table 15.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

SIMILARITY INDICES: I. WHAT DO THEY MEASURE?

INTRODUCTION

This report investigates 25 similarity indices to specify the property of an ecological community measured by each index. Although the indices discussed have been used or suggested for use in studies attempting to determine whether or not dumping physical or chemical waste has caused a change in a biological community, this study stops short of such a grandiose goal. The goal is to clarify what information each index uses and how that information is summarized in the index. It is expected that from the analysis of the type of data and calculational formula used, it will become evident that similarity indices can only indicate a change in a very limited aspect of the ecological community under investigation.

The question of whether the data adequately characterize an ecological community is sidestepped by operationally defining an ecological community to consist of the attributes (species in these data) a specific sampling method could collect for measurement. It is understood that this kind of definition makes the sampling method the controlling consideration in the definition of an ecological community, thus partitioning the community into the various components available to the particular sampling methods used.

The 25 similarity indices to be discussed were selected to demonstrate the wide variety of measures of the undefined concept "similarity." Apparently, similarity can only be defined operationally as "what a similarity index measures." Since there is great diversity in what similarity indices measure, it is important that users of these indices have a clear understanding of what the index they are using does measure.

All of the indices considered here fall under our definition of a similarity index: a single number which is a function of the scores

resulting from the comparison of the value measured for an attribute (species) in one sample with the value measured for the same attribute in a second sample. This definition will be clarified with several examples in what follows. The definition restricts consideration to those indices applicable to exactly two samples from a community. The indices considered are all based on a two step process. First, the observed values for each attribute are compared and an attribute similarity score determined. Second, these attribute similarity scores are summed, or otherwise combined to provide the index of paired sample similarity, the similarity index.

The word community is used in the narrow sense implied by the definition given earlier. A sample is indicative of the status of the community at, or over a particular time period and place. A similarity index based on the two samples is calculated in the hope that it will indicate the degree of resemblance between the two ecological populations represented by the samples. If the resemblance is "high" the samples may* be judged to come from the same population. If it is "low" the populations may be judged to be different. If the judgment is "different" some might say that the two communities (under a restricted, more ecological definition of community) are different.

The problems encountered are complex enough so that treating them in the abstract will only complicate matters further. Consequently, data on the study of the borers and foulers (periphyton) community sampled using exposure panels set in Niantic Bay near the NUSCO Power Station at Millstone Point, Connecticut, will be used to clarify ideas. The objective of this report is to investigate the usefulness of similarity indices, not to do an analysis of the Millstone data.

---

*This is not a permissive "may" only a factual "may". Such decisions usually have no objective statistical basis.

The next section discusses the kind of data base frequently subjected to a similarity index analysis. The third section defines the similarity indices discussed and gives examples of their calculation. The fourth section analyzes what the indices actually measure and points out their shortcomings. The Summary is contained in the final section. An appendix contains listings of some Millstone data and other information used in the examples.

A CAUTION

The title of this paper has a Roman numeral I, implying that there will be at least a Roman numeral II on the subject of similarity indices. The specific subject of the second paper is the usefulness of similarity indices as the statistic in a discrimination rule. The criterion of usefulness is that the decision rule have adequate power to detect a change when in fact there is one. The main conclusion in that paper (Johnston, 1976) is that similarity indices are virtually useless in the statistical discrimination problem.

If the reader is looking for methods which will help him decide whether or not two samples come from the same population, he is now advised to look elsewhere. He will not find it in the similarity indices discussed here. If he is interested in finding out why similarity indices have poor power to detect change, he can read this and follow it up with "Similarity Indices: II" to get a statistical demonstration of their lack of power. If he is interested in using similarity indices in descriptive ecology or taxonomy, the following sections should help give an understanding of the construction of many similarity indices.

THE DATA BASE

Studies of ecological communities could be based on many different types of data. This report considers only the (fairly common) type in which a particular instance of a community's status can be represented by a vector of measurements on the attributes used to define the community.

For the Millstone borers and foulers data, (Battelle 1975, and Brown, R. T. and S. F. Moore, 1976) the attributes measured were the species attached to the exposure panels. The species were not preselected for measurement, but were accumulated on a species list as the species were observed. From the start of panel collection in July, 1968, through December, 1975, 203 different species were observed. The measurements made were the percentage of the panel covered by each microscopic species and the number of individuals for each of the macroscopic species. The panels had been in the water for 12 months before being collected. One panel was removed from a rack holding 13 panels each month and replaced by a fresh panel. For this example, four different sampling locations will be considered. Two locations, FN and MH, were in the effluent plume. WP was somewhat removed and the last, GN, was quite removed from the effects of the plume. The data selected to exemplify a typical data base are listed in the Appendix, Table A-1. Only 2 years (24 monthly vectors) were selected for each site. The first year, 1970, was before plant operation began and the second, 1972, was during normal operation of the plant. Only 97 of the 203 total reported species were observed in at least one of the 96 (4 sites x 24 months) samples. Further explanation of the data base is contained in the Appendix.

The data base can be viewed as a set of 96, (4 x 24), vectors of dimension 203 (the total number of species observed over the period). There is a vector for each site by time classification, each vector having 203 elements, many of which are zero. The elements correspond to the

attributes (species). The observation for any species by time by site measurement can be specified by x_{ijk} where:

$i = 1, 2, \ldots, I = 4$ sites

$j = 1, 2, \ldots, J = 24$ time periods

$k = 1, 2, \ldots, N = 203$ species.

For example, if the four sites are assigned the i-subscripts:

Site	FN	WP	MH	GN
i	1	2	3	4

and the months correspond to j in temporal order, and the k subscript is assigned according to the listing order in Appendix Table A-1, then $x_{1,6,2}$ would be the observation for site FN in June of 1970 and the species BALE (Balanus eburneus). The values for $x_{1,6,2}$ is 0.25. For $X_{1,6,4}$ it is 23. Each sample vector of observations will be designated by using a capital X and dropping the k subscript, that is, X_{ij} is the N dimensional vector for site i at time j. In this view the data can be characterized as 4 multivariate time series, one series of 24 time periods for each site.

The data can be divided into the months before full scale operation began (January, 1971, when the plant attained a power level of 200 MW for the first time) and the months after plant operation began. The pre-operational or operational classification of the months, along with the In Plume and Out of Plume classification of sites, partitions the X_{ij} data vectors into four mutually exclusive sets as follows.

	Operating Status	
Plume Location	Pre-Op $1 \leq j \leq 12$	Operating $13 \leq j \leq 24$
Out $i = 1,2$	I	III
In $i = 3,4$	II	IV

The preoperational, in the plume location vectors (set 11), represent the locations which will be in the plume once operation begins. In this form the data are classified according to the paradigm analogous to the Pre-Post, Control-Treated experimental design frequently used by social scientists.

The final consideration for the characterization of the data base is the nature of the observations, x_{ijk}. As pointed out above for the Millstone data, some species provided percentage coverage data, others provided counts of individuals data. Generally, studies of ecological change provide measurement types ranging from attribute presence or absence [variously referred to as binary, dichotomous, binomial or (1,0) data] through the truly continuous type such as biomass. In between these extremes each attribute may provide categorical (multinomial or classificatory) data, ordered categories, or positive integral counts. Of course, continuous or counting data may be reduced to classificatory or dichotomous data by grouping into cells.

Some data specific problems frequently encountered, and present in the Millstone data, are:

- Different measurement types, e.g., counts of individual periphyton per species for some species and percentage of panel coverage for others.
- Missing data due to lost or destroyed samplers.
- A large fraction of attributes (species) are not observed very frequently.
- Different individuals with varying degrees of skill classifying the species and measuring (counting or estimating percent coverage).

Methods for resolving these problems are not discussed, but should be developed whenever such data are to be used.

To provide the reader with an example which could be used to check out ideas by hand and carried along in the text, a subset of the Millstone data consisting of 20 samples by 10 species was selected. The species

were selected to satisfy the a priori condition that the first 10 samples come from one population and the second 10 from a different population. Call these populations A and B. The populations were defined by the probabilities that a species be observed in each sample. As it turned out, the a priori conditions were very closely approximated by species from site FN, the first 10 samples being from 1970 and the second 10 from 1972. If we let p_{Ak} be the probability that species k be present in a sample from population A, similarly for p_{Bk}, and $p_{.k}$ be the probability that a sample from either population contain species k, then the populations for the example are specified by the probabilities in Table 1. The species were selected by looking in Appendix Table A-2 for species which came closest to these a priori probabilities. Table 1 assigns a species

TABLE 1. Population Probabilities of Observing Species in a Random Sample

Species (k)	1	2	3	4	5	6	7	8	9	10
Code	SERW	CREF	OBEX	GRAI	CODD	LEPS	CORC	SERP	CERR	CHAA
ID No.	123	132	68	56	5	105	162	69	21	2
Meas. Type*	C	C	F	F	F	C	C	F	F	F
p_{Ak}	.5	.5	.1	.1	.5	.5	.9	.9	.5	0
p_{Bk}	.1	.1	.5	.5	.9	.9	.5	.5	0	0
$p_{.k}$	.3	.3	.3	.3	.7	.7	.7	.7	.25	0

*For the measurement type; C implies counts, and F fraction coverage.

sequence number (k) to each species, gives the corresponding alphabetic code (Table A.3 gives the full name of the abbreviated species), the numerical ID code and the measurement type--C implies counts of individuals, F implies fraction coverage. If a species with a probability of being observed of 0.9 is classified as a predominant species, of 0.5 as a common species and of 0.1 as a rare species, then it will be noted common species in population A are either rare or predominant in population B, and those common in population B are either rare or predominant in population A. The overall effect when both populations are combined is to give half the species a probability of being observed of 0.3 and the other half a probability of 0.7. This holds except for species 9 and 10, included to show what happens when a species has zero probability of being observed in one or both populations. It is expected that the two populations defined by this construction would be judged to be different.

Table 2 lists the data for these species taken from March through December of (Table A.l) for site FN and 1970 and 1972. Some data were added or set to zero in order to make the samples exactly reflect the population probabilities. Table 2 has the same column headings as Table 1, with the addition of a column assigning sample numbers *(a)* to each sample. (It is hoped that the fact that species head columns and rows correspond to samples in Table 2 but Appendix Table A-1 has the opposite format will not cause the reader problems.) This selected data base will be used in the next section to illustrate the calculations involved in the various similarity indices.

TABLE 2 Basic Data for Example

Species (k)		1	2	3	4	5	6	7	8	9	10	Sample
Code		SERW	CREF	OBEX	GRAI	CODD	LEPS	CORC	SERP	CERR	CHAA	No. (ℓ)
ISU No.		123	132	68	56	5	105	162	69	21	2	
Meas. Type		C	C	F	F	F	C	C	F	F	F	
FN 70 POP.A	1.M	0	1	0	.01	0	0	50	0	0	0	1
	2.A	1	0	0	0	0	2	100	.20	.01	0	2
	3.M	1	0	0	0	0	3	30	.25	.01	0	3
	4.J	3	0	0	0	0	4	23	.05	0	0	4
	5.J	3	0	0	0	0	0	40	.01	0	0	5
	6.A	0	0	0	0	.04	0	30	.01	0	0	6
	7.S	0	1	0	0	.06	0	30	.02	.01	0	7
	8.O	2	1	0	0	.15	4	40	.10	0	0	8
	9.N	0	2	0	0	.01	5	160	.05	.01	0	9
	10.D	0	8	.01	0	.01	0	0	.01	.01	0	10
No. of Times Species Observed		5	5	1	1	5	5	9	9	5	1 0	
FN 70 cum POP.	1	0	0	0	.03	.01	1	0	.01	0	0	11
	2	0	0	0	.02	.01	2	0	.01	0	0	12
	3	0	0	0	.06	.01	1	0	.01	0	0	13
	4	0	0	0	.13	0	0	0	.01	0	0	14
	5	4	0	.01	0	.01	2	0	.01	0	0	15
	6	0	0	0	.01	.02	1	80	0	0	0	16
	7	0	0	.01	0	.01	2	200	0	0	0	17
	8	0	0	.03	0	.01	1	200	0	0	0	18
	9	0	0	.01	0	.01	2	300	0	0	0	19
	10	0	1	.01	0	.01	1	100	0	0	0	20
No. of Times Species Observed		1	1	5	5	9	9	5	5	0	0	
$p_{.k}$		.3	.3	.3	.3	.7	.7	.7	.7	.25	0	

INDICES OF SIMILARITY

GENERAL DEFINITION OF A SIMILARITY INDEX

The use of the word "similarity" is not intended to exclude indices of dissimilarity. One is the logical complement of the other in the sense that similarity indices indicate how "close" two samples are to one another and dissimilarity indices how "far apart" they are. Indices which have been classified under either name will be considered but all of them will be called similarity indices. However, only those indices which use intermediate similarity scores defined on the pairwise comparisons of the values for each community attribute in the two samples will be discussed. This excludes the type of index which only summarizes the attribute values for each sample into a single index and then makes comparisons between these indices. Such indices, called measures of community structure by Pinkham and Pearson (1976) and indices of sample diversity by others, summarize each sample by a single number, thus removing any possibility of considering the relative number of individuals per species and species identity. As will become evident in what follows, many of the indices based on pairwise comparisons of attributes also suffer from this drawback of giving identical values to the similarity measure when species numerosity and/or species identity have obviously different structures over the attributes considered.

The basic definition of a similarity index is here restricted to be a single number which is a function of the pairwise comparisons of the values for each attribute for two samples. Generally, we consider the "attribute space" or "universe of comparison" to be the complete set of attributes for which comparisons are possible. In the Millstone example the attribute space was determined as the study progressed and consists of the 203 species observed to date. A general similarity index can be defined on two N dimensional vectors, say X_{ℓ} and $X_{\ell'}$, where ℓ and ℓ' are two selections of location by time (ij) community samples,* as

*The subscripting of vectors is changed from ij to ℓ merely to avoid a confusing array of subscripts.

$$S_{\ell\ell'} = F(s_{\ell\ell'k})$$

where F is some function of the N pairwise comparisons, $s_{\ell\ell'k}$, and

$$s_{\ell\ell'k} = g(x_{\ell k}, x_{\ell' k})$$

where g is some function of the attribute values observed in the two samples.

Any particular similarity index, $S_{\ell\ell'}$, can be defined by specifying the functions g and F. Someone has proposed almost all of the possible basic operations on two variables for use as g or as part of a two-step calculation of $s_{\ell\ell'k}$. Usually F is a simple sum, or average, over the N attributes, although weighted sums and more complicated functions are sometimes encountered. The functions g and F will be specified for each of the similarity indices in the discussion which follows.

A number of $S_{\ell\ell'}$ are calculated based on a transformation of the basic data to presence-absence form by

$$\delta_{\ell k} = \begin{cases} 1 \text{ if } x_{\ell k} > 0 \\ 0 \text{ if } x_{\ell k} = 0 \end{cases}$$

The data matrix in this case, or when binary data are collected, consists of vectors of 0's and 1's. Table 3 illustrates the application of the $\delta_{\ell k}$ transformation to the data of Table 2.

It is also possible to reduce individual counts and continuous data to more than two categories by defining cells, usually of equal or logarithmic intervals, and classifying the data into these cells. In this case, the transformation is of the form

TABLE 3. The Operator $\delta_{\ell k}$ Applied to the Basic Data for Example

Sample (ℓ)	Species (k) 1	2	3	4	5	6	7	8	9	10	No. Obs.
1	0	1	0	1	0	0	1	0	0	0	3
2	1	0	0	0	0	1	1	1	1	0	5
3	1	0	0	0	0	1	1	1	1	0	5
4	1	0	0	0	0	1	1	1	0	0	4
5	1	0	0	0	0	0	1	1	0	0	3
6	0	0	0	0	1	0	1	1	0	0	3
7	0	1	0	0	1	0	1	1	1	0	5
8	1	1	0	0	1	1	1	1	0	0	6
9	0	1	0	0	1	1	1	1	1	0	6
10	0	1	1	0	1	0	0	1	1	0	5
											45
11	0	0	0	1	1	1	0	1	0	0	4
12	0	0	0	1	1	1	0	1	0	0	4
13	0	0	0	1	1	1	0	1	0	0	4
14	0	0	0	1	0	0	0	1	0	0	2
15	1	0	1	0	1	1	0	1	0	0	5
16	0	0	0	1	1	1	1	0	0	0	4
17	0	0	1	0	1	1	1	0	0	0	4
18	0	0	1	0	1	1	1	0	0	0	4
19	0	0	1	0	1	1	1	0	0	0	4
20	0	1	1	0	1	1	1	0	0	0	5
											40

$$C_{\ell k} = \begin{cases} 1 \text{ if } 0 \leq x_{\ell k} < C_1 \\ 2 \text{ if } C_1 \leq x_{\ell k} < C_2 \\ \vdots \qquad \vdots \qquad \vdots \\ m \text{ if } C_{m-1} \leq x_{\ell k} < C_m. \end{cases}$$

An illustration of this type of transformation will be found in Table 10.

It is possible that a mixture of binary, classificatory, counting and/or continuous data may be encountered in a single data set. Some $S_{\ell\ell'}$ are constructed to handle such mixed data sets, but most require transformation of the data to the measurement level they were designed to handle.

SPECIFIC SIMILARITY INDICES

The number of similarity indices used or proposed for use in studies of the type of data structures exemplified by Table A-1 and Table 2 is no doubt in excess of 50. Pinkham and Pearson (1976) attribute this proliferation to "the general dissatisfaction with the indices and the complexity of the problem," and propose a "new coefficient of similarity" which is claimed to be more sensitive to differences in species identity and numbers of individuals (or biomass) per species than some commonly used similarity indices. This recent addition to the literature is adduced to show that dissatisfaction with current measures of similarity used to describe community changes in pollution surveys is still with us.

There are many ways of classifying similarity indices. Sokal and Sneath (1973) discuss over 30 coefficients classified into: Distance Measures, Association Coefficients, Correlation Coefficients, and Probabilistic Coefficients. G. H. Ball (1966) lists 14 unclassified coefficients used in cluster analysis. They could be classified according to

the measurement level for which they are appropriate. Here we will classify them according to the function, g, used to make the initial pairwise comparison.

Table 4 lists 25 similarity indices classified by seven basically different ways of calculating the similarities between pairs of attributes. The first column of Table 4 gives the class (1 thru 7) for the index. The second column gives the attribute similarity function, g, and other algebraic manipulations of the data required to calculate the attribute similarity score, $s_{\ell\ell'k}$. Under the column headed "Name of Index" the name of the author of the coefficient is given along with the date of a reference in which the author discussed the index. Those indices which have acquired an accepted name in the literature have that name given under the author's name. The next column gives the formula for combining the attribute similarity scores into the index of sample similarity, $S_{\ell\ell'}$. These symbols are given in the last column. The symbols are composed of a letter and a subscript. The subscript is either the initial(s) of the author or the common name of the index. The letter "K" is used for indices based on binary data, "D" for indices commonly referred to as distance measures, "P" for indices based on the fraction of total individuals (or total biomass, etc.) in a sample belonging to a particular species (attribute), and "S" is used for the other indices. An exception to this symbol assignment scheme is made for the formula defining the "Pearson Product Moment Correlation Coefficie'nt" which is almost universally designated by the lower case "r." This is subscripted with a question mark since the algebraic manipulation indicated does not produce Pearson's r.

Indices for Binary Data-

The first class of indices, containing 11 coefficients, is based on the operation of reducing the data to binary presence-absence form, then

TABLE 4. Some **Similarity Indices**

Class	Attribute Similarity Function, g	Name of Index: Author	Date	Computing Formula for $S_{\ell\ell'}$	Symbol
1	δ_ℓ followed by	P. Jaccard (Coefficient of Community)	1908	a/(atbtc)	K_J
	$g(\delta_{\ell k}, \delta_{\ell' k}) = (a,b,c,d)_{\ell\ell'}$	L. R. Dice and T. Sorenson	1945 1948	2a/(2a+b+c)	K_D
	where				
	a = No. of (1,1)	L. W. Watson, W. T. Williams and G. N. Lance (Nonmetric Coefficient)	1966	(b+c)/(2a+b+c)	K_W
	b = No. of (1,0)				
	c = No. of (0.1)	M. Levandowsky (Binary Application)	1971	(b+c)/(atbtc)	S_L
	d = No. of (0,0)				
	N = a+b+c+d	M. D. Mountford*	1962	K_I in the solution to $\text{Exp}[(a+b)K_I] + \text{Exp}[(a+c)K_I] = 1 + \text{Exp}[(a+b+c)K_I]$ $K_I \doteq 2a/(ab+ac+2bc)$	K_I
		R. R. Sokal and C. D. Michener (Simple Matching or Affinity)	1958	(a+d)/N	K_{SM}
		D. J. Rogers and T. T. Tanimoto	1960	(a+d)/(a+d+2b+2c)	K_{RT}
		U. Haman	1961	[(a+d)-(b+c)]/N	K_H
		Yule (Q Coefficient)	1900	(ad-bc)/(ad+bc)	K_Y

*The relation ∎ means "is approximated by."

TABLE 4. Some Similarity Indices (Cont'd.)

Class	Attribute Similarity Function, g	Name of Index: Author	Date	Computing Formula for $S_{\ell\ell'}$	Symbol
1 (Cont.)		Yule (Colligation Coefficient)	1912	$\left[1-\left(\frac{bc}{ad}\right)^{1/2}\right]\Big/\left[1+\left(\frac{bc}{ad}\right)^{1/2}\right]$	K_{YC}
		Binary Product Moment Correlation (See Kendall and Stuart 1973)	1973	$\frac{(ad-bc)}{[(a+b)(a+c)(c+d)(b+d)]^{1/2}}$	K_B
2 a	$\lvert x_{\ell k}-x_{\ell' k}\rvert$	(City Block or Manhattan Metric) (See Sokal and Sneath)	1973	$\sum\lvert x_{\ell k}-x_{\ell' k}\rvert$	D_M
		Czekanowski (Mean Character Distance)	1909	$\frac{1}{N}\sum_{k=1}^{N}\lvert x_{\ell k}-x_{\ell' k}\rvert$	D_{MCD}
b	$x_{\ell k}+x_{\ell' k}$	G. N. Lance and W. T. Williams	1967	$\sum[\lvert x_{\ell k}-x_{\ell' k}\rvert/(x_{\ell k}+x_{\ell' k})]$	D_C
c	$R_k = \max x_{\ell k} - \min x_{\ell k}$	J. C. Gower	1971	$\frac{1}{N}\sum s_{\ell\ell' k}$	S_G
d	$1-\frac{\lvert x_{\ell k}-x_{\ell' k}\rvert}{R_k} = s_{\ell\ell' k}$				

TABLE 4. Some Similarity Indices (Cont'd.

Class	Attribute Similarity Function, g	Name of Index: Author	Date	Computing Formula for $S_{\ell\ell'}$	Symbol
3 a	$(x_{\ell k}-x_{\ell' k})^2$	Euclidean Distance		$\left[\sum(x_{\ell k}-x_{\ell' k})^2\right]^{1/2}$	D_E
		R. R. Sokal (Average Euclidean Distance)	1961	$\left[\frac{1}{N}\sum(x_{\ell k}-x_{\ell' k})^2\right]^{1/2}$	$D_{\overline{E}}$
		Clark (Coeff. of Divergence)	1952	$\left[\frac{1}{N}\sum\left(\frac{x_{\ell k}-x_{\ell' k}}{x_{\ell k}+x_{\ell' k}}\right)^2\right]^{1/2}$	D_{CD}
		Cattell (Coeff. of Pattern Similarity)	1949	$\frac{2\chi_{.5}^2(N)-D_E}{2\chi_{.5}^2(N)+D_E}$	S_C
4 a	$\min(x_{\ell k},x_{\ell' k})$	M. Levandowsky	1971	$1 \quad \frac{\sum\min(x_{\ell k},x_{\ell' k})}{\sum\max(x_{\ell k},x_{\ell' k})}$	S_L
b	$\max(x_{\ell k},x_{\ell' k})$				
c	$\frac{\min(x_{\ell k},x_{\ell' k})}{\max(x_{\ell k},x_{\ell' k})}$	Pinkham & Pearson	1976	$\frac{1}{N}\sum\frac{\min(x_{\ell k},x_{\ell' k})}{\max(x_{\ell k},x_{\ell' k})}$	S_{PP}
5 a	$(x_{\ell k}-\bar{x}_{\ell .})(x_{\ell' k}-\bar{x}_{\ell' .})$	Pearson (Product Moment Correlation)	1896	$\frac{\sum(x_{\ell k}-\bar{x}_{\ell .})(x_{\ell' k}-x_{\ell' .})}{\left[\sum(x_{\ell k}-\bar{x}_{\ell .})^2+\sum(x_{\ell k}-\bar{x}_{\ell' .})^2\right]^{1/2}}$	$r_?$
b	$(x_{\ell k}-\bar{x}_{\ell .})$				

TABLE 4 Some Similarity Indices (Cont'd.)

Class	Attribute Similarity Function, g	Name of Index: Author	Name of Index: Date	Computing Formula for $S_{\ell\ell'}$	Symbol
6 a	$q_{\ell k} = x_{\ell k}/x$	M. G. Johnson & R. O. Brinkhurst (Percentage Similarity of Community)	1971	$\sum_k \min(q_{\ell k}, q_{\ell' k})$	P_J
b	$\min(q_{\ell k}, q_{\ell' k})$				
c	$\lambda'_\ell = \sum_k q_{\ell k}^2$	M. Morisita	1959	$\frac{2\sum_k q_{\ell k}\, q_{\ell' k}}{\lambda'_\ell + \lambda'_{\ell'}}$	P'_M
d	$q_{\ell k}\, q_{\ell' k}$				
7	Complicated function based on ordering all outcomes of each possible pairwise attribute comparison by summing the "probabilities" of observing a less likely outcome.	D. W. Goodall	1966	Complicated calculation based on ordering all possible pairs of sample vectors by summing the "probabilities" of observing a less likely pair than the sample pair actually being compared.	S_P

counting the number of pairwise attribute comparisons falling into each cell of the two by two table:

	Sample a'	
Sample ℓ	.	∪
-		
0	c	

The notation attempting to specify this conceptually simple operation in Table 4 makes the attribute match score implicit, which may lead to some confusion. An example, using the data of Table 2, should clarify the procedure. The δ_{jk} function was applied to Table 2 resulting in Table 3. Suppose that the indices of the first class are to be calculated for samples 7 and 8 (a = 7 and a' = 8). From Table 3 the observed sample vectors are:

k	1	2	3	4	5	6	7	8	9	10
$\delta_{7,k}$	0	1	0	0	1	0	1	1	1	0
$\delta_{8,k}$	1	1	0	0	1	1	1	1	0	0

There are four kinds of results for the comparisons of the attributes, namely (1,1), (1,0), (0,1), (0,0).

The attribute similarity function can be explicitly defined in terms of these four kinds of results by considering four counters (a, b, c, d) which are set to zero at the start of each paired sample comparison. Then,

$$g(\delta_{\ell k}, \delta_{\ell' k}) = \begin{cases} a+1 & \text{for } (1,1) \\ b+1 & \text{for } (1,0) \\ c+1 & \text{for } (0,1) \\ d+1 & \text{for } (0,0) \end{cases}$$

That is, 1 is added to the appropriate counter based on the result of the attribute comparison. Essentially, the attribute similarity function assigns each attribute to one of the four categories. The number of

attributes in each of these four categories is then determined and put into the two by two table from which all similarity indices based on such two by two tables (not just the 11 discussed here) are calculated. Table 5 explicitly shows how the counters cumulate to the final result $(a, b, c, d)_{\ell,\ell'}$ for a = 7 and a' = 8. The sample vectors and the value in each counter after each attribute comparison is given. The final result is (4, 1, 2, 3) and these numbers are positioned in the appropriate cells of the two by two table. The last row of Table 5 gives the values of each of the 11 presence-absence similarity indices for samples 7 and 8.

Since many of the coefficients of the first class use some of the same intermediate quantities, it is efficient and instructive to calculate these first. They are, in addition to the marginal totals of the two by two tables:

a, the number of positive matches,

m = a + d, the number of positive and negative matches

n = b + c, the number of mismatches of either kind

N' = a + b + c = N - d, the number of attributes (species) present in at least one of the two samples

ad-bc, the determinant (differences between the product of the matches and the product of the mismatches) of the two by two table.

For the example,

$(a,b,c,d)_{7,8} = (4,1,2,3)$

and these quantities are:

a	m	n	N'	ad-bc
4	7	3	7	(12-2) = 10

It is then obvious that 5 of the 11 indices do not use d, the number of negative matches. These are:

TABLE 5. Example of Calculations for Indices of the First Class.

Species k	δ_{7k}	δ_{8k}	Comparison Counter (1,1) a	(1,0) b	(0,1) c	(0,0) d
			0	0	0	0
1	0	1	0	0	1	0
2	1	1	1	0	1	0
3	0	0	1	0	1	1
4	0	0	1	0	1	2
5	1	1	2	0	1	2
6	0	1	2	0	2	2
7	1	1	4	0	2	3
8	1	1	4	0	2	2
9	1	0	4	1	2	2
10	0	0	4	1	2	3

$$(a, b, c, d)_{7,8} = (4, 1, 2, 3)$$

Two by Two Table

Sample **7**	Sample 8: 1	Sample 8: 0	
1	a=4	b=1	a+b = *5*
0	c=2	d=3	c+d = *5*
	a+c =6	b+d =4	N = 10

Index	K_J	K_D	K_W	K_{SM}	K_{RT}	K_H	S_L	K_Y	K_{YC}	K_B	K_I*
$S_{7,8}$	.571	.727	.273	.700	.538	.400	.429	.714	.420	.408	.500

*The approximation K_I 2a/(ab+ac+2bc) was used.

$K_J = a/N' = 4/7 = 0.571$

$K_D = 2a/(2a+n) = 8/11 = 0.727$

$K_W = n/(2a+n) = 3/11 = 0.273 = 1 - K_D$

$S_L = n/N' = 3/7 = 0.429 = 1 - K_J$

and $K_I \quad 2a/(ab + ac + 2bc) = 8/(4+8+4) = 8/16 = 0.5$

Jaccard's and Dice's (Sorenson's) indices, K_J and K_D, attain a maximum value of 1.0 when all species present in one sample are also present in the other samples. The complements of these indices, S_L and K_W, respectively, attain a maximum of 1.0 when the number of mismatches, n, equals N'. A value of 1.0 for K_J or K_D indicates maximum similarity, but zero indicates maximum similarity for S_L or K_W.

The calculation of Mountford's Index, K_I, is quite complicated, requiring an iterative process, so only the approximation to K_I given by

$K_I' = 2a/(ab+ac+2bc)$

will be discussed here. (The exact formula will be discussed in the fourth section.) Sokal and Sneath (1973, p. 137) suggest that

$K_I \doteq 2a/(ab+ac+bc)$

is a "good approximation" to K_I, but the factor of 2 multiplying bc is algebraically implied by Mountford's equation for calculating the initial approximation in his iterative calculation of the exact K_I. Mountford's first approximation (in our notation) is

$K_I' = 2a/[2(a+b)(a+c)-(a+b+a+c)a]$

$= 2a/[2N_\ell N_{\ell'} - (N_\ell + N_{\ell'})a]$

where N_ℓ and $N_{\ell'}$ are the numbers of species observed in samples ℓ and ℓ' respectively.

The approximation (and the exact formula) for Mountford's K_I has some algebraic problems at the extreme of complete agreement, or complete

disagreement. In practice having the number of positive matches equal to zero or N (the total number of species observed in either sample) would be rare, but investigating these cases determines the minimum and maximum value of the index.

Generally, if n is the number of mismatches, then the two by two table is

	Sample a'		
Sample a	1	0	
1	N'-n	b	N'-n+b = N_a
0	n-b	(d)	
	N'-b = $N_{\ell'}$		N

and

$$K'_I = \frac{2(N'-n)}{2(N'-n+b)(N'-b) - (N'-n+b+N'-b)(N'-n)}$$

$$= \frac{2(N'-n)}{n(N'-n) + 2b(n-b)}$$

$$= \frac{2a}{na + 2bc}$$

In the case of no agreement, a is zero and K'_I is zero unless b or c is zero. If b or c is zero (and a is still 0), then K_I is the indeterminate quotient, (0/0). The case of a equal to zero and also b or c equal to zero implies that one of N_ℓ or $N_{\ell'}$ is zero; that is, no species were observed in one of the samples. Excluding this trivial case, K'_I has a minimum value of zero. In the case of perfect agreement, a is N' and b = c = 0. Then K'_I is

$$\frac{2N'}{0 \cdot N' + 2 \cdot 0} = \frac{2N'}{0}$$

which is undefined. If there is just one mismatch so that, say, $b = 1 = n$, then $a = N'-1$ and $c = 0$ and K_I' is

$$\frac{2(N'-2)}{2(N'-2) + 2 \cdot 2 \cdot 0} = 1,$$

but if $b = c = 1$, then

$$K_I' = \frac{2(N'-2)}{2(N'-2) + 2 \cdot 1 \cdot 1}$$

$$= \frac{2(N'-2)}{2N' - 2}$$

$$= \frac{N' - 2}{N' - 1}$$

It requires exactly one mismatch for K_I' to attain its maximum defined value of 2. If there are two mismatches, K_I' equals 1 when one of b or c is zero, but its value is dependent on N' when both b and c are unity.

If either b or c is zero, say c, then

$$K_I' = \frac{2a}{na + 2n(0)} = \frac{2}{n}$$

that is, 2 divided by the number of mismatches independent of the total number of different species, N, observed in both samples.

If the number of mismatches, n, is even and is split evenly between b and c then

$$b = c = n/2,$$

$$K_I' = \frac{2a}{na + 2(n/2)^2} = \frac{2a}{n(a+n/2)}$$

$$= \frac{2}{n}\frac{(N'-n/2) - n/2}{(N'-n/2)} = \frac{2}{n}\left(1 - \frac{n/2}{N'-n/2}\right)$$

$$= \frac{2}{n} - \frac{1}{N'-n/2} = \frac{2}{n} - \frac{2}{N'+a}$$

$$= 2\left(\frac{1}{n} - \frac{1}{N_\ell + N_{\ell'}}\right) = \frac{2}{n} - \frac{1}{N_\ell}$$

using the relations

$$a = N'-n,\ N'+a = N_\ell + N_{\ell'},$$

and

$$N_\ell = N_{\ell'} \text{ whenever } b = c.$$

The relation

$$K_I' = \frac{2}{n} - \frac{1}{N_\ell}$$

shows that K_I' will be quite small in rather moderate practical situations. For example, a table with intuitively high association such as

	1	0	
1	80	10	90
0	10		
	90		

has

$$K_I' = \frac{2}{20} - \frac{1}{90} = 0.1 - 0.0111 = 0.0889$$

whereas, for example,

$$K_J = 0.8.$$

If (a,b,c) were (8,1,1), a reduction by a factor of 10, then

$$K_I' = \frac{2}{2} - \frac{1}{9} = 1 - 0.111 = 0.889$$

an increase of a factor of 10.

The three indices which use the total number of matches of either kind, (a+d) = m, were calculated as follows for the example $(a,b,c,d)_{7,8} = (4,1,2,3)$.

$$K_{SM} = m/N = 7/10 = 0.7$$

$$K_{RT} = m/(m+2n) = 7/(7+2\cdot3) = 7/13 = 0.538$$

$$K_H = (m-n)/N = 4/10 = 0.4.$$

Note that all three indices attain a maximum value of 1.0 when m = N, i.e., whenever a match of either kind occurs for all species, since in that case n is zero. The indices K_{SM} and K_{RT} are zero when no matches of either kind occur, but Haman's K_H is negative whenever more mismatches than matches occur and attains its minimum value of -1.0 when m = 0 and n = N.

Haman's K_H uses the difference between the sum of the matches and the sum of the mismatches of both kinds. The last three indices (K_Y, K_{YC} and K_B) use the difference between the product of the two kinds of matches and the product of the two kinds of mismatches in the numerator of their computing formula (at last implicitly). Consequently, these indices, like K_H, vary between -1 for complete disagreement to +1 for complete agreement. Both of Yule's indices use the sum of the products of the matches and mismatches in the denominator, K_{YC} taking the square roots before summing. The binary product moment correlation K_B uses the product of the marginal totals from the two by two table in the denominator. The relation

$$NK_B^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}$$

shows that K_B is formally related to the usual formula for calculating the chi-square statistic, χ^2_{OBS}, for two by two contingency tables. This fact is pointed out merely to indicate that a calculational routine giving χ^2_{OBS} may be used, along with the relation

$$\pm K_B = \left(\frac{1}{N}\chi^2_{OBS}\right)^{1/2}$$

to calculate K_B, provided the sign of (ad-bc) is saved and applied to the result. This fact should not be used to calculate χ^2_{OBS} in an attempt to judge the statistical significance of the degree of association or dissociation since the two by two table generated by the two samples is not a contingency table. More will be said about this in the next section. This concludes the discussion of the calculation of the indices based on the two by two table.

Indices Based on the Absolute Difference

Four examples of the coefficients based on the absolute value of the differences between attribute values as the measure of attribute similarity are given to represent the second class of similarity measures. The first three of these (D_M, D_{MCD} and D_C) are distance (dissimilarity) measures attaining their maximum value when the pair of data vectors are most separated. The fourth, Gower's S_G, is constructed to attain its maximum value when the two vectors are identical so that $|x_{\ell k}-x_{\ell' k}| = 0$. These coefficients can be used with binary or categorical data, but are usually applied to $x_{\ell k}$ which result from counts or continuous measurements of the attributes.

The computing formulas will be applied to samples 7 and 8 of Table 2. Table 6 gives these samples in column vector form and the intermediate values required for calculating the indices. The indices are calculated

TABLE 6. Example of Calculations for Indices of the Second Class

	Column							
	1	2	3	4	5	6	7	8
Species k	x_{7k}	x_{8k}	$\lvert x_{7k}-x_{8k}\rvert$	$x_{7k}+x_{8k}$	R_k	$\frac{\lvert x_{7k}-x_{8k}\rvert}{x_{7k}+x_{8k}}$	$\frac{\lvert x_{7k}-x_{8k}\rvert}{R_k}$	
1	0	2	2	2	4	1	.5	
2	1	1	0	2	8	0	0	
3	0	0	0	0	.03	0	0	
4	0	0	0	0	.13	0	0	
5	.06	.15	.09	.21	.15	.429	.6	
6	0	4	4	4	5	1	.8	
7	30	40	10	70	300	.143	.033	
8	.02	.10	.08	.12	.25	.667	.320	
9	.01	0	.01	.01	.01	1	1	
10	0	0	0	0	0	0	0	
Total				16.18		4.239	3.253	

		Based On	
		N=10	N=7
D_M	$= \sum_{k=1}^{N} \lvert x_{\ell k}-x_{\ell' k}\rvert =$	16.18	16.180
D_{MCD}	$= D_M/N = 16.18/10 =$	1.618	2.311
D_C	$= \sum_{k=1}^{N} \frac{\lvert x_{\ell k}-x_{\ell' k}\rvert}{x_{\ell k}+x_{\ell' k}} =$	4.239	4.239
S_G	$= \frac{1}{N} \sum_{k=1}^{N} \left[1 - \frac{\lvert x_{\ell k}-x_{\ell' k}\rvert}{R_k}\right]$		
	$= 1 - \frac{1}{N} \sum_{k=1}^{N} \frac{\lvert x_{\ell k}-x_{\ell' k}\rvert}{R_k} =$	1 - .325 = 0.675	1 - .465 = 0.535

at the bottom of the table. As with the indices based on presence-absence data, it is possible to exclude species which are not present in either sample being compared by not counting those species which have $x_{\ell k} = x_{\ell' k} = 0$. This reduces the denominator in D_{MCD} and S_G from N = 10 to $N = 7$, and the number of terms in the sum to 7 for all 4 indices. Of course, skipping the species with $(x_{\ell k}, x_{\ell' k}) = (0,0)$ does not change the sums when the quotient (0/0) is defined to be zero.

Calculation of Gower's S_G requires determining the range, over all samples under study, for each attribute. For our 20 sample example, each attribute (k) has $x_{\ell k}$ equal to zero for at least one sample (a), so that the range, R_k, is the maximum value observed. These R_k are given for each species in Table 6. The R_k are used to "range" the absolute differences. Since the absolute difference on a particular species (k) for two samples is less than or equal to the range, the ranged difference lies between zero and unity, inclusive. Subtracting the ranged difference from unity makes Gower's index a measure of the degree of closeness rather than degree of separation and dividing by N assures that S_G will be between zero and unity, inclusive.

It is instructive to consider the application of these coefficients to binary data. For binary data $|x_{\ell k} - x_{\ell' k}|$ is 1 for a mismatch, (1,0) or (0,1), and 0 for (1,1) or (0,0), thus

$$\sum_{k=1}^{N} |x_{\ell k} - x_{\ell' k}| = b + c = D_M \qquad (1)$$

and

$$|x_{\ell k} + x_{\ell' k}| = \begin{cases} 0 \text{ for } (0,0) \\ 1 \text{ for } (1,0) \text{ or } (0,1) \\ 2 \text{ for } (1,1) \end{cases} \qquad (2)$$

Also, for any data type,

$$\max_{\ell} |x_{\ell k} - x_{\ell' k}| = \max_{\ell}(x_{\ell k}) - \min_{\ell}(x_{\ell k}) = R_k \qquad (3)$$

so that for binary data $R_k = 1$. We define $N' = a+b+c$ to be the number of valid matches so that negative matches (or missing data problems) can be eliminated from consideration when they are not considered appropriate to include. The coefficients of the second class, based on binary data, are then related to the coefficients of the first class as follows.

$$D_{MCD} = \frac{1}{N} D_M = \frac{1}{N} (b+c) = \frac{N+(a+d)}{N} = 1 - K_{SM}$$

If negative matches are excluded,

$$D_{MCD} = \frac{1}{N'} D_M = \frac{b+c}{a+b+c} = S_L = 1 - K_J.$$

For the Canberra metric, the ratio

$$\frac{|x_{\ell k} - x_{\ell' k}|}{x_{\ell k} + x_{\ell' k}} = \begin{cases} 1 \text{ for } (1,0) \text{ or } (0,1) \\ 0 \text{ for } (1,1) \\ \text{undefined for } (0,0) \end{cases}$$

so that the attribute match score is the same as for D_M. Dividing the absolute value by the sum has no effect for binary data. Similarly, Gower's "ranging" procedure has no effect since the range of binary attributes is 1, provided the attribute is present in at least one sample vector; otherwise $s_{\ell\ell' k}$ would involve a division by zero. When the range is 1 for all attributes, d must be zero so that

$$S_G = \frac{1}{N'} \sum_{k=1}^{N'} (1 - |x_{\ell k} - x_{\ell' k}|) = \frac{1}{N'} (N' - D_M)$$

$$= \frac{(a+b+c)-(b+c)}{a+b+c} = K_J = 1 - S_L$$

These considerations have shown how the binary coefficients can be made into distance measures defined on the unit hypercube, and point out some not immediately apparent implications of the use of K_{SM}, K_L and K_J and the binary application of S_L or S_G.

When unordered categorical data (e.g., attribute k is color with categories Red, White, Blue) define the measurement type of attribute k, Gower recommends using his coefficient by defining

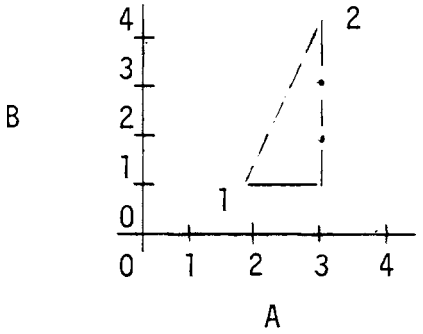
$$= \begin{cases} 1 \text{ if } x_{\ell k} = x_{\ell' k} \\ \text{Ootherwise} \end{cases}$$

With ordered categorical data, each category of the attribute can be assigned its integer rank, or the cell midpoint for categories constructed from counts or continuous data, and any similarity measure appropriate for continuous data used.

Indices Based on the Squared Difference

The third class of indices is exemplified by three more distance measures (D_E, $D_{\overline{E}}$ and D_C) and S_C a function of D_E. Here the measure of attribute similarity is the square of the difference between attribute scores. The first of these indices D_E is the Euclidean distance obtained as the square root of the sum of the squared attribute score differences.

At this point, a few words about the meaning of "distance" might help clarify things. The distance referred to is the distance between observed samples in the N-dimensional attribute space. Since only two dimensions can be easily represented on a sheet of paper, suppose we have only two attributes to consider, and call them attributes A and B. Figure 1 shows such a space with possible attribute values being the integers (0,1,2,3,4) for each attribute. Suppose the two samples (2,1) and (3,4) are observed. These samples determinne the points 1 and 2 in the attribute space. The distance, or degree of dissimilarity, between these two points could be measured in many ways. The two ways used here are City Block Distance



Sample	A	B
1	2	1
2	3	4

$$D_M = 2-3 + 1-4 = 1+3 = 4$$

$$D_E = [(2-3)^2 + (1-4)^2]^{1/2} = [(-1)^2 + (-3)^2]^{1/2} = (10)^{1/2} = 3.16$$

FIGURE 1. A Two-Dimensional Attribute Space

using the absolute value function, and straight line, Euclidean Distance using the squared difference function. The City Block Distance acquires its name from the fact that it can be determined from the "city block map" by counting the minimum number of blocks (units) which must be traversed along the "streets" to go from one point to the other. The Euclidean Distance on the other hand is the straight line shortest distance between the two points. Obviously, the City Block Distance will be greater than the Euclidean Distance, except in the case when all but one of the attributes scores are the same. Then D_M equals D_E, or in the case when the points are identical, and both D_M and D_E are zero.

The picture can be extended to three dimensions (attributes or species) by envisioning a skyscraper covering each block with the stories numbered with the integers to correspond to the number of individuals in the species. Suppose that two observed three-dimensional samples are:

Sample	A	B	C
1	2	1	3
2	3	4	1

Then,

$$D_M = |2-3| + |1-4| + |3-1| = 1+3+2 = 6$$

$$D_E = (1^2+3^2+2^2)^{1/2} = (1+9+4)^{1/2} = (14)^{1/2} = 3.74$$

The conceptualization of the three-dimensional case should require only an elevator for the City Block Distance but it requires the capability to fly in a straight line (through buildings) for the Euclidean Distance. The distance saved by the "flying capability" increases rapidly with the extension to more than three dimensions. This extension is algebraically straightforward, but conceptually difficult. There are no great problems either in extending both kinds of distance measure to continuous data.

Returning to the calculation of the third class of indices, note that the first two, D_E and $D_{\overline{E}}$, are analogous to D_M and D_{MCD}, and D_{CD} is analogous to D_C, except that D_{CD} is an average distance, whereas D_C is a simple sum of attribute similarities.

Cattell's Coefficient of Pattern Similarity is not a distance measure but more of a constructed correlation coefficient. It is assigned to the third class because it is dependent on D_E, the value $2\chi^2_5(N)$ being a constant for N (twice the median value of the chi-square distribution with N degrees of freedom).

The preliminary calculations of Table 6 can be used to calculate the indices based on the square of the difference between attribute values for samples 7 and 8. The Euclidean Distance measure is simply the square root of the sum of the squares of the numbers in Column 3 of Table 6,

$$D_E = [4 + 0.0081 + 16 + 100. + 0.0064 + 0.0001]^{1/2} = [120.0146]^{1/2} = 10.955$$

The Average Euclidean distance is

$$D_{\overline{E}} = [\frac{1}{10}(120.0146)]^{1/2} = (\frac{1}{10})^{1/2}(10.955) = 3.464.$$

Clark's Coefficient of Divergence is the square root of the average of the squares of the numbers in Column 6 of Table 6,

$$D_{CD} = [\frac{1}{10}(1 + 0.1840 + 1 + 0.0204 + 0.4449 + 1)]^{1/2} = [\frac{1}{10}(3.6494)]^{1/2} = 0.6041$$

Reference to a table of the cumulative chi-square distribution (e.g., Ostle, 1963, p. 533) gives,

$$2\chi^2_{0.5}(10) = 9.34,$$

then

$$S_C = \frac{(9.34 - 120.0146)}{(9.34 + 120.0146)} = -0.856.$$

It should be obvious that D_M, D_{MCD}, D_E, $D_{\overline{E}}$ and S_C are all sensitive to the different orders of magnitude of the values for different attributes. For example, the values for attribute 5 are 0.06 and 0.15, so that the difference relative to 0.15 is 60 percent. For attribute 7, the values are 30 and 40, a relative difference of only 25 percent. Attribute 5 contributes 0.09 and attribute 7 contributes 10 to D_M, and 0.09 is only 0.6 percent of D_M but 10 is 61.8 percent. The situation is even worse when

the differences are squared. Attribute 5 contributes only 0.0081 and attribute 7 contributes 100 to D_E^2, or 0.007 percent and 83.3 percent, respectively.

The problem of this implicit weighting of attributes with values orders of magnitude larger than other attributes is usually handled by somehow "normalizing" the individual attribute comparisons so that the order of magnitude differences from attribute to attribute are eliminated. The Canberra Metric, D_C, and the Coefficient of Divergence, D_{CD}, accomplish this normalization by dividing each attribute difference by the attribute sum before doing the other operations, so that the attribute similarity score lies between zero and unity. Gower's S_G divides by the range of the attribute values to accomplish the same end. Cattell's S_C involves normalizing D_E, not the attribute similarities, so that it is also unduly affected by large attribute to attribute variability.

When D_M, D_{MCD}, D_E, $D_{\overline{E}}$ or S_C are to be used, statistical standardization is recommended. (Sokal and Sneath, 1973). This is usually accomplished by calculating the arithmetic average and standard deviation over all samples in the study for each attribute (species). That is, for example, the data of Table 2 are transformed from $x_{\ell k}$ to

$$z_{\ell k} = (x_{\ell k} - \bar{x}_{.k})/sd_k$$

where, for the example,

$$\bar{x}_{.k} = \frac{1}{20} \sum_{\ell=1}^{20} x_{\ell k}$$

$$sd_k = \left[\frac{1}{19} \sum_{\ell=1}^{20} (x_{\ell k} - \bar{x}_{.k})^2 \right]^{1/2}$$

If we again envision our three dimensional city block with skyscraper model, the unstandardized data could be represented by a city with North-South streets 1 mile apart, East-West streets 1 foot apart and skyscrapers with stores 1 inch apart, with only the numbers and not the units given Standardization puts all attributes into the same units, namely standard deviation units and translates the point from which measurements are made from the origin [the vector (0,0,0) in three dimensions] to the average vector $[(\bar{x}_{.1}, \bar{x}_{.2}, \bar{x}_{.3})]$. The average value of the $z_{\ell k}$ for attribute k is zero. In the transformed (standardized) attribute space, values greater than the average are positive and values less than the average are negative The conceptual analogue is to put the origin at the center of the city instead of at a point on the Southwest corner, and to allow the skyscrapers to have basements of adequate depth. Standardization does not guarantee a symmetrical "city" unless the original attribute data was symmetrical.

Table 7 contains the average, $\bar{x}_{.k}$, the standard deviation, sd_k, and the standardized data, $z_{\ell k}$, for the data of Table 2. The averages are indicative of the range of orders of magnitude in the data. Attribute 9 has the smallest average, 2.5×10^{-3}, and attribute 7, the largest, 6.9×10^{1}, so that the data span at least 4 orders of magnitude. Note that Table 7 contains a large number of negative values for $z_{\ell k}$. Most of these result from the $x_{\ell k}$ which are zero, for then

$$z_{\ell k} = \frac{-\bar{x}_{.k}}{sd_k} .$$

Table 8 contains the calculations for the indices based on the two kinds of distance measures. Each index has two values reported. The first, under Z, is the value for the standardized data, and the second, under X, is the value obtained previously using the basic data.

TABLE 7. Standardized Data ($z_{\ell k}$)

Species (k)	1	2	3	4	5	6	7	8	9	10
Average $\bar{x}_{.k}$	0.7	0.7	0.004	0.013	.0185	1.550	69.15	.0375	.0025	0
Standard Deviation sd_k	1.261	1.809	.00754	.0313	.0342	1.5035	84.385	.06904	.01936	0
Sample (R)										
1	-.555	.166	-.531	-.096	-.541	-1.031	-.227	-.543	-.129	-
2	.238	-.387	-.531	-.415	-.541	.299	.366	2.354	.387	-
3	.238	-.387	-.531	-.415	-.541	.964	-.464	3.078	.387	-
4	1.824	-.387	-.531	-.415	-.541	1.630	-.547	.181	-.129	-
5	1.824	-.387	-.531	-.415	-.541	-1.031	-.345	-.398	-.129	-
6	-.555	-.387	-.531	-.415	.628	-1.031	-.464	-.398	-.129	-
7	-.555	.166	-.531	-.415	1.213	-1.031	-.464	-.253	.387	-
8	1.031	.166	-.531	-.415	3.842	1.630	-.345	.905	-.129	-
9	-.555	.718	-.531	-.415	-.248	2.295	1.077	.181	.387	-
10	-.555	4.035	.796	-.415	-.248	-1.031	-.819	-.398	.387	-
11	-.555	-.387	-.531	.543	-.248	-.366	-.819	-.398	-.129	-
12	-.555	-.387	-.531	.224	-.248	.299	-.819	-.398	-.129	-
13	-.555	-.387	-.531	1.501	-.248	-.366	-.819	-.398	-.129	-
14	-.555	-.387	-.531	3.737	-.541	-1.031	-.819	-.398	-.129	-
15	2.618	-.387	.796	-.415	-.248	.299	-.819	-.398	-.129	-
16	-.555	-.387	-.531	-.096	.044	-.366	.129	-.543	-.129	-
17	-.555	-.387	.796	-.415	-.248	.299	1.551	-.543	-.129	-
18	-.555	-.387	3.449	-.415	-.248	-.366	1.551	-.543	-.129	-
19	-.555	-.387	.796	-.415	-.248	.299	2.736	-.543	-.129	-
20	-.555	.166	.796	-.415	-.248	-.366	.366	-.543	-.129	-

TABLE 8. Example of Calculation of "Distance Measures" from Standardized Data.

Species k	z_{7k}	z_{8k}	$z_{7k}-z_{8k}$	$z_{7k}+z_{8k}$	R'_k	$\frac{\lvert z_{7k}-z_{8k}\rvert}{z_{7k}+z_{8k}}$	$\frac{\lvert z_{7k}-z_{8k}\rvert}{R'_k}$
1	-.555	1.031	1.586	.476	3.173	3.332	.500
2	.166	.166	0	.332	4.422	0	0
3	-.531	-.531	0	-1.062	3.980	0	0
4	-.415	-.415	0	-.830	4.152	0	0
5	1.213	3.842	2.629	5.055	4.383	0.520	.600
6	-1.031	1.630	2.661	.599	3.326	4.442	.800
7	-.464	-.345	.119	-.809	3.283	-0.147	.036
8	-.253	.905	1.158	.652	3.621	1.776	.320
9	.387	-.129	.516	.258	0.516	2.000	1.000
10	0	0	0	0	0	0	0
Total	-1.483	6.154	8.669			11.923	3.256

		Based On Z	Based On X
D_M	=	8.669	16.180
D_{MCD}	=	0.867	2.311
D_C	=	11.923	4.239
S_G	=	0.674	0.675
D_E	=	4.258	10.955
$D_{\overline{E}}$	=	1.346	3.464
D_{CD}	=	1.957	0.604
S_C	=	-0.320	-0.856

$$\sum_k (z_{7k}-z_{8k})^2 = 18.1293$$

$$\sum_k \frac{z_{7k}-z_{8k}}{z_{7k}+z_{8k}}^2 = 38.2798$$

Again considering the contribution of attributes 5 and 7 it is seen that:

1) Attribute 5 contributes 2.629 of D_M or 30.3 percent, and attribute 7 contributes 2.661 or 30.7 percent, which is intuitively more reasonable than the 0.6 versus 61.8 percent contributions obtained earlier.

2) Attribute 5 contributed 38.1 percent and attribute 7 contributed 39.1 percent of D_E^2, which is again more reasonable intuitively.

However, standardization has a bad effect on D_C and D_{CD}. This is due to the fact that negative values for $z_{\ell k}$ are possible so that

$$\frac{|z_{\ell k} - z_{\ell' k}|}{z_{\ell k} + z_{\ell' k}}$$

can be greater than unity. This happens for Species 1, 6, 8 and 9 in Table 8. It is also possible for the above ratio to be negative, as for species 7, since the denominator can be negative. The difference between S_G computed on Z and on X is small enough to be attributable to rounding error. This is indicative of the robustness and effectiveness of Gower's ranging procedure against orders of magnitude differences between attribute values. Cattell's S_C indicates a more moderate degree of disassociation between samples 7 and 8 than was indicated by the S_C calculated for the basic data.

Indices Based on the Minimum and Maximum of an Attribute Pair

The first coefficient of the fourth class, based on using the smallest and largest values observed in a pairwise attribute comparison, sums the minimum and maximum values separately before taking the ratio, and the second takes the ratio for each attribute and then sums over the attributes. Levandowsky calls his index "the 1-complement of a modified Jaccard's index." His modification consists in using four ordered categories (0,1,2,3) instead of just (0,1) based on the level of abundance of

a species of plankton (attribute) in each sample. After taking the ratio it is subtracted from unity. If S_L were applied to (0,1) data, the sum over the attribute of $\min(x_{\ell k}, x_{\ell' k})$ would be the number of positive matches (1,1) and the denominator the total number of species observed in either sample (N'). Negative matches are excluded (0,0) having max (0,0) = 0, so that nothing is added to the denominator for such pairwise attribute comparisons. For the binary case

$$S_L = 1 - \frac{a}{a+b+c} = 1 - K_J$$

as pointed out in the discussion of the first class of indices. Presumably Levandowsky's index can be used for any measurement type. Its nature as a "1-complement" makes it a measure of the separation, rather than the closeness, of two samples.

Pinkham and Pearson's index measures the average attribute similarity ratio. For (0,1) data

$$\frac{\min(x_{\ell k}, x_{\ell' k})}{\max(x_{\ell k}, s_{\ell' k}} = \begin{cases} 1 \text{ for } (1,1) \\ 0 \text{ for } (1,0) \text{ or } (0,1) \\ \text{undefined for } (0,0). \end{cases}$$

Again, negative matches are excluded from Consideration. Summing these ratios over all attributes present in either sample gives the denominator, N'. Thus, for the binary case

$$S_{pp} = a/N'$$

the proportion of positive matches. Pinkham and Pearson point out that (0,0) matches can be included by formally defining the ratio of zero to zero to be unity.

The application of S_L and S_{pp} to samples 7 and 8 from Table 2 is illustrated in Table 9. It is obvious that Levandowsky's S_L, like the distance measures D_M and D_E, is unduly affected by order of magnitude

TABLE 9. Example of Calculations of Indices of the Fourth Class

Species k	Original Data					Standardized Data				
	x_{7k}	x_{8k}	min	max	$\frac{\min}{\max}$	z_{7k}	z_{8k}	min	max	$\frac{\min}{\max}$
1	0	2	0	2	0	-.555	1.031	-.555	1.031	.538
2	1	1	1	1	1	.166	.166	.166	.166	1
3	0	0	-	-	-	-.531	-.531	-.531	-.531	1
4	0	0	-	-	-	-.415	-.415	-.415	-.415	1
5	.06	.15	.06	.15	.40	1.213	3.842	1.213	3.842	.316
6	0	4	0	4	0	-1.031	1.630	-1.031	1.630	-.632
7	30	40	30	40	.75	-.464	-.345	-.464	-.345	1.345
8	.02	.10	.02	.10	.20	-.253	.905	-.253	.905	-.280
9	.01	0	.01	.01	1	.387	-.129	-.129	.387	-.333
10	0	0	-	-	-	-	-	-	-	1
Total			31.09	47.26	3.35			-1.999	6.670	4.954

$$S_L = 1 - \frac{31.09}{47.26}$$
$$= 1 - .658$$
$$= 0.352$$

$$S'_L = 1 - \frac{-1.999}{6.670}$$
$$= 1 - (-.2997)$$
$$= 1.300$$

$$S_{PP} = \frac{1}{7}(3.35)$$
$$= 0.479$$

$$S'_{PP} = \frac{1}{10}(4.954)$$
$$= 0.495$$

Without Species 7

$$S_L = 1 - \frac{1.09}{7.26}$$
$$= 0.850$$

$$S_{PP} = \frac{1}{6}(2.6)$$
$$= 0.433$$

differences between attribute ranges. For example, attrtbute 7 contributes 30, or 96.5 percent, of the value of the numerator and 40, or 84.6 percent to the value of the denominator of S_L. Pinkham and Pearson's S_{PP} does not suffer from this sensitivity since the ratio of the values for each attribute is taken before the summation, resulting in the summed scores being normalized to be between zero and unity.

The right half of Table 9 shows what happens when normalization is attempted through standardization to $z_{\ell k}$. The negative values cause acute problems in the calculation of S'_L, based on the standardized data with the result that S'_L exceeds unity. The effect is somewhat masked in the calculation of S'_{PP} where the three species for which no individuals were observed were given attribute similarity scores of 1.0, since for

$$x_{\ell k} = 0, \qquad z_{\ell k} = \frac{-\bar{x}_{.k}}{sd_k} .$$

If these three values are removed from consideration, S'_{PP} is reduced from

$$4.954/10 = 0.495$$

to

$$1.954/7 = 0.279.$$

It is obviously possible to construct a case for which S'_{PP} would be negative.

A "normalization procedure" inferable from Levandowsky's use of classifying the data is attempted in Table 10. The categories are defined by looking at the range for each attribute (R_k, taken from Table 6), retaining a value of zero for $x_{\ell k} = 0$ and partitioning the other possible $x_{\ell k}$ values into categories with values 1, 2 and 3 as equally as possible. (No claim is made that this is in any sense, other than immediate expediency, an optimal partitioning into categories.) The columns headed C_{7k} and C_{8k} give the values assigned to the observed $x_{\ell k}$ by the categorizing

TABLE 10. Indices of the Fourth Class' Based on Categorized Data.

Species k	R_k	Category Definition 0	1	2	3	C_{7k}	C_{8k}	Min	Max	Min/Max
1	4	0	1	2,3	4	0	2	0	2	0
2	8	0	1,2,3	4,5,6	7,8	1	1	1	1	1
3	.03	0	.01	.02	.03	0	0	-	-	-
4	.13	0	.01-.04	.05-.08	.09-.13	2	3	2	3	0
5	.15	0	.01-.05	.06-.10	.11-.15	2	3	2	3	.667
6	5	0	1,2	3,4	5	0	2	0	2	0
7	300	0	1-100	101-200	201-301	1	1	1	1	1
8	.25	0	.01-.08	.09-.16	.17-.25	1	2	1	2	.5
9	.01	0	.01			1	0	0	1	0
10	0	0	-		-	0	0	-	-	-
Total								5	12	3.167

$$S''_L = 1 - \frac{5}{12}$$
$$= 1 - .417$$
$$= 0.583 \qquad S_L = 0.352$$

$$S''_{PP} = (1.167)/7$$
$$= 0.452 \qquad S_{PP} = 0.479$$

Without Species 7

$$S''_L = 1 - \frac{4}{11}$$
$$= 0.636$$

$$S''_{PP} = (2.167)/6$$
$$= 0.361$$

transformation. For example, x_{7k} was zero so C_{7k} is also zero and x_{8k} was 2 which is one of the values (2 and 3) which is assigned a category value of 2. Similarly, for attribute 5, $x_{7,5}$ equal to 0.06 implies C_{7k} is 2 and $x_{8,5}$ equal to 0.15 implies $C_{8,k}$ is 3. For attribute 7, both $x_{7,5}$ and $x_{8,5}$ get $C_{\ell,5}$ equal to 1.

The indices are calculated at the bottom of Table 10 and the previously calculated values given for comparison. The indices S_{PP} and S''_{PP} are both slightly less than 0.5 indicating (perhaps*) that the association between samples 7 and 8 is slightly less than might be expected by chance. The index S''_L also indicates a degree of association slightly less than might be attributed to chance. (Remember S_L is the 1-complement of the sum of the attribute absolute difference scores so that it increases with increasing separation.) But S_L is only 0.352 indicating (under the wild assumption in the last footnote) that the association is somewhat more than could be expected by chance. This is a rather tenuous basis for arguing that categorizing the data and calculating S''_L resulted in a better indicator of the similarity between samples 7 and 8 than did S_L, but it is hoped that it at least indicates some of the problems of comparing indices.

Pearson's Product Moment Correlation Coefficient

Pearson's Product Moment Correlation Coefficient, r, is the only representative considered in the fifth class of indices. This coefficient has a long history of use in biological studies. The population** correlation coefficient, ρ, is defined for two variables which have a mean and a

*The potential for error in comparing values of two difference coefficients in an intuitive way is quite high, even when the indices are calculated on the same data. The comparison is based on the wild assumption that the distribution of S_L'' and S''_{pp}, and of S_L and S_{pp} are symmetric about 0.5, their mid-range.

**The term "population" is here used in the statistical sense to specify that a population parameter, not a statistic estimated from a sample, is being discussed.

variance. It is a normalized measure of the degree of covariance between the two variables, the normalization causing r to be limited to lie between the values -1 and +1 inclusive. In order to calculate an estimate of ρ, a number of instances of the joint occurrence of the two variables are measured so that a pair of numbers is associated with each instance. A common example of two variables is height and weight and instances where these variables might be found are as properties of human beings. An example closer to the subject of this paper would be the measurement of the number of "Serpolid Tubes" (SERW) and Crepidula fornicata (CREF)* observed on an exposure panel. A number of exposure panels (instances) would need to be observed in order to determine r, the sample based estimate of ρ. For the example of Table 2, r would be calculated from the 20 pairs of monthly observations under species 1 and 2.

This rather basic discussion of how data for estimating the most familiar correlation coefficient arises, is intended to provide a basis for pointing out that the correlation coefficient is not even defined for the type of pairwise comparison attempted by similarity indices. The algebra is still valid, and the condition

$$-1 \leq r \leq 1$$

still pertains, but the result is not an estimate of ρ and equating it to Pearson's Product Moment Correlation Coefficient can be very misleading.

Table 2 provides measurements of 20 instances in which ten, not two variables (species) are measured. The algebraic manipulation required to calculate r are carried out on two row vectors, not the column vectors of Table 2. The resulting pair of averages and pair of sums of squares used in the calculation of r are not the averages and sums of squares of two variables. They are a mishmash taken over all 10 variables, and become meaningless when the variables are not commensurate in regard to order of magnitude.

---

*It is true the attributes ("species") are not species specific for the Millstone data.

The usual computing formula for r is

$$r = \frac{\sum x_{\ell k} x_{\ell' k} - x_{\ell .} x_{\ell' .} / N}{[(\sum x_{\ell k}^2 - x_{\ell .}^2 / N)(\sum x_{\ell' k}^2 - x_{\ell' .}^2 / N)]^{1/2}}$$

This formula is more convenient computationally, but it masks the basic indicator of attribute similarity which is explicit in the definitional formula:

$$r = \frac{\sum (x_{\ell k} - \bar{x}_{\ell .})(x_{\ell k} - \bar{x}_{\ell' .})}{[\sum (x_{\ell k} - \bar{x}_{\ell .})^2][\sum (x_{\ell' k} - x_{\ell' .})^2]^{1/2}}$$

Here it is obvious that attribute similarity is the product of the deviations. The quantities in the two equations for r are defined as follows:

$$x_{\ell .} = \sum_{k=1}^{N} x_{\ell k} = \text{total for sample a}$$

$$\bar{x}_{\ell .} = x_{\ell} / N = \text{average value for sample } \ell$$

$$\sum_{k=1}^{N} (x_{\ell k} - \bar{x}_{\ell})^2 = \text{sum of squared deviations from the average for sample a} = \sum_{k=1}^{N} x_{\ell k}^2 - x_{\ell}^2 / N$$

$$\sum_{k=1}^{N} (x_{\ell k} - \bar{x}_{\ell})(x_{\ell' k} - \bar{x}_{\ell' .}) = \text{sum of cross products of deviations from sample averages for samples a and a'} = \sum_{k=1}^{N} x_{\ell k} x_{\ell' k} - x_{\ell .} x_{\ell' .} / N$$

The equivalence of the two forms for the sums of squares and crossproducts can be proved by simple algebra (if one understands the meaning of the summation sign and recognizes that $x_{\ell.} = N\bar{x}_{\ell}$).

Table 11 gives an example of the calculation of $r_?$ for samples 7 and 8. Again, Species 7 contributes an unduly large amount to the value of $r_?$. The contributions to $r_?$ were determined by dividing each crossproduct, $(x_{7k}-\bar{x}_7)(x_{8k}-\bar{x}_8)$, by the common denominator, 1060.32765. Species 7 contributes almost 90 percent of the value of $r_?$ and the other 9 species contribute slightly more than one percent each. At the bottom of Table 11 it is pointed out that without Species 7 $r_?$ drops to 0.0365. When the attributes are normalized by categorizing or standardizing before calculating $r_?$, the results are both slightly less than 0.5. Those in the habit of interpreting r, might say that samples 7 and 8 have a moderate degree of positive association. They certainly wouldn't throw away an observation nor attempt to calculate r, without some kind of normalization of the attributes before applying the sample normalization implicit in the use of $r_?$.

Indices Based on Sample Fractions

The sixth class of coefficient depends on calculating the fraction of the total individuals (or total biomass, etc.) in each sample falling into each attribute. These indices are based on the relative abundances of the species (relative to the total for each individual sample) in each sample. This fraction will be called $q_{\ell k}$ and calculated as $q_{\ell k} = x_{\ell k}/x_{\ell.}$, where x_{ℓ} is the same as above, i.e., the sum of the $x_{\ell k}$ for sample ℓ.

The first of these "percentage" indices was used by Johnson and Brinkhurst (1971) in conjunction with Jaccard's K_J in a study of macroinvertebrates in Lake Ontario. It is simply the sum, over all attributes present in either sample, of the smallest of the two fractions for each attribute. When the two samples have no species in common, $\min(q_{\ell k}, q_{\ell' k})$ will be zero for all N' attributes so that P_J will be zero. P_J attains its

TABLE 11. Example of Calculation of "Pearson's Product Moment Coefficient"

Species k	x_{7k}	x_{8k}	$(x_{7k}-\bar{x}_{7.})$	$(x_{8k}-\bar{x}_{8.})$	$(x_{7k}-\bar{x}_{7.})(x_{8k}-\bar{x}_{8.})$	Contribution to $r_?$
1	0	2	-3.109	-2.725	8.4720	.00799
2	1	1	-2.109	-3.725	7.8560	.00741
3	0	0	-3.109	-4.725	14.6900	.01385
4	0	0	-3.109	-4.725	14.6900	.01385
5	.06	.15	-3.049	-4.575	13.9492	.01316
6	0	4	-3.109	-0.725	2.2540	.00213
7	30	40	26.891	35.275	948.5800	.89461
8	.02	.10	-3.089	-4.625	14.2866	.01347
9	.01	0	-3.099	-4.725	14.6428	.01381
10	0	0	-3.109	-4.725	14.6900	.01385
Total	31.09	47.25'	0.0	0.0		
$\bar{x}_{\ell.}$	3.109	4.725	0.0	0.0		
$\sum(x_{\ell k}-x_{\ell.})^2$			804.3453	1397.7762	1054.11075	$\sum_{k=1}^{10}$ = .99414

$$\sum_{k\neq 7} = 0.09953$$

$$r_? = \frac{1054.11075}{[(804.3453)(1397.7762)]^{1/2}} = \frac{1054.11075}{1060.32765} = .99414$$

Without Species 7, $r_? = .0365$

Based on C_{7k}, C_{8k} of Table 10, $r_? = .4910$

Based on z_{7k}, z_{8k} of Table 9, $r_? = .4879$

maximum value, 1, when $q_{\ell k}$ equals $q_{\cdot, k}$ for all N' attributes, that is, when each sample has the same relative numerosity for each attribute. The more similar the relative numerosities, the closer to unity P_J will be.

The calculation of Morisita's index, P_M, is somewhat more complicated than P_J. Due to certain complexities in Horn's (1966) exposition of P_M, the exact formula for Morisita's P_M as quoted by Horn, and Sokal and Sneath (1973, p. 137), is not given in Table 4. Horn defines P_M (his C_2), using our notation, as

$$P_M = \frac{2\sum_{k=1}^{N} x_{\ell k} x_{\ell' k}}{(\lambda_\ell + \lambda_{\ell'}) x_{\ell \cdot} x_{\ell' \cdot}}$$

where

$$\lambda_\ell = \frac{\sum_{k=1}^{N} x_{\ell k}(x_{\ell k} - 1)}{x_{\ell \cdot}(x_{\ell \cdot} - 1)}$$

Horn points out that λ_ℓ "is Simpson's (1949) index of diversity." Note that the denominator is a constant as far as k is concerned and that subtracting unity in λ_ℓ is based on unbiased estimation arguments which are practically irrelevant. Then we can write:

$$\lambda'_\ell = \sum_{k=1}^{N} \frac{x^2_{\ell}}{x^2_{\ell \cdot}} = \sum_{k=1}^{N} q^2_{\ell k}$$

and

$$P'_M = \frac{2}{(\lambda'_\ell + \lambda'_{\ell'})} \sum_{k=1}^{N} \frac{x_{\ell k}}{x_{\ell \cdot}} \frac{x_{\ell' k}}{x_{\ell' \cdot}}$$

$$= \frac{2\sum_{k} q_{\ell k} q_{\ell' k}}{\lambda'_\ell + \lambda'_{\ell'}}$$

which is the formula appearing in Table 4.

In providing an example of the calculation of these two percentage indices, having two different kinds of measurements (counts and fraction coverage) creates a problem. Previously only r, required calculation of $x_{\ell\cdot}$, and there the "dimensionless" nature of Pearson's r let the question of mixing the two data types slip by. (Note that r, was far from being independent of the dimension used. When the dimensions were those in the basic data, counts and fractions, $r_?$ was 0.9941. But when the dimensions were transformed to attribute standard deviation units, $r_?$ was reduced to 0.4879. This parenthetical digression is purposely placed here to emphasize another important reason for putting the question mark on r as applied in similarity studies.)

In order to apply P_J and P'_M a total, x_ℓ, is required which should be a total of consistent units. For the sake of providing an example, the fractional data was multiplied by 100 and treated as though the result represented counts so that all species have counting data. This was done in Table 12. Again Species 7 unduly affected the results. The perfect match for Species 2 contributed a relatively miniscule amount to both indices and the not so perfect match for Species 7 contributed 72 percent of P_J and 92 percent of P'_M. Without Species 7, the indices were considerably reduced.

Goodall's Probabilistic Index

The final example of a similarity measure is provided by Goodall's "probability based" index. Goodall's (1966) introduction of his index in Biometrics is so complicated that a brief summary of the procedure is extremely difficult. Goodall's mind-boggling definition of pairwise attribute similarity (which must be related to our function, g, if our objective is to be attained) doesn't help much: "Once all pairs of values for an attribute have been ordered in respect of similarity, the similarity index for each pair is defined as the complement of the probability that a random sample of two will have a similarity equal to, or greater than, the pair in question." Emphasis is added to stress the circularity. The

TABLE 12. Example of Calculation of P_J and P_M'.

Species k	Counts x_{7k}	Counts x_{8k}	q_{7k}	q_{8k}	min	$q_{7k}q_{8k}$	Contribution to P_M'
1	0	2	0	.028	0	0	0
2	1	1	.025	.014	.014	.00035	.00073
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	6	15	.150	.208	.150	.03120	.06469
6	0	4	0	.056	0	0	0
7	30	40	.750	.556	.556	.41700	.86462
8	2	10	.050	.139	.050	.00695	.01441
9	1	0	.025	0	0	0	0
10	0	0	0	0	0	0	0
Sum	40	72	1.000	1.001	.770	.45550	.94444
λ_ℓ'			.5888	.3758			

$$P_J = 0.770$$

$$P_M' = \frac{2(.45550)}{.5888+.3758} = \frac{.9110}{.9646} = .9444$$

Without Species 7

$$P_J = 0.7000$$

$$P_M' = 0.9153$$

only hope is to ignore most of what Goodall says and try to figure out what he does. However, this will not be done here.* The interested reader is referred to Goodall's (1966) article. He will there find that the application of the "probabilistic" index is not based on any theory of the probability of an observed outcome (event or sample point) but rather on the observed frequency of the pairwise matches that do occur. The entire set of N samples are treated as if they constituted the statistical population. Goodall's index is based on the observed frequencies of the possible kinds of matches not "probability" in the statistical sense. The computational effort involved to arrive at an index which only has validity as a descriptive measure of similarity makes Goodall's index hardly worth the effort since there is already an ample number of descriptive indices.

This discussion of similarity measures has ignored some statistics frequently employed, or which could be employed, as measures of similarity. Among these is the distance measures of Mahalanobis (1930), excluded because its calculation is based on the comparison of two samples of size greater than one from two (supposedly) different multivariate populations. Excluded for the same reason were Sanghri's (1953), Crovello's (1968) and Pearson's (1926) multi-sample indices. Measures based on ranks, e.g., Kendall's τ and Mann-Whitney's U (Siegel, 1956) were also excluded mainly due to lack of time for adequate exposition.

---

*Such an explanation would require adding more complexity to our notation and unprofitably expand the bulk of this paper.

WHAT DO THE SIMILARITY INDICES MEASURE?

A DEFINITION OF MEASUREMENT

"Measurement is the process of ... assigning numbers to objects or events." (Siegel, 1956). This very broad definition of measurement is adequate for introducing the discussion of what similarity indices measure. The event to which a similarity index assigns a number is the comparison of two vectors, each vector itself consisting of measurements on N attributes. It is desired to measure the closeness, in some sense, of the two more basic events (vectors). Two preliminary measurement operations are required to provide this measure of closeness. First, each sample (exposure panel) is observed and a number assigned to each of the N objects (attributes, species) defining the community under study. This process produces the basic data. The second measurement operation assigns a number to each attribute based on the event of comparing a number assigned to a particular attribute in one sample with the number assigned to the same attribute in another sample. This process produces the attribute similarity score. Finally, the results of the assignments for all N attributes are used to assign a number to the pair of vectors for which the N attribute similarity scores were measured. The "objects or events" to which numbers are assigned in our example are, clearly; species, pairs of species measurements, and vectors or pairwise species scores.

The second essential in particularizing the general definition of measurement to our example is the specification of the "process" for assigning numbers at each stage. In the physical sciences there is usually much concern that the measurement process be in "statistical control" so that if the process is applied to the same object or event a second (third, fourth, ...) time, the number assigned will be within the measurement capability of the process with a high probability, that is, the process is repeatable. Instruments of suitable accuracy are selected and used to assign numbers to the particular properties of the objects that the

instruments are designed to measure. For example, the objects measured might be metal rods of half-inch diameter and the process might be placing one end of the rod against a butt plate so that it is aligned with a meter ruler, zeroed to the butt plate, and the length of the rod recorded by reading the ruler to the last centimeter before the other end of the rod. No doubt this measurement process would almost always give the same results within plus or minus one cm. whether the same or different individuals did it on the same rod. This would be an adequate process for classifying rods for sale at different nominal lengths. However, it would not be adequate if the rods were being studied to determine the characteristic thermal expansion coefficient of such rods.

This example of measuring the length of rods points up some problems with measurements made in ecological studies. First of all, if we want to determine whether or not two rods are in the same lenath category, we can do this by laying them side by side against the butt plate and noting whether or not their end points fall between the same two sequential centimeter marks. If so, they are the same with regard to sale price. The previous measurement process would provide us with numbers which would make the physical comparison of the two rods unnecessary provided each rod was labeled with its nominal length. A similar operation could be defined with respect to diameter providing us with the potential for the bivariate measurement of length and diameter for each rod. The point is that rods can be physically juxtaposed for comparison. Ecological communities, as the term is often and loosely interpreted, cannot. Under the narrow operational definition of ecological community (p. 1) of this report, the communities (exposure panels) could be physically placed next to one another and the relative quantity of each species compared. The resulting measurements will be properties of the exposure panels. Any extrapolation to a broader interpretation is bound up with the adequacy of the sampling scheme for sampling the more broadly defined community and the biological theory supporting the broader definition.* The first important distinction

*Discussions of sampling schemes and broader definitions of community are beyond the scope of this paper.

is that the measurement of "ecological communities" is mediated by sampling methods. We only have fundamental measurement (Campbell, 1928) of the exposure panels, not of the broader community.

The second distinction is that the measurement process is repeatable only for the exposure panels, not for the broader community. "Ecological communities" (broadly interpreted) are complexes of living, interacting entities and so in a constant state of flux.* The "community" can change in the process of being measured, or the measurement process can bias our picture of what the "community" (broad sense) looks like. The only part of the measurement process which can be tested for statistical control is the repeatability of the measurements made by the scientists applying the classification, counting, and percent coverage procedures. It is usually impossible to make truly repeated measurements of the same time and place "community" (broad sense). This is no problem with rods and rulers.

A third problem is that the effect of extraneous variables on the property measured is not predictable within specifiable error limits, based on substantive theory. For example, the thermal expansion of steel rods can be accurately predicted, but the effect of heating the water a number of degrees on the numerosity of a particular species of algae fouling exposure panels cannot. Further, the extraneous variables which must be controlled or measured to make a theoretical determination of the length of a steel rod under various values for these variables are known. Such is not the case for exposure panels. The arrival of one type of organism on the panel may affect the numerosity of another type, just as an increase in temperature may affect the length of the rod, but such relationships are only rarely quantifiable for exposure panels (and other types of environmental samples).

---

*Depending on the particular type of community and the attributes measured, this flux might not be apparent under "normal" conditions for a short time, e.g., forests and communities of larger animals exhibit fairly predictable changes over time.

The point of this discussion is that measurement in ecological studies is quite different from measurement in the physical sciences in three important respects. The object of interest, a broadly defined (or, more frequently, undefined) community, is not directly measurable, consequently, the properties measured are properties of the sample. Since repeated measurements of the same "community" are impossible in most cases, there is no way of estimating sampling error or bias leaving the validity of the measurements as representative of the "community" open to question. Finally, the theory relating concomitant variables to the measurement of interest is not well developed, making inferences regarding causes of observed changes in "communities" open to a multitude of alternatives. (See Brown and Moore, 1976, Section 3 for a discussion of the problems involved in applying the Island Colonization Theory to the Millstone data.)

The process through which numbers are assigned effectively restricts the meaningfulness of any analysis done using the numbers to the particular samples collected. Brown and Moore (1976, p. 31) essentially concede this when they raise the question of how the community on the panel represents the "total community." Their best "hope" is that "the exposure panel is a sort of measure of the adaptiveness of those species arriving on the panel." The restrictive, operational definition of community is the best we can do.

Siegel's (1956) definition of measurement which introduced this section leaves out one important word, "properties." Measurement assigns numbers to properties of objects or events. Table 13 specifies the processes by which the properties of the three kinds of objects or events are assigned numbers. Fundamental measurement occurs when numbers are initially assigned to the species by observing the exposure panel. The processes are particularized for the Millstone data. The measurement of species presence uses species classification alone, assigning a value of unity if the species is observed and a value of zero if it is not. The measurement of numerosity uses the processes of species classification and counting the

TABLE 13. Measurements Made in Calculating Similarity Indices.

Object or Event	Properties	Processes
Species	Presence Numerosity Percent Coverage	Species Classification, Counting, "Eyeball" estimation
Pairs of Species Measurements	Closeness	$g(x_{\ell k}\ x_{\ell' k})$
Vector of Pairwise Species Scores	Magnitude	Sum or weighted Sum

number of individuals of the species. (Some authors use "species numerosity" to mean the total number of different species observed on a panel.) Percent coverage was determined by species classification and estimating the fraction of the panel's area covered by the species.

The second step provides measurements of the property of the "closeness" of two exposure panels with respect to each attribute. The nature of the "closeness" measure is determined by the function g defined on the pair of attribute values (in Table 4). The final object of measurement is the vector of pairwise species scores. A magnitude is assigned to this vector by the sum or weighted sum as specified for each similarity index in Table 4. The end product of this three step measurement process is a measurement of the closeness of two exposure panels. The rest of this section will be devoted to clarifying the meaning of "closeness" for each index.

INFORMATION LOSS COMMON TO ALL SIMILARITY INDICES

The entire set of data taken monthly over many years provides a vast amount of information which could be extracted using many different kinds of statistical manipulations. Reducing the data to a matrix of similarity indices causes much of this potential information to be lost. The exposure panel data potentially contain information on not only the numerosity of each species but also on the time sequence and location for each sample. When the entire data set is viewed as a matrix, say D_{LxN}, consisting of L rows, 1 row for each sample, and N columns, 1 column for each species then there is a potential for using the time series information content for each species (how the species numerosity changes over time). If each sample is viewed as a multivariate observation, then the pattern of species numerosity could be analyzed.

Similarity indices ignore both the time series and response pattern information in the data. They are all defined on an arbitrary pair of sample vectors. These vectors are identifiable as representing particular time and location samples. The samples may be selected for comparison by a similarity index to reflect a potential change between two time periods at the same location or a difference between two locations for the same time period. But the indices are algebraically symmetric so that the fact that one was taken before the other has no impact on the calculation. All the functions assigning pairwise attribute similarity scores are symmetric so that

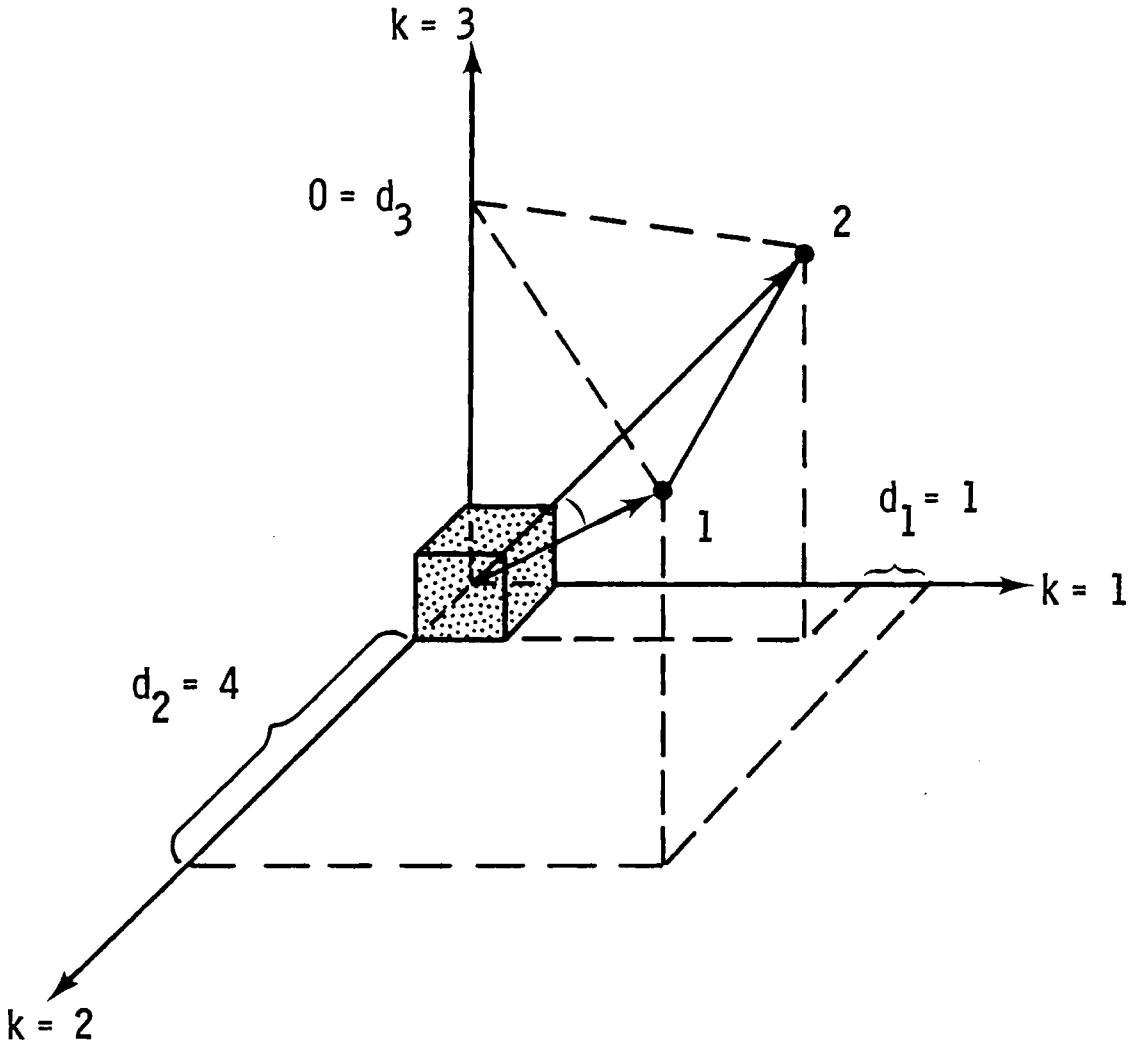
$$g(x_{\ell k}, x_{\ell' k}) = g(x_{\ell' k}, x_{\ell k})$$

The pattern of species numerosity is lost in the calculation of similarity indices since the order in which the species appear in the sample data vector is irrelevant as long as it is consistent from sample to sample. (Goodall's index does involve ordering the possible pairwise comparison for each attribute based on the fraction of samples in which each attribute value appears. The species are thus differentially weighted to give

different contributions to S_p. The less common and less numerous species get relatively more weight. The pattern is not preserved but at least different species are differently weighted.) No matter what the species order in the data vector, the species which contributed to the value of the similarity index are lost in the process of summation.

A MODEL FOR DEFINING "CLOSENESS"

Use of similarity indices implies the decision to investigate how near, in some sense, two samples are to each other. The concept of nearness or closeness connotes a potential for measuring the distance between the two objects that are close. The fact that the objects are vectors encourages use of a vector space in which the degree of closeness can be measured. The geometric model to be used to clarify the meaning of closeness is the attribute space of N dimensions discussed earlier in connection with distance measures. Such a space of two dimensions was illustrated for two species in Figure 1. Figure 2 illustrates a three dimensional space. The two points, numbered 1 and 2, are the locations of the end points of the vectors (samples) 1 and 2 with the three attribute values $X_1 = (6,5,5)$ and $X_2 = (5,1,5)$. The heavy black line represents the Euclidean distance between the two points, and the heavy black arc represents the angle between the two vectors. The dashed lines are the projections of the points onto the (k=1, k=2) plane and onto the three axes. The distances between the projections on the axes is a measure of attribute (species) closeness. The angle between the vectors from the origin to the two points is also a measure of the degree of closeness of the two points. The length of the vectors can be determined by applying the Euclidean Distance formula, D_E, to the null vector and X_1 or X_2. This produces 9.27 as the length of X_1 and 7.14 as the length of X_2. The Euclidean distance between X_1 and X_2 is $(1^2 + 4^2 + 0^2)^{1/2} = 4.12$. This model will be referred to as each index is discussed.



ℓ / k	1	2	3
1	6	5	5
2	5	1	5
d_k	1	4	0

FIGURE 2. Example of Two Vectors in Three-Space

THE PROPERTY OF THE SAMPLES MEASURED BY THE INDICES

Presence, Absence Indices

All of the indices of the first class disregard the information on species identity and number of individuals per species. Their use in the study of the effects of pollution thus implies acceptance of the thesis (or assuming) that two communities can be differentiated based on only the number of species classified into the four cells of the two by two table. This is an ecological question currently under dispute and so could be a valid thesis as far as an ecological layman (statistician) is concerned. However, the theoretical considerations of Maillfeur (way back in 1929) and the practical examples of Pinkham and Pearson (1976) plus many intervening criticisms of this type of coefficient, particularly Jaccard's (e.g., Morisita, 1959; Levandowsky, 1971; Ashby, 1935, Williams, 1949), wouldlead a novice to wonder whether the Jaccard coefficient K_J, is useful for anything since it is so dependent on size of the sample and definition of the population sampled. The other coefficients based on the ratio of sums of the (a,b,c,d) from the two by two table would seem to be open to the same type of criticism, given R. A. Fisher's, et al., (1943) theory that the frequency distribution of genera with different numbers of species can be represented by a logarithmic series (see e.g., Williams, 1949).

Negative Matches Excluded

These five indices, Jaccard's K_J, Dice's K_D, the Nonmetric Coefficient K_W, Levandowsky's S_L and Mountford's K_I, consider only whether each species is present or absent and exclude from consideration in any pair of vectors all species which are not present in either vector. The numbers of individuals of each species is thus lost.

The full (N dimensions) attribute space is contained in the unit hypercube of N dimensions. Such a hypercube has 2^N points. For the three dimensions of Figure 2 each possible vector would have its end point at one of the eight vertices of the unit cube sketched in Figure 2. For the full 203 species space of the Millstone data there would be 1.3×10^{61} possible vectors. For the example of 10 species, there are 1024 possible vectors.

However, the fact that negative matches (a species is not present in either sample) are excluded will cause the dimension of the space to be less than full (N) whenever such species are excluded. The dimensions excluded may differ from pair to pair, so that not only the dimensionality but the specific dimensions removed may change. This raises the question of what space is involved when a matrix of these similarity indices is subjected to cluster, factor, or principal components analysis. It also makes the meaning of the comparison of two similarity indices from the same set of data questionable.

For example, if N=100 species were observed in a data set, and each is observed in at least one of the two sample vectors and 25 are observed in both, then Jaccard's $K_J = 0.25$. But if only 50 species (of the 100 in the full attribute space) are observed in at least one of the two samples being compared, and again 25 species are observed in both, K_J would be 0.5. Yet in both cases 25 positive matches were observed. Since the attribute space is reduced by effectively excluding from it species which are not observed in a particular pair of samples, these five indices are actually increased in proportion to the number of attributes excluded. Continuing this example, let d be the number of negative matches. Multiplying K_J by unity in the form of N/N gives

$$K_J = \frac{N}{N}\frac{a}{(N-d)} = \frac{N}{(N-d)}\frac{a}{N} .$$

It is obvious that if the number of positive matches remains fixed, then K_J is going to increase as d increases since N/(N-d) is greater than or equal to unity, equality holding when d is zero. Jaccard's K_J is thus the product of the fraction of positive matches, (a/N), and the inverse of the fraction of species which are observed in at least one of the two samples in the full attribute space.

Table 14 gives an example of the use of the proportionality factor to calculate K_J for four values of d. The number of positive matches was held constant at 25 and K_J was calculated using the formula

$$K_J = a/N' \ .$$

For d=0, N'=N and K_J = 0.25 which is the fraction of positive matches in the full attribute space. Then the inverse of the fraction of species observed in either sample was calculated and multiplied by 0.25 giving the same result as the previous calculation of K_J. Obviously, excluding negative matches has a pronounced effect on the value of K_J in direct proportion to the number of negative matches. This is true also for the other four indices, K_D, K_W, S_L and K_I'. By excluding negative matches, all five of these indices are effectively weighted by the number of negative matches. This changing of the dimensionality of the full attribute space based on the two particular sample vectors being measured for similarity leads to problems of comparability of the resulting values. It also raises the statistical problem of using the sample results to define the analysis.

These five coefficients are related to the vector space model (see Figure 2) as follows. The full attribute space is the unit hypercube in N-space. Figure 3 gives an example of the unit cube in 3-space. The Euclidean distance between any two points, say A, and $A_{\ell'}$, is given by

$$\sum_{k=1}^{N} \left[(\delta_{\ell k} - \delta_{\ell' k})^2 \right]^{1/2}$$

Now

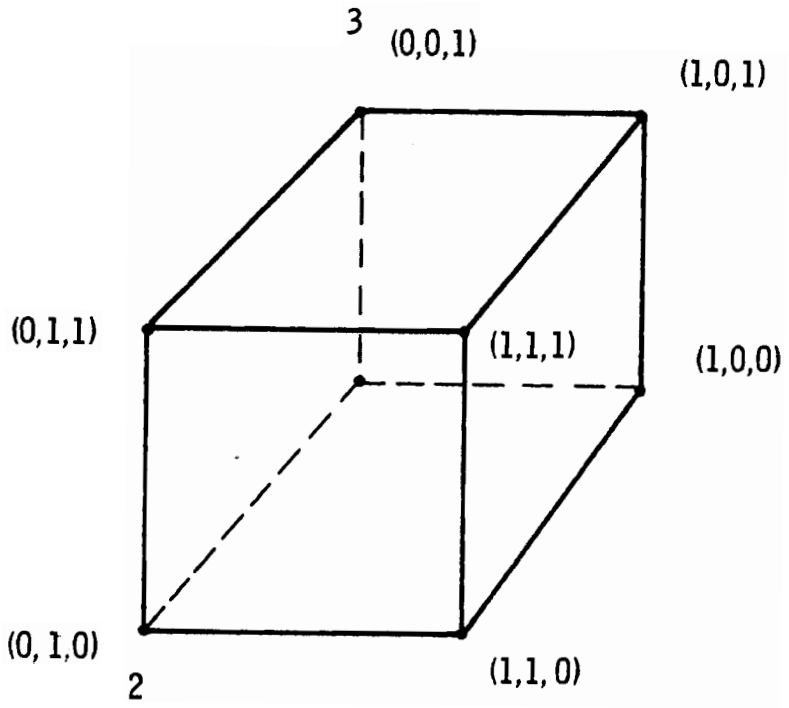
$$(\delta_{\ell k} - \delta_{\ell' k})^2 = \begin{cases} 1 & \text{for } (1,0) \text{ or } (0,1) \\ 0 & \text{for } (1,1) \text{ or } (0,0) \end{cases}$$

so that, also

$$(\delta_{\ell k} - \delta_{\ell' k})^2 = |\delta_{\ell k} - \delta_{\ell' k}|$$

TABLE 14. Example of Calculation of K_J Using Proportionality Factor Based on d.

	d=0	d=25	d=50	d=62
a	25	25	25	25
N' = N-d	100	75	50	38
K_J	.25	.333	.5	.658
N/(N-d)	1	1.3333	2	2.6316
.25N/(N-d)	.25	.3333	.5	.658
2a	50	50	50	50
2a+n	125	100	75	63
K_D	.4	.5	.667	.794
K_W	.6	.5	.333	.216
	.75	.667	.5	.342
	38	25	13	7
	37	25	12	6
	.0107	.0200	.0534	.1222



THREE DIMENSIONAL POINTS (Δ_ℓ)

	k		
	1	2	3
0	0	0	0
1	0	0	1
2	0	1	0
3	0	1	1
4	1	0	0
5	1	0	1
6	1	1	0
7	1	1	1

FIGURE 3. The Unit Cube and Its Eight Binary Points

and

$$\sum_{k=1}^{N} (\delta_{\ell k} - \delta_{\ell' k})^2 = \sum_{k=1}^{N} |\delta_{\ell k} - \delta_{\ell' k}| = b+c = n,$$

the number of mismatches of either kind.

The number of positive matches can be obtained as the scalar product of the two vectors

$$a = \sum_{k=1}^{N} \delta_{\ell k} \delta_{\ell' k} = \Delta_{\ell} \cdot \Delta_{\ell'}$$

since

$$\delta_{\ell k}\ \delta_{\ell' k} = \begin{cases} 1 & \text{for } (1,1) \\ 0 & \text{for } (1,0),\ (0,1) \text{ or } (0,0). \end{cases}$$

Then, the number of species observed in either sample ℓ or ℓ' is

$$N' = a + b + c = a + n$$

and

$$d = N - N'.$$

It is then apparent that

$$S_L = \frac{n}{N'} = \frac{\sum_{k=1}^{N} |\delta_{\ell k} - \delta_{\ell' k}|}{N'}$$

which is the Mean Character Distance of Czekanowski, D_{MCD}, in the reduced attribute space. S_L is thus a measure of the degree of separation of the two samples. Since the square of the difference and the absolute difference are the same for binary data, S_L is also D_E^2 in the reduced attribute space. Straightforward algebra shows,

$$K_J = 1 - S_L$$

so that Jaccard's K_J is the complement of a measure of separation and measures the closeness of the two samples.

The indices K_D and K_W distort the distance interpretation by giving double weight to positive matches. The impact of this weighting can be seen by comparing K_D with K_J and K_W with S_L. The indices K_J and K_D are the same except that Kg uses 2a instead of a. We can write

$$\frac{K_J}{K_D} = \frac{a}{a+N} \frac{2a+n}{2a} .$$

Then,

$$\frac{K_J}{K_D} = \frac{2a+n}{2(a+n)} = \frac{a+N'}{2N'} = \frac{1}{2} \left(\frac{a}{N'} + \frac{N'}{N}\right) = \frac{1}{2}(K_J+1),$$

so that,

$$K_D = \frac{2}{K_J+1} K_J .$$

Since K_J cannot be greater than 1

$$\frac{2}{K_J+1} \geq 1$$

so that

$$K_D \geq K_J$$

equality holding when $a = N'$ and $K_D = K_J = 1$ or when $a = 0$ and $K_D = K_J = 0$. Weighting a by 2 results in making the measure of closeness (K_D) larger than (K_J) when such weighting is not done. Weighting positive matches thus makes the measure of closeness, K_D, nearer to its maximum value, unity. The opposite holds for K_W, which increases with greater separation. Writing

$$\frac{S_L}{K_W} = \frac{n}{a+n}\,\frac{2a+n}{n} = \frac{2a+n}{a+n} = \frac{a+N'}{N'} .$$

Then

$$K_W = \frac{S_L}{1+K_J}$$

Since $1+K_J$ is greater than or equal to unity, K_W will be less than S_L unless both equal 1 or zero. The weighting of a thus increases the distance measure making K_W closer to zero than S_L when the vectors are closer in the Euclidean sense.

All four of these indices can be written in terms of a, N' and K_J so that K_J contains all of the objective information summarized by any one of them.

$$K_J = a/N$$

$$K_D = 2K_J/(K_J+1)$$

$$K_W = (1-K_J)/(1+K_J)$$

$$S_L = 1-K_J .$$

These coefficients were calculated in Table 14 for comparison.

The relationship of Mountford's K_I to the vector space model is beyond the understanding of the author. However, the following facts about K_I might be helpful. K_I is based on the theory that the "species frequency distribution found in random samples of natural populations could be well described by the logarithmic-series distribution" (Mountford, 1962), a theory proposed by Fisher, Corbet and Williams (1943). Under this theory the logarithmic series

$$\alpha x,\ \alpha\frac{x^2}{3},\ \alpha\frac{x^3}{3},\ \ldots,\ \frac{\alpha x^{\nu}}{\nu},\ \ldots$$

gives, in the first term, the number of species with one individual; in the second term, the number of species with two individuals; and in general, the ν^{th} term gives the number of species with ν individuals. The value of x varies from sample to sample.

K_I is the inverse of the "Index of Diversity," a, which is the parameter of the logarithmic series and which "is a constant for all samples of whatever size from the same population" (Mountford, 1962). Algebraic manipulation of the expected number of species and expected total of individuals per sample, under the assumption that the logarithmic series is appropriate, results in elimination of the number of individuals from the algebra and a relation in only the number of species and the inverse of α.

This is

$$\exp[(a+b)K_I] + \exp[(a+c)K_I] = 1 + \exp[(a+b+c)K_I].$$

If we note that $(a+b) = N_\ell$, the number of species observed in the first sample and $(a+c) = N_{\ell'}$, the number observed in the second sample, then the defining relation is

$$\exp[N_\ell K_I] + \exp[N_{\ell'} K_I] = 1 + \exp[N' K_I]$$

Mountford suggests that K_I can be found by "substituting the particular values of a (our N_ℓ), b (our $N_{\ell'}$) and j (our a so that $N' = N_\ell + N_{\ell'} - a$) in the above equation and interpolating within the table of exponentials," using as the first value in our notation

$$K_I' = \frac{2a}{ab+ac+2bc} = \frac{2a}{2N_\ell N_{\ell'} - (N_\ell + N_{\ell'})a}$$

Mountford (1962, p. 45) also provides a nomograph for determining K_I. The approximation K_I' overestimates K_I but provides the starting value for the iterative procedure which attempts to balance the defining relation above.

For the two-way table used in the calculation of K_I'

T1	1	0	
1	80	10	90
0	10		
	90		

K_I' was 0.0889. Using 0.08 as the initial trial value gives

$$\exp[N_{\ell}K_I] + \exp[N_{\ell}'K_I] - \exp[N'K_I]$$

$$= 267.86 - 2980.96 = -302.1.$$

This is less than unity indicating 0.08 is too large. Trying 0.07 gives

$$1089.14 - 1096.63 = -7.5$$

indicating 0.07 is too large. Trying 0.065 gives

$$694.47 - 665.14 - 29.33$$

which is greater than unity indicating 0.065 is too small. Continuing in this fashion it is found that $K_I = 0.0692$ gives

$$1013.48 - 1012.32 = 1.16$$

which is close enough to unity, giving K_I to three significant digits.

Investigating the algebraic expected value* of K_I under the logarithmic series assumption sheds light on its meaning. In two samples of the same size,** the two by two table would be

*This is not a statistical expectation since no statistical distribution is involved.

**Sample size here refers to the size of the sampling units, e.g., a one square meter area or an exposure panel left in the water for 12 months.

T2	1	0	
1	N_ℓ-αlog 2	b	N_ℓ
0	c		
	N_ℓ		

where

$$b = c = \alpha \log 2 = 0.693\ a$$

based on the results of Williams (1949). Applying the defining relation for K_I to T2, we get the surprising results

$$2 \exp[N_\ell K_I] = 1 + \exp[(2N_\ell - a)K_I]$$

but $K_I' = 1/a$ and $a = N - \alpha \log 2$, so that

$$2 \exp[N_\ell/\alpha] - \exp[(2N_\ell - N_\ell + \alpha \log 2)/\alpha] = 1$$

$$\exp[N_\ell/\alpha]\left[2 - \exp[(\alpha \log 2)/\alpha]\right] = 1$$

$$\exp[N_\ell/\alpha][2-2] = 1$$

$$0 = 1$$

This contradiction comes from using Williams' result for "a" which only holds in the limit. His expression for the number of species in two Samples of the same size is

$$S_2 = \alpha \log(1 + 2M/\alpha)$$

where M is the total number of individuals in a single sample. The expression for the number of individuals in a single sample is

$$S_1 = \alpha \log(1 + M/\alpha).$$

This is our N_ℓ. The increase in the number of species is then

$$S_2 - S_1 = \alpha \log\left[\frac{1 + 2M/\alpha}{1 + M/\alpha}\right]$$

$$= \alpha \log\left[\frac{\alpha + 2M}{\alpha + M}\right]$$

$$= \alpha \log\left(1 + \frac{M}{\alpha+M}\right)$$

It is true that in the limit

$$\lim_{M\to\infty}\left(1 + \frac{M}{\alpha+M}\right) \to 2$$

but for any given M, the difference from 2 is

$$\delta = 2 - \left(1 + \frac{M}{\alpha+M}\right) = 1 - \frac{M}{\alpha+M} = \frac{\alpha}{\alpha+M}$$

Then the expected number of positive matches (the number of species common to both samples) is exactly

$$a = N_\ell - \alpha \log\left(1 + \frac{M}{\alpha+M}\right)$$

under the logarithmic assumption and Williams' algebra. The defining relation for K_I then reduces to

$$\exp[N_\ell/\alpha]\left[2 - \left(1 + \frac{M}{\alpha+M}\right)\right] = 1$$

or

$$\exp[N_\ell/\alpha] = 1$$

From which, taking logarithms,

$$(N_\ell/\alpha) + \log \delta = 0$$

and

$$K_I = \frac{1}{a} = -\log \delta/N_\ell$$

The algebraic expected value of K_I is thus a function of how close the ratio $M/(\alpha+M)$ is to one [or $\alpha/(\alpha+M)$ is to zero]. The exponent is simply the number of species ($N_\ell = N$ by the logarithmic assumption) observed in each sample divided by the index of diversity. It does not seem reasonable to base a measure of species presence/absence similarity on how closely a ratio based on the number of individuals in the samples approaches unity, the empirical evidence for the applicability of the logarithmic series notwithstanding. At least a great regard to potential rounding error must be given wnenever K_I is used.

Using the two by two table T1, which has $K_I = 0.0692$ and $a = 14.45$, it can be determined that M is 7311 under the logarithmic assumption. Then

$$\exp[N_\ell/\alpha][2 - (1 + \frac{M}{\alpha+M})]$$

$$= \exp[90/14.45][2 - (1 + \frac{7311}{7325.45})]$$

$$= 506.93[2 - 1.99802]$$

$$= 1013.86 - 1012.86 = 1.00.$$

(Fortuitous rounding error makes this result exactly 1.00 whereas it was 1.16 for the previous calculation to determine $K_I = 0.0692$.) If 6 were (2 - 1.999) = 0.001 instead of 0.002 as above, M would have to be almost doubled to 14,435.55* for α to still be 14.45. If M were 14,435.55 then

$$N_\ell = 14.45 \log(1 + \frac{14,435.55}{14.45}) = 99.82$$

and, as before, b = c = 10.01 so that a = 89.81 determining the tables:

---

*Two decimal places will be carried for the expected counts to avoid rounding error.

T3	1	0	
1	89.81	10.01	99.82
0	10.01		
	99.82		

T3'	1	0	
1	90	10	100
0	10		
	100		

For the two by two table T3, the defining relation is

$$\exp[99.82/14.45][2 - 1.999]$$

$$= 1000.20(0.001) = 1.0002$$

and K_I can be determined by

$$K_I = -\text{en}\ \delta/N = -\ell n(0.001)/99.82 = 0.0692$$

as for TI. When the expected counts are rounded as in T3', K_I is 0.06908, resulting in loss of the third significant digit.

This shows that K_I is not sensitive to doubling the sample size but is very sensitive to the value of δ. In practical situations where the numbers of individuals may actually vary considerably in two samples from the same population, this sensitivity to δ may be important. In any case careful investigation of the appropriateness of the logarithmic series assumption for the samples under study should be made before using Mountford's K_I.

It is the author's conjecture that K_I, being the inverse of α, is just the inverse of the Index of Diversity for the two sample vectors combined and so not a good tool for indicating the degree of similarity between the two vectors in the absence of a comparison of calculated K_I with the expected value of K_I. The truth or falsity of this conjecture and its implications were not investigated due to already excessive bulk of the paper and lack of time.

In summary, K_I depends on M, the number of individuals observed in the sample, and its expected maximum value decreases with increasing M. Mountford's K_I could not be related to the vector space model, but it is related to the index of diversity, α, under the logarithmic series assumption. The property of the two samples measured by K_J, K_D, K_W and S_L is the separation or, its complement, the closeness of the binary vectors representing the samples in the unit hypercube of dimension N'. The major drawback to using these coefficients is that the dimension, N', can change from pairwise comparison to pairwise comparison. This makes the measure of closeness, or separation, a function of the number of species in the full attribute space which are excluded because they were not observed in the particular pair of vectors for which the index was calculated. If one still has reason to use one of these indices, K_J is recommended for its simplicity.

Negative Matches Included--Additive Functions

These indices are K_{SM}, K_{RT} and K_H. The simplest of these is K_{SM}, the Simple Matching or Affinity Coefficient of Sokal and Michener. It is the proportion of matches of either kind in the full attribute space. K_{SM} increases with increasing closeness and K_{RT} and K_H are related to K_{SM} as follows:

$$K_{SM} = m/N$$

$$K_{RT} = \left(\frac{N}{N+n}\right)K_{SM}$$

$$K_H = 2K_{SM}-1.$$

The information used by these indices is contained in m, the number of matches of either kind, and N, the dimension of the attribute space, since $n = N-m$. .

It was shown above (p. 31) that

$$K_{SM} = 1 - D_{MCD}$$

where D_{MCD} is the Mean Character Distance defined on the full attribute space. Also,

$$K_{SM} = 1 - D_E^2$$

in the unit hypercube of dimension N. K_{RT} will be closer to zero than K_{SM} since N/(N+n) is less than unity unless n is zero, so that a perfect match on all attributes occurs. K_H merely uses the fact that K_{SM} attains a maximum of 1 to provide an index which indicates degrees of closeness ranging from -1 to +1. Values of K_{SM} less than 0.5 will make K_H negative and values of K_{SM} greater than 0.5 will result in positive K_H. Since it is not generally true that K_{SM} is symmetrically distributed about 0.5 (See Goodall, 1967), one fails to see the logic in constructing an index apparently based on the assumption that it does. There is no assurance that K_H will be negative half the time and positive the other half unless the probability of observing each of the N species is 0.5. This is invariably far from the case with the type of data considered here. The effect of K_H is simply to spread K_{SM} over the range -1 to +1 instead of 0 to 1.

The Simple Matching Coefficient captures all of the information contained in m and N. Spreading it out over -1 to +1 by using K_H or modifying the scale by multiplying it by N/(N+n) only confuses its interpretation. Unlike the indices which exclude negative matches K_{SM} (and the functions of K_{SM}, K_{RT} and K_H) are defined on the same attribute space from pair to pair in the same data set. This holds as long as the attribute space is fixed for the study. If, as in the Millstone data, the species list is accumulated as the study progresses, the K_{SM} calculated for a pair of vector samples early in the study would not be comparable to a calculation later in the study.

In summary K_{SM} and the function of K_{SM}, K_{RT} and K_H, measure the closeness of the two binary vectors representing the samples in the unit hypercube of dimension N. Since the dimension remains the same from pairwise comparison to pairwise comparison within the same data set, the resulting similarity indices have a common basis for comparison.

Negative Matches Included--Multiplicative Functions

These three indices, Yule's K_Y and K_{YC} and the Binary Product Moment Correlation Coefficient, K_B, have the determinant of the two by two table in their numerator. These coefficients were designed to behave like correlation coefficients in that a value of -1 indicates complete negative association (if species k is present in sample a it is absent in sample a' or vice versa), a value of zero indicates no association between the samples, and a value of +1 indicates complete positive association (if species k is present in sample a it is also present in sample a', and absence in a implies absence in a').

Kendall and Stuart (Vol. 2, third edition, 1973) give a good discussion of these three indices in the first 27 Sections of Chapter 33. However, once one understands that their discussion and results are concerned with the association between two attributes based on a sample of size N, he will realize that this wealth of information is not generally applicable to the type of two by two table arising from the comparison of binary vectors under study here. Some of the algebra is applicable. In terms of a geometric model they are concerned with N points in two-space, we are concerned with two points in N-space. Fienberg and Gilbert (1970) give a geometric model for two by two contingency tables but, again, we don't really have a contingency table. A two by two contingency table classifies N sampled individuals into the cells of the table based on the observation of presence or absence for each of two dichotomized attributes. In calculating these indices for the periphyton data, N dichotomized attributes (species) are classified into the cells of the table based on the presence and absence pattern observed for each attribute in the two samples. The

resulting two by two tables look the same but the interpretations are quite different. In the case of a contingency table "a" is the number of individuals in the sample of size N which have both attributes, and a/N is an estimate of the probability that individuals in the population from which the individuals were sampled will have both attributes. In the table based on the periphyton data "a" is the number of species which are present in both samples. The quantity a/N estimates, if anything, the average probability that in two randomly selected samples an arbitrary species will be present in both. The practical meaning and utility of such an estimate is dubious when the probability of a positive attribute match for the individual species varies from zero to 0.81 (p_{Ak}^2 or p_{Bk}^2 from Table 1).

For example, consider again samples 7 and 8 from Table 3. The two by two table was given in Table 5 and had

$$(a,b,c,d) = (4,1,2,3)$$

$$(a+b) = 5, \quad (a+c) = 6$$

and

$$a/N = 4/10 = 0.4.$$

Both of these samples were from Population A of Table 1 by construction. The probabilities that two samples randomly selected from Population A will have species k represented in both samples is given by p_{Ak}^2 and are:

Species k	1	2	3	4	5	6	7	8	9	10
p_{Ak}^2	.25	.25	.01	.01	.25	.25	.81	.81	.25	0

The average of these probabilities is 0.289. This is the true (population) value estimated by a/N = 0.4. The fact that 0.289 implies we should expect about three positive matches and we get four for samples 7 and 8 is beside the point. The point is that 0.289 is 0.289 above the smallest true probability and 0.521 below the largest true probability. In a true contingency table the probability that both attributes will be present in

a sampled individual is constant for all individuals in the population, and under the assumption of independence of attributes this probability does not change from individual to individual as it does for the different species in the periphyton data.

The understanding of what is measured by these three indices is elucidated only by counterexample when the two by two contingency table model is entertained. Our two by two table should be considered as just a convenient way of summarizing the four possible kinds of matches and mismatches of two vectors on N binary attributes. Consequently, we return to the geometric model considered previously.

For the two by two table,

	Sample ℓ'		
Sample ℓ	1	0	
1	a	b	$a+b = N_\ell$
0	c	d	$c+d = N-N_{\ell'}$
	$a+c = N_{\ell'}$	$b+d = N-N_{\ell'}$	N

the marginal totals can be written as above in terms of the total number of species observed in each sample, N_ℓ and $N_{\ell'}$, and N the total number of different species in the study.

In terms of the N dimensional unit hypercube model, these coefficients (K_Y, K_{YC} and K_B) are functions of the angle $\theta_{\ell\ell'}$ between the two vectors, say A_ℓ and $A_{\ell'}$. This can be seen as follows. From the definition of the scalar (dot) product in vector algebra we have: (see, e.g., Schwartz, Green and Rutledge, 1960, p. 16)

$$\Delta_\ell \cdot \Delta_\ell = \left(\sum_{k=1}^{N} \delta_{\ell k}^2 \right)^{1/2} \left(\sum_{k=1}^{N} \delta_{\ell' k}^2 \right)^{1/2} \cos \theta_{\ell\ell'}$$

$$= (N_\ell N_{\ell'})^{1/2} \cos \theta_{\ell\ell'} .$$

Also,

$$\Delta_{\ell} \cdot \Delta_{\ell'} = \sum_{k=1}^{N} \delta_{\ell k} \delta_{\ell' k} = a,$$

so that

$$a = (N_{\ell} N_{\ell'})^{1/2} \cos \theta_{\ell\ell'} .$$

If we note that

$$Na - N_{\ell} N_{\ell'} = a^2 + ab + ac + ad - (a+c)(a+b) = ad - bc,$$

then by substitution

$$ad - bc = N[(N_{\ell} N_{\ell'})^{1/2} \cos \theta_{\ell\ell'}] - N_{\ell} N_{\ell'} .$$

It follows that the determinant is a function of the marginal totals and the cosine of the angle between the two vectors. Then

$$K_Y = \frac{N[(N_{\ell} N_{\ell'})^{1/2} \cos \theta_{\ell\ell'}] - N_{\ell} N_{\ell'}}{ad + bc}$$

and

$$K_B = \frac{N[(N_{\ell} N_{\ell'})^{1/2} \cos \theta_{\ell\ell'}] - N_{\ell} N_{\ell'}}{[N_{\ell} N_{\ell'} (N-N_{\ell})(N-N_{\ell'})]^{1/2}}$$

Kendall and Stuart (1973, Vol. II. p. 559) point out that

$$K_Y = \frac{2K_{YC}}{1+K_{YC}^2}$$

It follows that K_{YC} is also an angular coefficient.

These angular coefficients measure the separation between the samples as a function of the angle between the two vectors representing the samples in the N dimensional vector space. The relative magnitudes (lengths) of the vectors is not completely ignored in K_Y as it is in K_B, but both emphasize angular separation rather than magnitude of separation.

Absolute Difference Indices

The data for these indices, D_M, D_{MCD}, D_C and S_G, can be thought of as vectors in an N-dimensional space, but now the N axes are not restricted to define a unit hypercube. The axis for the k^{th} attribute ranges from zero to the largest value observed for the k^{th} attribute. Since there are N such attributes with the maximum value for one attribute being potentially orders of magnitude greater than for some other attribute, the resulting vector space may be far from homogeneous in directional magnitude. This was pointed out in the discussion of the calculation of these indices.

The sum of the absolute differences is the definition of the City Block distance between the two points defined by the sample vectors. This is exactly D_M. The Mean Character Distance is simply

$$D_{MCD} = D_M/N$$

which is the average absolute difference between species scores. This average is not meaningful, nor is the total, D_M, when there are order of magnitude differences in the species numerosities. It is recommended that the data be standardized if either D_M or D_{MCD} is used. This makes each axis comparable in length since all axes are transformed to attribute standard deviation units.

Lance and William's D_C normalizes each absolute difference by dividing the difference by the sum of the values which are differenced so that the resulting attribute similarity score is between zero and unity and the vector space becomes the unit hypersphere. Since the normalizing sum

changes for each attribute from pairwise comparison to pairwise comparison of vectors, D_C calculated for one pair of vectors is not generally comparable to D_C on another pair of vectors. For example, using data from Table 2, the similarity score for species 7 when samples 7 and 8 are compared is

$$|30-40|/70 = 10/70 = 0.143$$

but when samples 8 and 9 are compared it is

$$|40-160|/200 = 120/200 = 0.6.$$

In the first case, 70/70 is mapped onto unity, in the second unity is equivalent to 2001200.

Gower's S_G does not have this problem since the same normalizing factor, the attribute range, is used for a given attribute no matter what pair of sample vectors are being compared. For example, species 7 has a range of 300 (in Table 2) so that the attribute similarity score for samples 7 and 8 is

$$s_{7,8,7} = 1 - \frac{|30-40|}{300} = 1 - \frac{10}{300} = 1 - 0.0333 = 0.9667.$$

and for samples 8 and 9 it is

$$s_{8,9,7} = 1 - \frac{|40-160|}{300} = 1 - \frac{120}{300} = 1 - 0.4 = 0.6.$$

(Subtracting the "distance measure" from unity in Gower's S_G makes it a measure of closeness). For Gower's S_G the units for each axis are the same no matter what pairwise vector comparison is being made. However, normalizing by the range does make S_G sensitive to a single unusually large observation as, for example, species 2 in sample 10 of Table 2. The largest value $x_{10,2} = 8$ is four times the next largest value, $x_{9,2} = 2$, making the smallest possible similarity score be

$$1 - \frac{|0-2|}{8} = 1 - 0.25 = 0.75$$

for any comparison which does not include sample 10.

The first two of these indices, D_M and D_{MCD}, measure the City Block Distance, or the average City Block Distance, in the N-dimensional attribute space. In order to avoid a distorted attribute space, which unduly weights the axes with large ranges, the distortion should be removed by standardizing the attribute scores. The method of normalizing D_C causes an interval of 0.1 units to represent different degrees of separation in each attribute when different pairs of vectors are considered. Consequently, D_C should not be used when similarity indices are to be compared, nor should negative matches ($x_{\ell k} = x_{\ell' k} = 0$) be eliminated since this changes the dimensionality of the attribute space.

There are many factors which recommend Gower's S_G for use rather than any of the other similarity indices considered in this report. It is a measure of closeness. The ranging of the absolute difjerences is simpler than standardizing and causes the resulting attribute similarity scores to be between zero and unity. Averaging the similarity scores makes S_G lie between zero and unity. The units in the attribute space do not change from pair to pair of vectors allowing the resulting indices to be meaningfully compared.

Squared Difference Indices

The same N dimensional attribute space as for the Absolute Difference Indices is applicable for D_E, $D_{\overline{E}}$, D_{CD} and S_C, the indices based on squared differences. The only change is that vector separation is measured by straight line Euclidean distance rather than City Block distance. The problem of attribute range heterogeneity must be solved by normalizing or standardizing the attributes.

Only D_{CD} has attribute normalization built into its calculational formula. It uses the same ratio as D_C, the potentially negative sign being taken care of by squaring the ratio. Since the normalization factors vary from pair to pair, the same lack of an equal interval property as explained for D_C is a problem in the use of D_{CD}. However, dividing the sum of the squared normalized differences by N causes D_{CD} to lie between zero and unity.

Cattell's S_C is a normalized index, but the normalization is not done on the attribute similarity scores. The Euclidean distance is subtracted from the median value of the chi-square distribution in the numerator and the two quantities are added in the denominator of S_C. Such a quantity lies between -1 and +1, but its meaning is distorted unless the attributes have been previously standardized. But even standardization cannot correct the skewness of the original distributions, as is obvious from Table 7 (p. 38). Using the deviation of D_E from the median chi-square has meaning when

$$\sum_{k=1}^{N} (x_{\ell k} - x_{\ell' k})^2$$

is a sum of the squares of standard deviates from a (approximately) normal (Gaussian) frequency distribution. Standardizing the attribute value does not make the distribution of the difference between the resulting z values approach the normal distribution. The distributional theory involved requires concepts of mathematical statistics, which is beyond the scope of this paper.

The indices D_E or $D_{\overline{E}}$ based on standardized data should be used if one wants a Euclidean distance measure. Both Clark's D_{CD} and Cattell's S_C have theoretical problems which make their interpretation questionable.

Minimum/Maximum Indices

Levandowsky's S_L and Pinkham and Pearson's S_{PP} both suffer from having potentially different normalizing factors for different pairs of sample vectors. These indices may thus have different units when indices from different pairs of vectors are calculated, making comparison of indices calculated from the same data set invalid. In addition, Pinkham and Pearson's index allows for eliminating species which appear in neither sample for a particular calculation of S_{PP}, thus changing the dimensionality of the attribute space from pair to pair of vectors. This is necessary to avoid indeterminate zero divided by zero quotients in calculating the attribute similarity score. Their suggestion to avoid this problem by assigning such negative matches a value of one is consistent with their other algebra. However, it leads to a lower limit on the index directly calculatable from the number of such negative matches as

(no. of (0,0) comparisons)/N.

The index S_{PP} also assigns an attribute similarity score of unity to any exact numerical match so that relative numerosity from species to species has no impact; that is, 2/2 is the same as (1000/1000) although (1/2) = 0.5 and 1/1000 = 0.001. This is a type of equal interval failing that cannot be corrected by the usual standardization procedure since then the attribute ratios may be negative causing confusion in the interpretation of S_{PP} (and S_L) as pointed out above (p. 43).

The data may be formally adjusted for lack of equal interval results by categorization into a number of cells as was done on p. 44 and in Table 10. Categorization makes the attribute space homogeneous in the axes. The attribute ratio scores are then comparable (of the same order of magnitude) from species to species over all samples.

These two indices are related to the N-dimensional vector space as follows. A pair of vectors defines a new pair of vectors through the min, max operations. Call these vectors $M_{\prime\prime\prime}$ and $M_{\prime\prime\prime}$. Then Levandowsky's

S_L is based on the sum of the components of each of these two vectors. The sum of the components of $MN_{\ell\ell'}$ is the city block length of $MN_{\ell\ell'}$, and the sum of the components of $MX_{\ell\ell'}$ is its city block length. S_L is thus a function of the ratio of the city block length of $MN_{\ell\ell'}$ to the city block length of $MX_{\ell\ell'}$. If the two original vectors are identical then $x_{\ell k} = x_{\ell' k}$ and

$$\min(x_{\ell k}, x_{\ell' k}) = \max(x_{\ell k}, x_{\ell' k})$$

for all attributes, k, so that

$$MN_{\ell\ell'} = MX_{\ell\ell'}$$

and the ratio is unity making S_L equal to zero. The only way S_L can be unity is for the sum of the elements of $MN_{\ell\ell'}$ to be zero. This can only happen if $\min(x_{\ell k}, x_{\ell' k})$ is zero for all attributes, that is, at least one of $x_{\ell k}, x_{\ell' k}$ is zero for each attribute.

Pinkham and Pearson's S_{PP} is the average of the ratios

$$mn_{\ell\ell' k}/mx_{\ell\ell' k}, \quad k = 1, \ldots, N$$

where $mn_{\ell\ell', k}$ and $mx_{\ell\ell', k}$ are the components of $MN_{\ell\ell'}$ and $MX_{\ell\ell'}$. The ratio of the minimum attribute score to the maximum attribute score corresponds to making each axis of unit length and assigning an attribute score based on the fraction of this unit length accounted for by the minimum value. S_{PP} is thus the city block length of the vector in the unit hypersphere of dimension N divided by N, or the average proportional attribute separation.

The vector space model shows explicitly how both S_L and S_{PP} may be differently normed for different pairs of sample vectors.

"Pearson's Product Moment Correlation Coefficient"

The unsuitability of this algebraic manipulation for indicating similarity was amply discussed under the subsection describing the calculation of r,. Here we merely point out that the N dimensional algebra does

apply and that r, is the cosine of the angle between the two vectors determined by projecting the original vectors onto the equiangular line along the unit vector of dimension N and translating the vectors so determined to the origin so that they lie in the plane perpendicular to the equiangular line. The details are given by Anderson (1958, p. 49). The correlation coefficient is insensitive to the relative magnitudes of the attribute scores so that

$$r(X_\ell, X_{\ell'}) = r(X_\ell, tX_{\ell'})$$

where t is an arbitrary constant. This means that r, would be the same if $X_{\ell'}$ were reduced by, say, a factor of ten since the angle between X_ℓ and $X_{\ell'}$ is the same as the angle between X_ℓ and $tX_{\ell'}$.

Sample Fractions Indices

Johnson and Brinkhurst's P_J and Morisita's P_M are based on transforming each sample vector from the original counts or biomass observations to fractions (or percentages) by dividing each attribute value by the total of all attribute values in the vector. This operation makes each attribute vector X_ℓ correspond to a point Q_ℓ in the positive quarter of the N dimensional unit hypersphere. Then

$$P_J = \sum_k \min(q_{\ell k}, q_{\ell' k})$$

is just the city block length of the vector $MN'_{\ell\ell'}$ defined by taking the smallest fraction observed on each of the N axes for the pair of sample vectors under consideration. The unit length of each axis corresponds to the total x_ℓ, and so the fractions are strictly comparable only for two vectors which have the same total. For example, the vectors

$$X_\ell = (4,6,10), \quad X_{\ell'} = (8,12,20)$$

would have

$$Q_\ell = (.2,.3,.5), \quad Q_{\ell'} = (.2,.3,.5)$$

and so correspond to the same point in the unit hypersphere. For these two vectors Pj would be unity. If x_ℓ were (6,4,10) and $X_{\ell'}$ the same as above, then P_J would be 0.9. For Pj to be zero at least one of $x_{\ell k}$ or $x_{\ell' k}$ must be zero for all species. When both $x_{\ell k}$ and $x_{\ell' k}$ are zero for species k, $q_{\ell k} = q_{\ell' k} = 0$ which makes a negative match equivalent to an attribute comparison in which only one of $x_{\ell k}$ or $x_{\ell' k}$ is zero. But a value of zero for an attribute has an impact on the $q_{\ell k}$ for the other attributes in the sample. For example,

$$X_\ell = (4,6,10) \text{ and } X_{\ell'} = (4,6,0)$$

would have

$$Q_\ell = (.2,.3,.5) \text{ and } Q_{\ell'} = (.4,.6,0),$$

so that

$$P_J = 0.5.$$

This reflects another kind of lack of the equal interval property necessary for meaningful comparisons of indices. In the last example above

$$x_{\ell 1} = x_{\ell' 1} = 4$$

but

$$q_{\ell 1} = 0.2 \text{ and } q_{\ell' 1} = 0.4.$$

The problem is that different sample vectors may be mapped onto different subspaces of the N dimensional unit hypersphere when $x_{\ell k}$ is zero for some species (k).

Turning to Morisita's P_M we note that it is based on the same transformation of the N dimensional space into the positive quarter of the unit N dimensional hypersphere. The same lack of the equal interval property caused by vectors with different totals and vectors with different numbers of zero observations is encountered.

In terms of vector algebra

$$P'_M = \frac{2\sum_k q_{\ell k} q_{\ell' k}}{\sum_k q^2_{\ell k} + \sum_k q^2_{\ell' k}} = \frac{2Q_\ell \cdot Q_{\ell'}}{|Q_\ell|^2 + |Q_{\ell'}|^2} = \frac{2|Q_\ell||Q_{\ell'}|\cos Q_{\ell\ell'}}{|Q_\ell|^2 + |Q_{\ell'}|^2}$$

where the quantity in the numerator is twice the scalar (dot) product of the vectors of fractions Q_ℓ and $Q_{\ell'}$ and the denominator is the sum of the squared lengths of the two vectors. In Figure 4 two general vectors, lying in the plane determined by the vectors under consideration, are represented. The line segment OM is the perpendicular projection of Q_ℓ onto $Q_{\ell'}$ and its length is given by $|Q_\ell|\cos Q_{\ell\ell'}$. The vector $Q'_{\ell'}$ is the vector $Q_{\ell'}$ translated so that its initial point lies at the end point of Q_ℓ and its end point determines the point R. Then the vector $\overline{OR}$ is the vector sum $Q_\ell + Q_{\ell'}$. Note that

$$(Q_\ell + Q_{\ell'})^2 = (Q_\ell + Q_{\ell'})\cdot(Q_\ell + Q_{\ell'}) = |Q_\ell|^2 + |Q_{\ell'}|^2 + 2|Q_\ell||Q_{\ell'}|\cos\theta_{\ell\ell'},$$

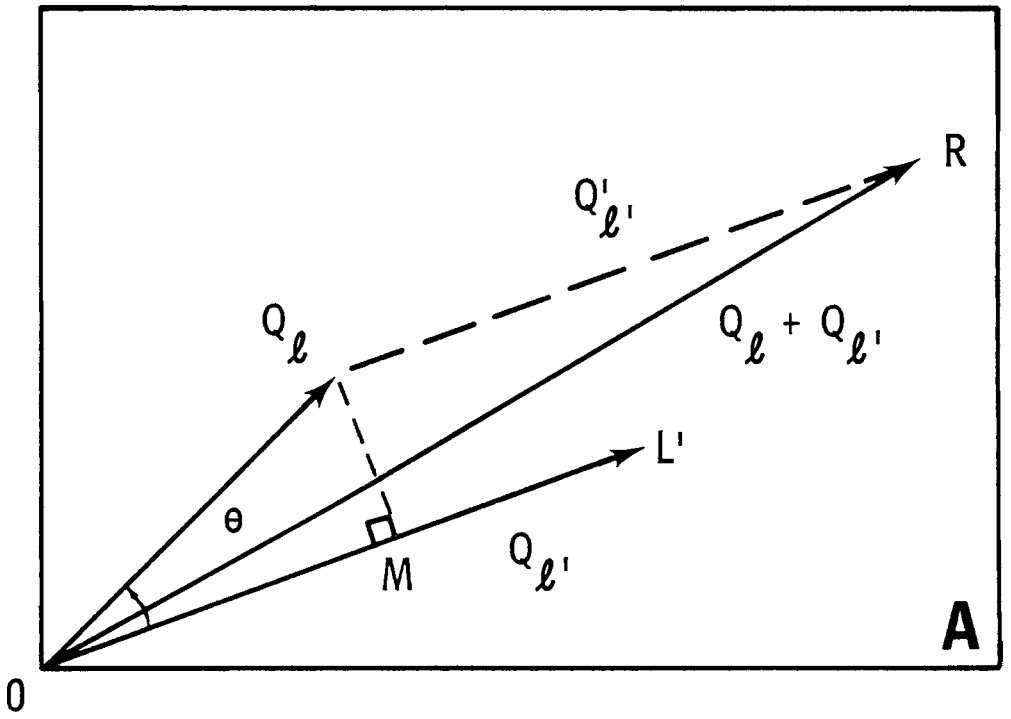
and

$$2|Q_\ell||Q_{\ell'}|\cos\theta_{\ell\ell} \leq |Q_\ell|^2 + |Q_{\ell'}|^2$$

equality holding when $Q_\ell = Q_{\ell'}$ and $\cos\theta = 1$ so that the vectors are identical in both magnitude and direction and lie along the vector $\overline{OR}$. In this case

$$P'_M = \frac{2|Q_\ell||Q_\ell|}{|Q_\ell|^2 + |Q_\ell|^2} = \frac{2|Q_\ell|^2}{2|Q_\ell|^2} = 1$$

If the vectors are perpendicular, $\theta_{\ell\ell'} = \pi/2$ and $\cos(\pi/2) = 0$ so that P'_M is zero regardless of the lengths of the vectors. Figure 4A shows the general case and Figure 4B the case when $|Q_\ell| = |Q_{\ell'}|$. Note that in



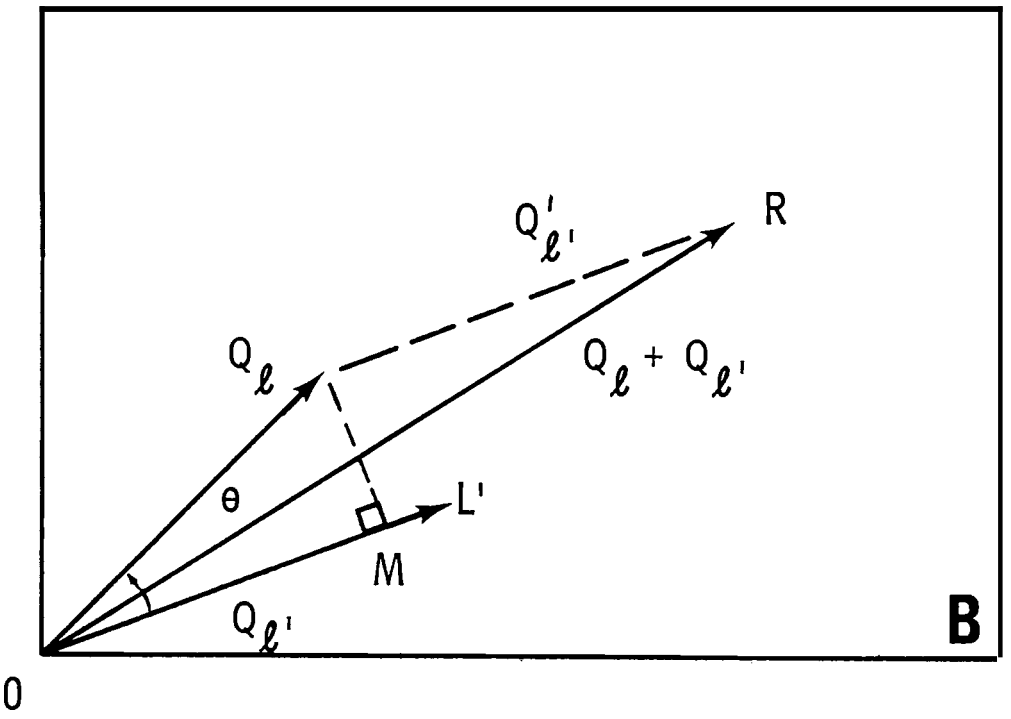


FIGURE 4. Illustrating the Vector Algebra Interpretation of P'_M

Figure 4B, the vector $\overline{OM}$ is still shorter than $Q_{\ell'}$, but that as the angle between Q_ℓ and $Q_{\ell'}$, is reduced $|Q_{\ell'}| - |\overline{OM}|$ approaches zero and in the limit the end points of Q_ℓ and $Q_{\ell'}$ would be the same point on $\overline{OR}$, then

$$2|\overline{OM}| = |Q_\ell + Q_{\ell'}| = |\overline{OR}|$$

and,

$$|Q_\ell + Q_{\ell'}|^2 = |Q_\ell|^2 + |Q_{\ell'}|^2 .$$

And

$$P'_M = \frac{2|Q_\ell||Q_{\ell'}|\cos\theta_{\ell\ell'}}{|Q_\ell|^2 + |Q_{\ell'}|^2}$$

$$= \frac{2|Q_\ell||\overline{OM}|}{|Q_\ell + Q_{\ell'}|^2},$$

so that the denominator is the squared length of the sum of the two vectors if they were colinear, i.e., if they had the same direction ($\theta_{\ell\ell'} = 0$). The numerator is twice the product of the length of one vector and the length of its projection on the other, that is, the length of Q_ℓ if it were projected onto $Q_{\ell'}$. Since

$$\overline{OM} = Q_{\ell'} - \overline{ML'}$$

where L' is the end point of $Q_{\ell'}$, we can write

$$P'_M = \frac{2|Q_\ell|}{|Q_\ell|^2 + |Q_{\ell'}|^2}\left[|Q_{\ell'}| - |\overline{ML'}|\right]$$

The first factor is twice the length of the first vector divided by the sum of the squares of the two vectors. The first factor thus depends on the relative lengths of the vectors alone. The second factor is dependent upon where the point M falls on $Q_{\ell\ell}$ and so is a function of the angle between Q_{ℓ} and $Q_{\ell\ell}$. Given a Q_{ℓ} of constant length, M will get closer to the origin as the angle increases. Morisita's P'_M thus measures the closeness of the two vectors in the N dimensional hypersphere as a fraction of what the squared length of $(Q_{\ell}+Q_{\ell'})$ would be under perfect agreement.

Using the same examples as for P_J we have

$$x_{\ell} = (4,6,10),\ x_{\ell'} = (8,12,20) \text{ and } x_{\ell''} = (6,4,10)$$

Then

$$P'_{M\ell\ell'} = \frac{2(.2^2+.3^2+.5^2)}{(.2^2+.3^2+.5^2)+(.2^2+.3^2+.5^2)} = \frac{2(.38)}{(.38)+(.38)} = 1$$

$$P'_{M\ell\ell''} = \frac{2[(.2)(.3)+(.3)(.2)+.5^2]}{0.76} = \frac{2(.37)}{.76} = \frac{.74}{.76} = 0.9737$$

For

$$x_{\ell} = (4,6,10) \text{ and } x_{\ell\ell} = (4,6,0)$$

$$P'_{M\ell\ell'} = \frac{2[(.2)(.4)+(.3)(.6)+(.5)(0)]}{(.38)+(.52)} = \frac{2(.26)}{.90} = \frac{.52}{.90} = 0.5778$$

Both P'_M and P_J are based on sample proportions so that the magnitude of the vectors in the original N-dimensional space are lost. That is, the fact that one sample has a total of, say, 5000 individuals and another only 100 is lost in the calculation of P_J and P'_M when the transformation to the unit hypersphere is made.

Goodall's "Probabilistic" Index

Goodall's S_p is based on using the entire set of sample vectors to determine the relative frequency with which the various matches for binary, classificatory, or ordered categorical data are observed. Goodall recommends that continuous data or counting data with a large range be made categorical by dividing the observed distribution into a convenient number of groups, and using the mean or median of each group as the value of the attribute for each individual included in it. (Goodall, 1966, p. 888) The relative frequencies observed in the total sample provide an empirical distribution function* which is used to order all pairwise comparisons for each attribute. The attribute similarity score for an observed pair of values is the one-complement of one, that is zero, if the attribute values are different, and of the cumulative relative frequency with which the match occurs if the attribute values are the same. Each possible pairwise sample vector comparison is then ordered by multiplying the N attribute similarity relative frequencies together and creating another empirical cumulative relative. frequency distribution. Vector similarity scores are assigned to an observed pair of vectors based on where the relative frequency score for the pair, calculated from the relative frequencies with which each attribute value was observed, falls on the vector empirical distribution.

---

*This "empirical distribution function" (e.d.f.) is not the same as the e.d.f. used in some non-parametric statistical procedures. The e.d.f. of the Komolgorov-Smirnov test for example is simply based on the cumulative relative frequencies with which a discrete set of values are observed. Goodall doesn't use the concept of an e.d.f. but his procedure is to construct something which looks like the classical e.d.f. but is based on treating the observed relative frequencies as if they were population probabilities, his use of carats over "estimated" values not really being carried to adequately interpreting what he has done. He claims his index "clears the way for numerical taxonomy controlled by specific significance levels" (1966, p. 897) which is eminently open to question. Our use of "e.d.f." here is merely to stress the fact that his results are empirical not probabilistic.

The explanation of how Goodall's index is constructed is complicated, but the calculation of S_p is complicated. Relating Goodall's index to the vector space model requires an NxL dimensional matrix sample space since his "probabilities" are based on the entire set of L vectors of dimension N. A truly probabilistic model would need to consider the set of all possible samples of L vectors. Goodall's simplification to a model involving only N dimensional vectors, not NxL dimensional matrices, makes his claim of providing "specific significance levels" (Goodall, p. 897) hollow. This is not to say that Goodall's S_p may not be useful in descriptive studies of similarity if one can bear the calculational labor. It does, more than any other similarity index considered in this paper, use the totality of the information contained in the entire set of L sample vectors. However, this may itself be a problem when there are really more than one population represented in the sample.

Looking at Table 1 (p. 8), Goodall's "estimate" of $p_{.k}$ would be 0.3 for K = 1,2,3, or 4 and 0.7 for K = 5,6,7 or 8 given the binary data of Table 3 (p.13). Ignorning species 9 and 10, his attribute similarity score would be as follows.

Observed Comparison	Species Set	
	k = 1,2,3,4	k = 5,6,7,8
(1,1)	.91	.42
(0,0)	.42	.91
(1,0) or (0,1)	0	0

If population A alone were considered then the attribute similarity scores would be:

Observed Comparison	Species Set		
	k = 1,2,5,6	k = 3,4	k = 7,8
(1,1)	.5	.99	.18
(0,0)	.5	.18	.99
(1,0) or (0,1)	0	0	0

These tables indicate that a positive match, when all L = 20 samples are considered gets an attribute comparison score of 0.91 for a (1,1) match for species 1,2,3 and 4. When only population A is considered, the match (1,1) gets a score of 0.5 for species 1 and 2 and .99 for species 3 and 4. Similarly, for the 20 sample exercise, the (1,1) match for species 7 and 8 is scored 0.42 but in population A it would be 0.18.

Obviously, Goodall's procedure has a smoothing effect making the estimate of the fraction of times a species match will be observed tend towards the average value of the population fractions when more than one population is represented in all the L samples. When these fractions are used to arrive at the attribute similarity scores for an observed pair of vectors of attribute values, the result is a vector of scores effectively based on the assumption that the two vectors come from the same population. The attribute similarity score is a function of the entire set of samples so that the pairwise comparisons of vectors from the whole set are not independent. The similarity score for each pairwise vector comparison is influenced by the pattern of attribute scores for the whole set of samples.

SUMMARY OF RESULTS

Table 15 summarizes 7 characteristics of each of the 25 indices plus an "Index of Choice" designating the author's preference. This table contains 200 bits of information and so must be quite cryptic. The cryptography will be explained for each column in turn.

PROPERTY MEASURED

Symbol	Means
D-1	Distance as measured in the N-dimensional unit hypercube.
D'-1	Same as D-1, but the dimensionality of the space may be less than N and vary from pair to pair of vectors.
D-N	Distance in the N dimensional vector space.
D"-1	Distance in the unit hypersphere but the dimensionality may be less than N and vary from pair to pair.
$\measuredangle$	The angle, rather than the magnitudes, of the vectors is the controlling property measured.
α	Mountford's K_I is equivalent to $1/\alpha$, where a is the "Index of Diversity."
RF	Goodall's S_p is a measure of the Eelative Frequency with which a less likely match would occur given the entire set of samples.

SPECIES NUMEROSITY

Symbol	Means
I	Ignored
U	Used
R	Made relative

TABLE 15. Some **Characteristics** of **the Indices**

Index	Property Measured	Species Numerosity	Attribute Space	Equal Interval	Range	Increases With	Standardize Data	Index of Choice
K_J	D'-1	I	V	N	0,1	C	N	X
K_D	D'-1	I	V	N	0,1	C	N	
K_W	D'-1	I	V	N	0,1	S	N	
S_L	D'-1	I	V	N	0,1	S	N	
K_I	α	I	V	N	0,u	C	N	
K_{SM}	D-1	I	F	Y	0,1	C	N	X
K_{RT}	D-1	I	F	Y	0,1	C	N	
K_H	D-1	I	F	Y	-1,1	C	N	
K_Y	4	I	F	N	-1,1	C	N	
K_{YC}	4	I	F	N	-1,1	C	N	
K_B	4	I	F	N	-1,1	C	N	
D_M	D-N	U	E	Y	0,∞	S	Y	
D_{MCD}	D-N	U	E	Y	0,∞	S	Y	
D_C	D"-1	R	E	N	0,N	S	N	
S_G	D"-1	R	E	Y	0,1	C	N	XX
D_E	D-N	U	E	Y	0,∞	S	Y	
$D_{\overline{E}}$	D-N	U	E	Y	0,∞	S	Y	X
D_{CD}	D"-1	R	E	N	0,1	S	N	
S_C	D$^{\pm}$1	U	E	N	-1,1	S	Y	
S_L	D"-1	U	E	N	0,1	S	C	
S_{PP}	D"-1	R	E	N	0,1	C	C	X
$r_?$	4	U	F	N	-1,1	C	Y	
P_J	D"-1	R	V	N	0,1	C	N	
P_M	D"-1	R	V	N	0,1	C	N	X
S_P	RF	R	F	Y	0,1	C	N	

ATTRIBUTE SPACE

Symbol	Means
V	The dimension of the vector space is variable from pair to pair of vectors.
F	The dimension is fixed at N for all pairs of vectors.
E	The dimension may be either fixed or variable depending on treatment of (0,0) attribute comparisons. If it is desired to compare similarity indices for different pairs of vectors the (0,0) matches should be included to maintain the same dimensionality base.

EQUAL INTERVAL

If the index maintains the equal interval property, it is coded Y for "Yes." If it fails to do so by: mapping different values onto 1 for unit hypersphere spaces, or normalizing attribute similarity scores by different values in either different attributes for the same pair of vectors or for different pairs of vectors, it is given an N for "No."

RANGE

This column gives the range of values the index can take on. It should be self explanatory, except for Mountford's K_I which has a range of 0, u indicating it is undefined for the case of perfect agreement.

INCREASES WITH

If the index attains its maximum when the two vectors are identical, it is considered a measure of closeness, coded C. If it attains its maximum when the two vectors are maximally different, it is considered a measure of separation and coded S.

STANDARDIZE DATA

Unless the index has a built in normalization, which corrects for the order of magnitude differences in the ranges of the axes in the vector

space, some attributes will be implicitly weighted. Each attribute value should be standardized by calculating the average and standard deviation of the values observed over all samples in the study for each attribute separately and subtracting the average from the observed value and dividing the result by the standard deviation. It is recommended that standardization be practiced for those indices which have a Y for "Yes" in this column. A "C" indicates that broad differences in ranges of attribute values can be eliminated by classifying each attribute into the same number of categories.

INDEX OF CHOICE

The author's preference in similarity indices, if he were forced somehow to use them, is given in the last column. Gower's S_G is preferred above all, when (0,0) matches are included in its calculation, because:

- each attribute is simply normed by the range for the attribute.
- it has a range of zero to one.
- it does not reduce the dimensionality of the attribute space if (0,0) matches are included.
- it maintains the equal interval property over all pairs in the sample since attribute similarities are normed by attribute range.
- it can be used for any measurement type.

Reasons for the other choices are as follows. Jaccard's K_J is preferred simply because it is simplest and the indices K_D, K_W and S_L (binary application) are just simple functions of K_J. However, K_J should be avoided in even such intuitive index comparison applications as cluster analysis since potentially gross differences in the underlying attribute space for each pairwise comparison removes the basis of comparability. Mountford's K_I requires assuming that the number of individuals per species can be characterized by a logarithmic series, in addition to the acceptance

of ignoring negative matches. The Simple Matching Coefficient, K_{SM} was selected for its simplicity since, again K_{RT} and K_H are simple functions of K_{SM}. It is to be preferred to K_J since the attribute space remains fixed from pair to pair of vectors. The angular coefficients, K_Y, K_{YC} and K_B are not recommended since their interpretation depends on being based on a two by two contingency table, which is not equivalent to our two by two table based on mutual presence of a number of different attributes. The indices based on City Block Distance, D_M, D_{MCD}, D_C and S_G are dominated in utility by S_G. If a distance measure is desired $D_{\bar{E}}$, the Average Euclidean Distance is recommended, provided the data are standardized first. Neither S_{PP} nor S_L is desirable because of the lack of the equal interval property, but S_{PP} is to be preferred either on categorized data or not, since it ratios the minimum and maximum for each attribute before summing. Categorization is definitely required if S_L is to be used. The use of the algebra for the Pearson Product Moment Correlation Coefficient on our data vectors can only be misleading, even if the attributes are standardized first. The relative frequency indices, P_J and P'_M make species numerosity relative to sample total so that gross changes in numbers of species can go undetected as long as proportions remain fairly constant. (This is also true of the angular indices.) Morisita's P_M is preferred because it makes more use of the data. Finally, Goodall's S_P smooths out the lack of similarity, if any, in the vectors before calculating the attribute similarity scores, and smooths again before combining these scores into his index. Besides, his index requires too much labor to understand and calculate, and once understood it is evident that he would do better using a (quite complicated) multinomial model.

LITERATURE CITED

Anderson, T. W. An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, New York, 1958.

Ball, G. H. A Comparison of Some Cluster-Seeking Techniques. Stanford Research Institute, RADC-TR-66-514, 1966.

Battelle Staff. A Monitoring Program on the Ecology of the Marine Environment of the Millstone Point Area. TR-14592. W. F. Clapp Lab., Duxbury, Mass, 1975.

Brown, R. T. and Moore, S. F. An Analysis of Exposure Panel Data Collected at Millstone Point, Connecticut. MIT, Cambridge, Mass, 1976. (DRAFT)

Campbell, N. R. Measurement and Calculation. Longmans, Green and Co., New York, 1928.

Cattell, R. B. rp and other coefficients of pattern similarity. Psychometrika 14:279-288, 1949.

Clark, P. J. An extension of the coefficient of divergence for use with multiple characters. Copeia 2:61-64, 1952.

Crovello, T. J. The effect of alteration of technique at two stages in a numerical taxonomic study. U. Kansas Sci. Bull. 47:761-786, 1968.

Czekkanowski, J. Zur differential diagnose der neandertalgruppe. Korespondenzblatt Deutsh Ges. Anthropol. Ethnol. Urgesh. 40:41-47, 1909.

Czekkanowski, J. "Coefficient of racial likeness" and "Jurchschnittliche differenz." Anthrop. Anz. 9:227-249, 1909.

Dice, L. R. Measures of the amount of ecological association between species. Ecology 26:297-302, 1945.

Fienberg, S. E. and J. P. Gilbert. The geometry of a two by two contingency table. JASA 65:694-701, 1970.

Fisher, R. A., A. S. Corbet and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. J. Animal Ecol. 12:42-58, 1943.

Goodall, D. W. A new similarity index based on probability. Biometrics 22: 882-907, 1966.

Goodall, D. W. The distribution of the matching coefficient. Biometrics 23: 647-656, 1967.

Gower, J. C. A general coefficient similarity and some of its properties. Biometrics 27:852-871, 1971a.

Haman, U. Merkmalbestand und verwardtschaftsbeziehungen der Farinosae. Willdenowia 2:639-768, 1961.

Horn, H. S. Measurement of "overlap" in comparative ecological studies. The Amer. Nat. 100:419-424, 1966.

Jaccard, P. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat. 44:223-270, 1908.

Jaccard, P. Lois de distribution florale dans la zone alpine. Bull. Soc. Vaud. Sci. Nat. 38:69-130, 1902.

Johson, M. G. and R. O. Brinkhurst. Association and species diversity in benthic micro-invertebrates of the Bay of Quinte and Lake Ontario. J. Fish. Res. Bd. Can. 28:1683, 1971.

Johnston, J. W. Similarity Indices II: The Power of Goodall's Significance Test for the Simple Matching Coefficient. BNWL-2152. Battelle-Northwest, Richland, WA, 1976.

Kendall, M. G. and A. Stuart. The Advanced Theory of Statistics. Charles Griffin and Co., Ltd., London, 1969, 1973, 1968 (3 vols).

Lance, C. N. and W. T. Williams. Mixed-data classificatory programs I. Agglomerative systems. Aust. Computer J. 6:15-20, 1967.

Levandowsky, M. An ordination of phytoplankton populations in ponds of varying salinity and temperature. Ecology 53:398-407, 1971.

Mahalanobis, P. C. On tests and measure of group divergence. J. Asiat. Soc. Beng. 26:541-588, 1930.

Maillfeur, A. Le coefficient generique de P. Jaccard et sa signification. Mem. Soc. Vaud. Sci. Nat. 3:113-183.

Morisita, M. Measuring of interspecific association and similarity between communities. Mem. Fac. Sci. Kyushu Univ. Ser. E. (Biol.) 3:65-80, 1960.

Mountford, M. D. An index of similarity and its application to classificatory problems, 43-50. In: Progress in Soil Zoology, P. W. Murphy (ed.), Butterworth, London, 1962.

Mountford, M. D. A test of differences between clusters; pp. 238-257. In: Statistical Ecology 3. G. P. Patil, E. C. Pielou, W. E. Waters (eds.), Penn. State U. Press, 1970.

Ostle, B. Statistics in Research. Iowa State U. Press, Ames, Iowa, 1963.

Pearson, K. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. Phil. Trans. A. 187:253-318, 1896.

Pinkham, C. F. and J. G. Pearson. Application of a new coefficient of similarity to pollution surveys. J. WCPF 48:717-723, 1976.

Rogers, D. J. and T. T. Tanimoto. A computer program for classifying plants. Science 132:1115-1118, 1960.

Sanghri, L. D. Comparison of genetical and morphological methods for a study of biological differences. Amer. J. Phys. Anthropol. 7:385-404, 1953.

Schwartz, M., S. Green and W. A. Rutledge. Vector Analysis. Harper and Bros., New York, 1960.

Siegel, S. Nonparametric Statistics. McGraw-Hill, New York, 1956.

Sokal, R. R. Distance as a measure of taxonomic similarity. Systematic Zool. 10: 70-79, 1961.

Sokal, R. R. and C. D. Michener. A statistical method for evaluating systematic relationships. U. Kansas Sci. Bull. 38:1409-1438, 1958.

Sokal, R. R. and P. H. A. Sneath. Numerical Taxonomy. W. H. Freeman and Co., 1973.

Sorenson, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. K. Danska Videnskab Selsk. Biol. Skr. 5(4):1-34, 1948.

Watson, L., W. T. Williams and G. W. Lance. Angiosperm taxonomy: a comparative study of some novel numerical techniques. J. Linn. Soc. (Bot.) 59: 491-501.

Williams, C. B. Jaccard's generic coefficient of floral community in relation to the logarithmic series and the index of diversity. Annals of Botany 13:53-58, 1949.

Yule, G. U. On the association of attributes in statistics. Phil. Trans. A. 194:257, 1900.

Yule, G. U. On the methods of measuring association between two attributes. J. R. Stat. Soc. 75:579-652, 1912.

REFERENCES

E. Ashby. The quantitative analysis of vegetation. Annals of Botany 69: 779-802, 1935.

E. W. Beals. Ordination: Mathematical elegance and ecological naviete. J. Ecol. 61:23-25, 1973.

J. Cairns, Jr. and R. L. Kaesler. Cluster analysis of fish in a portion of the Upper Potomac River. Trans. Amer. Fish. Soc. 100:750-756, 1971.

A. J. Cole (Ed.). Numerical Taxonomy (Colloq. Proc. U. St. Andrews, 1968) Academic Press, London, 1969.

L. C. Cole, The measurement of interspecific association. Ecology 30: 411-424, 1949.

E. Dahl. Some measures of uniformity in vegetation analysis. Ecology 41: 805-808, 1960.

H. G. Gauch, Jr., G. B. Chase and R. H. Whittaker. Ordination of vegetation samples by Gaussian species distribution. Ecology 55:1382-1390, 1974.

L. A. Goodman and W. H. Kruskal. Measures of association for cross classification. I. II, III. JASA 49:732; 54:123; and 58:310.

J. C. Gower. A comparison of some methods of cluster analysis. Biometrics 23:623-637.

J. C. Gower. A note on Burnaby's character-weighted similarity coefficient. Math. Geol. 2:39-45, 1970.

J. C. Gower. Measures of taxonomic distance and their analysis. In: Symposium on Assessment of Population Affinities. Oxford University Press, 1971.

P. Greig-Smith. Quantitative Plant Ecology, 2nd ed. Butterworth, London, 256 pp., 1964.

P. Greig-Smith, M. P. Austin and T. C. Whitmore. The application of quantitative methods to vegetation survey I. J. Ecol. 55:483-503, 1967.

J. A. Hendrickson and R. R. Sokal. A numerical taxonomic study of the genus Psorophora (Diptera:culicidae). Ann. Entomol. Soc. Amer. 61: 385-392, 1968.

D. M. Jackson and L. J. White. The weakening of taxonomic inferences by homological error. Mathematical Biosciences 10:63-89.

K. Matusita. Decision rule based on distance for the classification problem. Ann. Inst. Stat. Math. 8:67-77, 1957.

L. Orloci. Geometric models in ecology. I. The theory and application of some ordination methods. J. Ecol. 54:193-215, 1960.

L. Orloci. Data centering: a review and evaluation with references to component analysis. Syst. Zool. 16:208-212, 1967.

L. Orloci. An agglomerative method for classification of plant communities. J. Ecol. 55:193-206, 1967.

G. P. Patil, E. C. Pielou and W. E. Waters (eds.). Statistical Ecology. Penn. State Statistics Series, Penn. State U. Press, 1971.

K. Pearson. On the coefficient of racial likeness. Biometrika 18:105-117, 1926.

E. C. Pielou. An Introduction to Mathematical Ecology. Wiley-Interscience, New York, 1969.

E. H. Simpson. Measurement of diversity. Nature 163:168, 1949.

R. R. Sokal and F. J. Rohlf. Biometry, The Principal and Practice of Statistics in Biological Research. W. H. Freeman and Co., San Francisco, 776 pp., 1969.

W. T. Williams and M. B. Dale. Fundamental problems in numerical taxonomy. Ad. Bot. Res. 2:35-68.

W. T. Williams, J. M. Lambert and G. N. Lance. Multivariate methods in plant ecology. V. Similarity analysis and information analysis. J. Ecol. 54:427-445, 1965.

APPENDIX A

APPENDIX A

Millstone Data Exemplifying Type of Data Encountered in Similarity Studies

This Appendix contains an example of the type of data frequently subjected to a similarity index analysis. Table A-1 lists the basic monthly data for two years at Site FN. The species are listed in the order of most frequently observed to not observed in the set of 96 samples (4 sites with 24 monthly samples for each site). The percentage of the 96 samples in which each species was observed is given in the last column of Table A-1 headed 0/0 OBS. The first three columns contain species identification codes. The alphabetic codes, under CODE, are the same as on the species list given in Table A-3. (The SEQ NO. and COMB. numerical codes were for computer use.) The monthly data are given in the columns headed by the abbreviated names of the month: a number followed by a decimal point indicating counting data and a number with two digits after the decimal point indicating percent coverage data.

Table A-2 gives the percentage of samples in which the species were observed broken down for each year at each site. For each site the columns headed 70 and 72 give the percentages for the respective years (denominator of 12) and the column headed SUM, for both years combined (denominator of 24). An exception occurs for Site GN, 1972 for which the March and April panels were missing. The last 4 columns pertain to the total for all 94 samples present, 48 samples in 1970 and 46 samples in 1972.

It is interesting to note that only 97 species were observed in these 94 samples, implying an additional 106 species were observed in the samples not included in this two year selection.

TABLE A.1 Basic Data for Site FN Year 970

SEQ NO.	CODE	COMB.	JAN	FEB	MAR	APR	MAY	JUNE	JULY	AUG	SEPT	OCT	NOV	DEC	% OBS.
184	LIMG	12158	3000.	4300.	2900.	2800.	3000.	3 50.	3000.	3400.	4000.	2100.	2050.	840.	100.0
178	BALE	12144	.55	.40	.15	.20	.20	.25	.30	.06	.09	.40	.20	.30	89.4
82	CRYP	8031	.01	0.00	0.00	.02	.03	0.00	.02	.20	.30	.07	.07	.03	80.9
162	CORC	12051	20.	0.	50.	100.	30.	23.	40.	30.	30.	40.	160.	60.	74.5
161	CHET	12041	50.	1 0.	45.	50.	100.	30.	0.	0.	0.	180.	490.	15.	73.4
127	MUDW	9229	.01	.06	.05	.01	.03	.01	.03	.03	.03	.02	.01	0.00	72.3
69	SERP	5061	.01	.01	.02	.20	.25	.05	.01	.01	.02	.10	.05	.01	71.3
198	BOTS	14031	0.00	0.00	.01	.01	.03	.01	.05	0.00	0.00	.02	.01	.01	71.3
149	MYTE	11141	.01	.01	.02	.01	0.00	.01	0.00	0.00	0.00	.01	.02	.03	70.2
152	TERN	11162	.15	.02	.01	.05	0.00	.10	.20	.85	.30	.01	.01	.15	70.2
177	BALC	12143	.05	.05	.45	.20	.10	0.00	0.00	.04	.01	.05	.30	.15	70.2
5	CODD	1041	0.00	0.00	.01	.01	0.00	0.00	.01	.04	.06	.15	.01	.01	57.4
12	ULVL	1061	.01	.01	.07	.02	.05	.01	0.00	0.00	0.00	0.00	.01	.01	57.4
58	HALB	4021	.05	.02	.06	.12	.10	0.00	0.00	0.00	0.00	.01	.01	.01	57.4
105	LEPS	9101	0.	0.	2.	2.	3.	4.	0.	0.	0.	4.	5.	0.	46.8
182	LIMT	12152	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	46.8
183	LIMU	12153	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	45.7
181	LIML	12151	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	38.3
89	BUGS	8081	0.00	0 00	0.00	0.00	0.00	0 00	.50	.50	70	.05	.01	0.00	37.2
110	NERV	9123	2.	1.	0.	0.	0.	1.	0.	7.	8.	3.	1.	3.	33.0
172	MICX	12118	35.	9.	16.	30.	3.	0.	20.	25.	0.	35.	30.	0.	30.9
68	OBEX	5058	0.00	0 00	0.00	0.00	0.00	0 00	0.00	0.00	0 00	0.00	0.00	0.00	28.7
123	SERW	9199	0.	0.	0.	1.	1.	3.	3.	0.	0.	2.	0.	0.	28.7
132	CREF	11031	0.	0.	1.	0.	0.	0.	0.	0.	1.	1.	2.	8.	21.3
48	LAMA	3051	.01	0.00	0.00	0.00	0.00	.02	0.00	0.00	.01	.01	0.00	0.00	20.2
21	CERR	2052	0.00	0.00	0.00	.01	.01	0.00	0.00	0.00	.01	0.00	.01	.01	18.1
90	BUGT	8082	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	18.1
22	CERY	2058	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	17.0
56	GRAT	4011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	14.9
4	CLAX	1038	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	13.8
37	POLX	2148	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.7
176	BALB	12142	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	.10	11.7
199	CTOI	14041	0.00	0.00	.01	0.00	0.00	0.00	0.00	.01	.01	0.00	0.00	.01	11.7
8	ENTI	1053	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.6
158	CAPG	12031	0.	0.	0.	0.	0.	0.	0.	15.	25.	8.	0.	0.	10.6
1	BRYP	1011	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	9.6
200	MOLC	14051	0.00	0.00	0.00	0.00	0.00	0.00	.1	.01	.01	0.00	0.00	0.00	9.6
7	ENTC	1052	0.00	0.00	0.00	0.00	0.00	0.00	.0	.01	.01	0.00	0.00	0.00	8.5
141	COLL	11071	0.	0.	0.	3.	0.	0.	0	27.	40.	0.	0.	0.	8.5
11	ENTX	1058	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	6.4
148	MODM	11131	0.	0.	0.	0.	0.	0.	0	0.	0.	1.	0.	0.	6.4
41	RHOP	2171	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	5.3
75	STYE	6011	0.	0.	0.	0.	0.	0.	0	0.	0.	0.	0.	0.	5.3
83	ELEC	8041	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	5.3

TABLE A.1 Basic Data for Site FN Year 1970 (Cont'd)

111	NERX	9128	0.	0.	0.	0.	0.	0.	1.	0.	0.	0.	0.	0.	5.3
179	BALI	12145	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.3
62	LEUX	4038	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	.01	4.3
121	SABF	9189	0.	0.	0.	0.	0.	0.	0.	0.	2.	0.	0.	0.	4.3
143	UROC	11091	0.	0.	0.	0.	0.	0.	0.	0.	0.	1.	0.	0.	4.3
23	CHAP	2061	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	.01	0.00	0.00	3.2
29	GRAF	2102	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.2
125	TERF	9209	0.	0.	0.	0.	0.	1.	0.	0.	0.	0.	0.	0.	3.2
133	CREP	11032	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.2
160	CAPF	12039	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	80.	3.2
188	PANH	12191	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	2.	3.2
3	CHAX	1028	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	.01	0.00	0.00	2.1
15	AGAT	2021	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.1
26	CYSP	2081	0.00	0.00	.01	0.00	0.00	0.00	0.00	0.00	.01	0.00	0.00	0.00	2.1
57	GRAX	4018	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.1
70	SERT	5062	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.1
72	METD	5081	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.1
85	ELEM	8043	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.1
86	SCHU	8051	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.1
126	SPIW	9218	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.1
138	LITO	11061	0.	0.	0.	0.	0.	0.	0.	50.	0.	0.	0.	0.	2.1
180	BALX	12148	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.1
10	ENTP	1055	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
14	ACHX	2018	0.00	0.00	0.00	0.00	.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
18	CALB	2042	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
25	CHOX	2078	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
27	DAYP	2091	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
39	PORX	2158	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
40	RHOS	2161	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
42	RHOX	2178	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
43	DESV	3011	0.00	0.00	0.00	.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
54	PYLL	3091	0.00	0.00	.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
66	CAMX	5038	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
74	UNIH	5099	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
78	LEPF	6029	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
84	ELEP	8042	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
88	BOWG	8071	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	.01	0.00	1.1
91	BUGX	8088	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
98	EULV	9051	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
109	NERS	9122	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
112	NEPX	9138	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
115	PHYX	9158	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
128	NERM	9231	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	1.1
130	COLA	11011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
142	THAL	11081	0.	0.	0.	0.	0.	0.	2.	0.	0.	0.	0.	0.	1.1
145	AN08	11111	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	1.1
156	AMPX	12028	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
186	CARM	12171	0.	0.	0.	0.	0.	0.	0.	1.	0.	0.	0.	0.	1.1
187	EURD	12181	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
189	OECF	12209	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
191	IDOP	12212	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
193	TANC	12231	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.1
203	STYP	14061	0.00	0.00	0.00	0.00	0.00	0.00	.01	0.00	0.00	0.00	0.00	0.00	1.1
2	CHAA	1021	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
6	ENTA	1051	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
9	ENTL	1054	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
13	CHLF	1079	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
16	ANTX	2038	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
17	CALI	2041	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
19	CALX	2048	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0

TABLE A.1 Basic Data for Site FN Year 1970 (Cont'd)

20	CERO	2051	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
24	CHOR	2071	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
		2101	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
30	GRAC	2103	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
31	GRIA	2111	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
32	HERT	2121	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
33	LOMA	2131	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
34	LOMX	2138	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
35	POLI	2141	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
36	POLN	2142	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
38	PORU	2151	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
44	ECTX	3028	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
45	ELAX	3038	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
46	FUCE	3041	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
47	FUCX	3048	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
49	LAMX	3058	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
50	PEIL	3061	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
51	PHYA	3071	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
52	PUNL	3081	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
53	PUNX	3088	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
55	PYLX	3098	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
59	HALP	4022	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
60	HALX	4028	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
61	LEUA	4031	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
63	RENF	4049	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
64	UNIA	5019	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
65	ANTU	5029	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
67	DIAL	5041	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
71	TUBX	5078	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
73	METS	5082	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
76	LEPA	6021	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
77	LEPX	6028	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
79	RHYP	7019	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
80	CALA	8011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
81	CRIE	8021	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
87	TEGU	8061	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
92	AMPH	9018	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
93	AMPF	9019	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
94	CAPI	9029	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
95	CIRG	9031	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
96	CIRF	9039	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
97	EUCR	9041	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
99	EULX	9058	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
100	EUMX	9068	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
101	GLYF	9079	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
102	GLYX	9081	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
103	HYDD	9091	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
104	HYDX	9098	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
106	MARS	9111	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
107	MARX	9118	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
108	NERP	9121	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
113	NOTL	9141	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
114	PHYA	9151	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
116	PHYF	9159	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
117	PODO	9161	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
118	POLC	9171	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
119	POLF	9179	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
121	SARM	9181	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
122	SERV	9191	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
124	TERL	9201	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0

TABLE A.1 Basic Data for Site FN Year 1970 (Cont'd)

129	SIPX	10011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
131	CERG	11021	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
134	CREX	11038	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
135	HERX	11048	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
136	ILVO	11051	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
137	ILYX	11058	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
139	LITS	11062	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
140	LITX	11068	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
144	UNIN	11109	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
146	ANOX	11118	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
147	CRAV	11121	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
150	SAXA	11151	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
151	TERR	11161	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
153	TERE	11169	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
154	AEGL	12011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
155	AMPR	12021	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
157	AMPE	12029	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
159	CAPX	12038	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
163	ELAL	12061	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
164	GAMA	12071	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
165	GAML	12072	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
166	GAMX	12078	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
167	GAMF	12079	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
168	GRUC	12081	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
169	JASA	12091	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
170	MELD	12101	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
171	MELN	12102	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
173	UNIC	12129	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
174	UNCI	12131	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
175	BALA	12141	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
185	BRAX	12169	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
190	IDOB	12211	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
192	JAEM	12221	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
194	ASTF	13019	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
195	ASTO	13011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
196	ASCX	14018	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
197	AMAX	14028	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
201	MOLM	14052	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
202	MOLX	14058	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
TOTAL SPECIES OBSERVED			19/35	33	36	39	32	29	38	38	40	46	42	43	
O/O SPECIES OBSERVED			17.2	16.3	17.7	19.2	15.8	14.3	18.7	18.7	19.7	22.7	20.7	21.2	

TABLE A.1 Basic Data for Site FN **Year** 1972

sen NO.	CODE	COMB.	JAN	FEB	MAR	APR	MAY	JUNE	JULY	AUG	SEPT	OCT	NOV	DEC	0/0 OBS.
184	LEMG	12158	5800.	5400.	5900.	7600.	6800.	5700.	2200.	2050.	2700.	2900.	4000.	1350.	100.0
17A	BALE	12144	.01	.15	.03	.05	.02	0.00	.01	.05	.20	.55	.40	.40	89.4
82	CRYP	8031	n. 00	0.00	0.00	0.00	.01		.02	.50	.20	.01	.02	.01	80.9
162	CORC	12051	0.	0.	0.	0.	0.	0.	80.	200.	200.	300.	100.	200.	74.5
161	CHET	12041	200.	2000.	800.	100.	0.	0.	120.	400.	0.	280.	2000.	200.	73.4
127	MUDW	9229	.01	0.00	0.00	0.00	.01	.01	.01	.01	.01	.01	.01	.01	72.3
69	SERP	5061	.01	.01	.01	.01	.01	.01	.01	0.00	0.00	0.00	0.00	0.00	71.3
198	BOTS	14031	.01	.01	.01	.01	.18	.01	0.00	0.00	0.00	0.00	0.00	.01	71.3
149	MYTE	11141	.01	.01	0.00	0.00	0.00	0.00	.01	.01	.01	.04	.04	.01	70.2
152	TERN	11162	.01	.06	.09	.01	.50	.80	.60	.30	.01	.01	0.00	0.00	70.2
177	BALC	12143	.40	.36	.20	.35	.40		0.00	0.00	.06	.05	.15	.50	70.2
5	CODD	1041	.01	.01	.01	.01	.01	0.00	.01	.02	.01	.01	.01	.01	57.4
12	ULVL	1061	.01	.01	0.00	.01	.01	0.00	.01	0.00	.01	.01	.01	.01	57.4
58	HALB	4021	.01	.01	0.00	.03	.01	.01	0.00	0.00	0.00	0.00	.06	.01	57.4
105	LEPS	9101	1.	2.	1.	1.	0.	0.	0.	0.	0.	2.	1.	2.	46.8
182	LIMT	12152	2200.	1500.	2900.	4000.	3800.	4800.	1550.	2125.	1500.	2000.	1800.	825.	46.8
183	LIMU	12153	50.	80.	40.	200.	390.	90.	100.	225.	200.	103.	60.	45.	45.7
181	LIML	12151	100.	80.	100.	100.	80.	290.	350.	inn.	200.	10n.	100.	130.	38.3
89	BUGS	8081	0.00	0.00	0.00	0.00	0.00	0.00	.05	.30	.30	0.00	0.00	0.00	37.2
110	NERV	9123	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	2.	0.	33.0
172	MICX	12118	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	30.9
68	OBEX	5058	0.00	0.00	0.00	.01	.01	.01	.03	.01	.01	.01	.01	.01	28.7
123	SERW	9199	0.	0.	0.	0.	0.	0.	4.	1.	0.	0.	0.	0.	28.7
132	CREF	11031	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	1.	21.3
48	LAMA	3051	0.00	.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.2
21	CERR	2052	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	18.1
90	BUGT	8082	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	18.1
22	CERX	2058	0.00	0.00	0.00	0.00	0.00	0.00	.01	0.00	0.00	0.00	.01	.01	17.0
56	GRAT	4011	.07	.03	.02	.06	.13	0.00	.01	0.00	0.00	0.00	0.00	0.00	14.9
4	CLAX	1038	.01	.01	.01	0.00	0.00	0.00	.01	.01	.01	.01	0.00	.01	13.8
37	POLX	2148	0.00	.01	0.00	0.00	0.00	0.00	.01	.01	0.00	.01	.01	0.00	11.7
176	BALB	12142	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.7
199	CIOI	14041	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.7
8	ENTI	1053	0.00	0.00	0.00	0.00	0.00	.01	0.00	0.00	0.00	0.00	0.00	0.00	10.6
158	CAPG	12031	0.	0.	0.	0.	0.	0.	0.	0.	0.	100.	50.	0.	10.6
1	BRVP	1011	0.00	0.00	.01	.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	.01	9.6
200	MOLC	14051	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	.01	0.00	0.00	0.00	9.6
7	ENTC	1052	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.5
141	COLL	11071	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	8.5
11	ENTX	1058	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	.01	0.00	0.00	0.00	6.4
148	MODM	11131	1.	1.	0.	0.	0.	1.	0.	0.	0.	0.	0.	0.	6.4
41	RHOP	2171	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.3
75	STYE	6011	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	1.	3.	5.3
83	ELEC	8041	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	.01	0.00	0.00	0.00	5.3

TABLE A.1 Basic Data for Site FN Year 1972

20	CERO	2051	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
24	CHOR	2071	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
28	GRAC	2101	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
30	GRAS	2103	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
31	GRIA	2111	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
32	HERT	2121	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
33	LOMB	2131	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
34	LOMX	2138	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
35	POLI	2141	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
36	POLN	2142	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
38	PORU	2151	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
40	ECTX	3028	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
45	ELAX	3038	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
46	FUCE	3041	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
47	FUCX	3048	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
49	LAMX	3058	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
50	PETA	3061	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
51	PHYR	3071	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
52	PUNL	3081	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
53	PUNX	3088	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
55	PYLX	3098	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
59	HALP	4022	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
60	HALX	4028	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
61	LEUR	4031	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
63	RENF	4049	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
64	UNIA	5019	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
65	ANTU	5029	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
67	DIAL	5041	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
71	TUBX	5078	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
73	METS	5082	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
76	LEPA	6021	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
77	LEPX	6028	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
79	RHYP	7019	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
80	CALA	8011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
81	CRIE	8021	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
87	TEGU	8061	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
92	AMPH	9018	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
93	AMPF	9019	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
94	CAPI	9029	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
95	CIRG	9031	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
96	CINF	9039	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
97	EUCR	9041	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
99	EULX	9058	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
100	EUMX	9068	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
101	GLYP	9079	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
102	HART	9081	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
103	HYDD	9091	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
104	HYDX	9098	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
106	MARS	9111	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
107	MARX	9118	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
10A	NERB	9121	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
113	NOTL	9141	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
114	PHYA	9151	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
116	PHYF	9159	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
117	PODO	9161	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
11A	POLC	9171	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
119	POLF	9179	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
120	SARM	9181	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
122	SERV	9191	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
124	TERL	9201	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0

TABLE A.1 Basic Data for Site FN Year 1972

129	SIPx	10011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
131	CERG	11021	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
134	CREx	11038	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
135	HERx	11048	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
136	ILYO	11051	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
137	ILYx	11058	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
139	LITS	11062	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
140	LITx	11068	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
144	UNIN	11109	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
146	ANOx	11118	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
147	CRAV	11121	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
150	SAXA	11151	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
151	TERB	11161	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
153	TERE	11169	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
154	AEGL	12011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
155	AMPR	12021	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
157	AMPF	12029	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
159	CAPx	12038	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
163	ELAL	12061	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
164	GAMA	12071	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
165	GAML	12072	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
166	GAMx	12078	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
167	GAMF	12079	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
168	GRUC	12081	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
169	JASA	12091	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
170	MELD	12101	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
171	MELN	12102	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
173	UNIC	12129	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
174	UNCI	12131	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
175	BALA	12141	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
185	BRAx	12169	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
190	IDUB	12211	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
192	JAEM	12221	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
194	ASTF	13019	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
195	ASTO	13011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
196	ASCx	14018	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
197	AMAx	14028	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
201	MOLM	14052	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
202	MOLx	14058	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
TOTAL SPECIES OBSERVED			16	13	19	21	15	14	14	18	19	24	20	18	
0/0 SPECIES OBSERVED			0.9	6.4	9.4	10.3	7.4	6.9	6.9	8.9	9.4	11.8	9.9	8.9	

TABLE A.2 Percentage of Samples in Which Species Were Observed

SEQ NO.	CODE	COMB.	SITE FN 70	SITE FN 72	SITE FN SUM	SITE WP 70	SITE WP 72	SITE WP SUM	SITE MH 70	SITE MH 72	SITE MH SUM	SITE GN 70	SITE GN 72 Ma-Apr Mean	SITE GN SUM	TOTAL OF ALL SITES 70	TOTAL OF ALL SITES 72	TOTAL OF ALL SITES SUM	N
184	LIMG	12158	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94
178	BALE	12144	100.0	91.7	95.8	100.0	100.0	100.0	100.0	100.0	100.0	83.3	30.0	59.1	95.8	82.6	89.4	84
82	CRYP	8031	75.0	66.7	70.8	83.3	50.0	66.7	100.0	100.0	100.0	100.0	70.0	86.4	89.6	71.7	80.9	76
162	CORC	12051	100.0	50.0	75.0	100.0	50.0	75.0	91.7	50.0	70.8	91.7	60.0	77.3	95.8	52.2	74.5	70
161	CHET	12041	75.0	75.0	75.0	58.3	75.0	66.7	100.0	91.7	95.8	66.7	40.0	54.5	75.0	71.7	73.4	69
127	MUDW	9229	91.7	75.0	83.3	100.0	83.3	91.7	91.7	50.0	70.8	50.0	30.0	40.9	83.3	60.9	72.3	68
69	SERP	5061	100.0	58.3	79.2	91.7	58.3	75.0	100.0	41.7	70.8	91.7	20.0	59.1	95.8	45.7	71.3	67
198	BOTS	14031	66.7	58.3	62.5	83.3	75.0	79.2	83.3	83.3	83.3	75.0	40.0	59.1	77.1	65.2	71.3	67
149	MYTE	11141	66.7	66.7	66.7	100.0	100.0	100.0	91.7	75.0	83.3	25.0	30.0	27.3	70.8	69.6	70.2	66
152	TERN	11162	91.7	83.3	87.5	100.0	91.7	95.8	41.7	8.3	25.0	91.7	50.0	72.7	81.3	58.7	70.2	66
177	BALC	12143	83.3	83.3	83.3	66.7	100.0	83.3	50.0	25.0	37.5	66.7	90.0	77.3	66.7	73.9	70.2	66
5	CODD	1041	66.7	91.7	79.2	66.7	50.0	58.3	75.0	75.0	75.0	8.3	20.0	13.6	54.2	60.9	57.4	54
12	ULVL	1061	66.7	75.0	70.8	83.3	50.0	66.7	25.0	25.0	25.0	66.7	70.0	68.2	60.4	54.3	57.4	54
58	HALB	4021	66.7	58.3	62.5	58.3	83.3	70.8	66.7	50.0	58.3	41.7	30.0	36.4	58.3	56.5	57.4	54
105	LEPS	9101	50.0	58.3	54.2	58.3	41.7	50.0	33.3	41.7	37.5	41.7	50.0	45.5	45.8	47.8	46.8	44
182	LIMT	12152	0.0	100.0	50.0	0.0	100.0	50.0	0.0	100.0	50.0	0.0	80.0	36.4	0.0	95.7	46.8	44
183	LIMU	12153	0.0	100.0	50.0	0.0	100.0	50.0	0.0	100.0	50.0	0.0	70.0	31.8	0.0	93.5	45.7	43
181	LIML	12151	0.0	100.0	50.0	0.0	100.0	50.0	0.0	16.7	8.3	0.0	100.0	45.5	0.0	78.3	38.3	36
89	BUGS	8081	41.7	25.0	33.3	50.0	16.7	33.3	58.3	41.7	50.0	50.0	10.0	31.8	50.0	23.9	37.2	35
110	NERV	9123	66.7	8.3	37.5	91.7	16.7	54.2	41.7	0.0	20.8	33.3	0.0	18.2	58.3	6.5	33.0	31
172	MICX	12118	75.0	0.0	37.5	58.3	0.0	29.2	33.3	0.0	16.7	75.0	0.0	40.9	60.4	0.0	30.9	29
68	OBEX	5058	0.0	75.0	37.5	0.0	75.0	37.5	0.0	75.0	37.5	0.0	0.0	0.0	0.0	58.7	28.7	27
123	SERW	9199	41.7	16.7	29.2	41.7	41.7	41.7	16.7	16.7	16.7	16.7	40.0	27.3	29.2	28.3	28.7	27
132	CREF	11031	41.7	8.3	25.0	41.7	25.0	33.3	8.3	8.3	8.3	16.7	20.0	18.2	27.1	15.2	21.3	20
48	LAMA	3051	33.3	8.3	20.8	41.7	66.7	54.2	0.0	0.0	0.0	8.3	0.0	4.5	20.8	19.6	20.2	19
21	CERR	2052	41.7	0.0	20.8	58.3	0.0	29.2	33.3	0.0	16.7	8.3	0.0	4.5	35.4	0.0	18.1	17
90	BUGT	8082	0.0	8.3	4.2	41.7	33.3	37.5	8.3	0.0	4.2	16.7	40.0	27.3	16.7	19.6	18.1	17
22	CERX	2058	0.0	0	12.5	0.0	33.3	16.7	0.0	25.0	12.5	0.0	60.0	27.3	0.0	34.8	17.0	16
56	GRAI	4011	0.0	50.0	25.0	0.0	41.7	20.8	0.0	16.7	8.3	0.0	10.0	4.5	0.0	30.4	14.9	14
4	CLAX	1038	0.0	66.7	33.3	0.0	16.7	8.3	8.3	8.3	8.3	0.0	10.0	4.5	2.1	26.1	13.8	13
37	POLX	2148	0.0	41.7	20.8	0.0	41.7	20.8	0.0	8.3	4.2	0.0	0.0	0.0	0.0	23.9	11.7	11
175	BALB	12142	8.3	0.0	4.2	16.7	0.0	8.3	25.0	0.0	12.5	41.7	0.0	22.7	22.9	0.0	11.7	11
199	CIOI	14041	33.3	0.0	16.7	16.7	25.0	20.8	8.3	0.0	4.2	8.3	0.0	4.5	16.7	6.5	11.7	11
9	ENTI	1053	0.0	8.3	4.2	16.7	0.0	8.3	33.3	0.0	16.7	25.0	0.0	13.6	18.8	2.2	10.6	10
158	CAPG	12031	25.0	16.7	20.8	16.7	0.0	8.3	8.3	0.0	4.2	0.0	20.0	9.1	12.5	8.7	10.6	10
1	BRYP	1011	0.0	25.0	12.5	0.0	8.3	4.2	0.0	41.7	20.8	0.0	[illegible]	0.0	0.0	19.6	9.6	9
200	MOLC	14051	25.0	8.3	16.7	41.7	0.0	20.8	0.0	0.0	0.0	0.0	0.0	0.0	16.7	2.2	9.6	9
7	ENTC	1052	25.0	0.0	12.5	8.3	8.3	8.3	16.7	0.0	8.3	8.3	0.0	4.5	14.6	2.2	8.5	8
141	COLL	11071	25.0	0.0	12.5	16.7	0.0	8.3	0.0	0.0	0.0	8.3	20.0	13.6	12.5	4.3	8.5	8
11	ENTX	1058	0.0	8.3	4.2	0.0	8.3	4.2	8.3	16.7	12.5	0.0	10.0	4.5	2.1	10.9	6.4	6
148	MODM	11131	8.3	25.0	16.7	0.0	0.0	0.0	0.0	16.7	8.3	0.0	0.0	0.0	2.1	10.9	6.4	6

TABLE A. 2 Percentage of Samples in Which Species Were Observed (Cont'd)

41	RHOP	2171	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	50.0	22.7	0.0	10.9	5.3	5
75	STYF	6011	n.n	16.7	8.3	0.0	16.7	8.3	0.0	0.0	8.3	4.2	0.0	0.0	0.0	10.9	5.3	5
83	ELEC	8041	0.0	8.3	4.2	0.0	0.0	0.0	16.7	0.0	8.3	4.2	8.3	10.0	9.1	6.5	4.3	5
111	NERX	9128	0.0	8.3	4.2	n.n	8.3	4.2	0.0	8.3	4.2	0.0	20.0	9.1	0.0	10.9	5.3	5
179	BALI	12145	0.0	0.0	0.0	0.0	0.0	0.0	0.0	33.3	16.7	8.3	0.0	4.5	2.1	8.7	5.3	5
62	LEUX	4038	25.0	8.3	16.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.5	2.2	4
121	SABF	9189	0.0	8.3	4.2	0.0	5	12.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.7	4.3	4
143	UROC	11091	8.3	8.3	8.3	n.n	0.0	0.0	0.0	0.0	0.0	0.0	20.0	9.1	2.1	8.5	4.3	4
23	CHAP	2061	0.0	8.3	4.2	0.0	8.3	4.2	0.0	8.3	4.2	0.0	0.0	0.0	0.0	6.5	3.2	3
29	GRAF	2102	0.0	0.0	0.0	0.0	16.7	8.3	0.0	0.0	0.0	0.0	10.0	4.5	0.0	6.5	3.2	3
125	TCRF	9209	0.0	8.3	4.2	8.3	0.0	4.2	0.0	8.3	4.2	0.0	0.0	0.0	2.1	8.3	3.2	3
133	CREP	11032	0.0	0.0	0.0	0.0	16.7	8.3	8.3	0.0	4.2	0.0	0.0	0.0	2.1	8.3	3.2	3
160	CAPF	12039	0.0	8.3	4.2	0.0	8.3	4.2	0.0	8.3	4.2	0.0	0.0	0.0	0.0	6.5	3.2	3
188	PANH	12191	0.0	8.3	4.2	0.0	8.3	4.2	0.0	0.0	0.0	0.0	10.0	4.5	0.0	6.5	3.2	3
3	CHAX	1028	0.0	8.3	4.2	0.0	0.0	0.0	0.0	8.3	4.2	0.0	0.0	0.0	0.0	4.3	2.1	2
15	AGAT	2021	0.0	0.0	0.0	0.0	8.3	4.2	0.0	0.0	0.0	0.0	10.0	4.5	0.0	4.3	2.1	2
20	CYSP	2081	0.0	16.7	8.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.3	2.1	2
57	GRAX	4018	16.7	0.0	8.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.2	0.0	2.1	2
70	SERT	5062	0.0	0.0	0.0	0.0	16.7	8.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.3	2.1	2
72	METO	5081	0.0	0.0	0.0	-	o.n	4.2	8.3	0.0	4.2	0.0	0.0	4.2	0.0	4.3	2.1	2
85	ELL*	8041	0.0	0.0	0.0	0.0	8.3	4.2	0.0	0.0	0.0	0.0	10.0	4.5	0.0	4.3	2.1	2
86	SCHU	8051	0.0	0.0	0.0	16.7	0.0	8.3	0.0	0.0	0.0	0.0	0.0	0.0	4.2	0.0	2.1	2
126	SPIW	9218	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.7	8.3	0.0	0.0	0.0	0.0	4.3	2.1	2
138	LITO	11061	0.0	8.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	4.5	0.0	4.3	2.1	2
180	BALX	12148	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.3	10.0	9.1	2.1	2.2	2.1	2
10	ENTP	1055	0.0	0.0	0.0	0.0	0.0	0.0	8.3	0.0	4.2	0.0	0.0	0.0	2.1	0.0	1.1	1
14	ACHY	2018	0.0	8.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.1	1
18	CALB	2042	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	4.5	0.0	2.2	1.1	1
25	CHOX	2078	0.0	0.0	0.0	8.3	0.0	4.2	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	1.1	1
27	DAYP	2091	0.0	0.0	0.0	0.0	8.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.1	1
39	PORX	2158	0.0	0.0	0.0	0.0	8.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.1	1
40	RHOS	2161	0.0	6.n	0.0	0.0	8.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.1	1
42	RHOX	2178	0.0	0.0	0.0	9.3	0.0	4.2	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	1.1	1
a3	DESV	3011	0.0	n.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.1	1
54	PYLL	3091	0.0	8.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.1	1
66	CAMX	5038	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.3	4.2	0.0	0.0	0.0	0.0	2.2	1.1	1
74	UNIH	5099	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.3	4.2	0.0	0.0	0.0	0.0	2.2	1.1	1
78	LEPF	b028	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.3	4.2	0.0	0.0	0.0	0.0	2.2	1.1	1
84	ELEP	8042	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	4.5	0.0	2.2	1.1	1
88	BOWG	8071	0.0	8.3	u.?	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.1	1
91	BUGX	8088	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	4.5	0.0	2.2	1.1	1
98	EULV	9051	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	10.0	4.5	0.0	2.2	1.1	1
109	NERS	9122	0.0	0.0	0.0	0.0	n.n	0.0	0.0	0.0	0.0	0.0	10.0	4.5	0.0	2.2	1.1	1
112	NEPX	9138	0.0	0.0	0.0	8.3	0.0	4.2	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	1.1	1
115	PHYX	9158	0.0	0.0	0.0	0.0	8.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.1	1
128	NERM	9231	8.3	n.n	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	1.1	1
130	COLA	11011	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.3	4.5	2.1	0.0	1.1	1
142	THAL	11081	0.0	8.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.1	1
145	ANOS	11111	8.3	0.0	4.2	0.0	0.0	n.n	0.0	0.0	0.0	0.0	0.0	0.0	2.1	0.0	1.1	1
156	AMPX	12028	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	4.5	0.0	2.2	1.1	1
186	CARM	12171	0.0	8.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.1	1
107	EURD	12181	0.0	0.0	0.0	0.0	8.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.1	1
189	DECF	12209	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.3	4.5	2.1	0.0	1.1	1
191	IDOP	12212	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	4.5	0.0	2.2	1.1	1
193	TANC	12231	0.0	0.0	0.0	0.0	0.0	0.0	8.3	0.0	4.2	0.0	0.0	0.0	2.1	0.0	1.1	1
203	STYP	14061	0.0	8.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.1	1
2	CHAA	1021	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
b	ENTA	1051	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
9	ENTL	1054	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
13	CHLF	1079	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

TABLE A.2 Percentage of Samples in Which Species Were Observed (Cont'd)

TABLE A.2 Percentage of Samples in Which Species Were Observed (Cont'd)

120	SABM	9181	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
122	SERV	9191	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
124	TERL	9201	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
129	SIPX	10011	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
131	CERG	11021	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
134	CREX	11038	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
135	HERX	11048	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
136	ILYO	11051	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
137	ILYX	11058	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
139	LITS	11062	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
140	LITX	11068	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
144	UNIN	11109	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
146	ANOX	11118	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
147	CRAV	11121	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
150	SAXA	11151	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
151	TERB	11161	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
153	TERE	11169	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
154	AEGL	12011	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
155	AMPR	12021	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
157	AMPE	12029	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
159	CAPX	12038	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
163	ELAL	12061	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
164	GAMA	12071	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
165	GAML	12072	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
166	GAMX	12078	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
167	GAMF	12079	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
168	GRUC	12081	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
169	JASA	12091	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
170	MELD	12101	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
171	MELN	12102	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
173	UNIC	12129	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
174	UNCI	12131	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
175	BALA	12101	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
185	BRAX	12169	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
190	IDOB	12211	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
192	JAEM	12221	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
194	ASTF	13019	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
195	ASTO	13011	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
196	ASCX	14018	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
197	AMAX	14028	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
201	MOLM	14052	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
202	MOLX	14058	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

TABLE A.3 Millstone Point Exposure Panels

Species List

Alpha Code	Species Name	Data in % or #	Group
	Chlorophyta		
BRYP	*Bryopsis plumosa*	%	A
CHAA	*Chaetomorpha area*	%	A
CHAX	*Chaetomorpha* sp.	%	A
CLAX	*Cladophora* sp.	%	A
CØDD	*Codium fragile*	%	A
ENTA	*Enteromorpha clathrata*	%	A
ENTC	*Enteromorpha compressa*	%	A
ENTI	*Enteromorpha intestinalis*	%	A
ENTL	*Enteromorpha linza*	%	A
ENTP	*Enteromorphu prolifera*	%	A
ENTX	*Enteromorphu* spp.	%	A
ULVL	*Ulva lactuca*	%	A
CHLF	Unidentified Green Film (Chlorophyceae)	%	A
	Rhodophyta		
ACHX	*Achrochaetium* sp.	%	A
AGAT	*Agardhiella tenera*	%	A
ANTX	*Antihamnion* sp.	%	A
CALI	*Callithamnion baileyi*	%	A
CALB	*Callithamnion byssoidewn*	%	A
CALX	*Callithamnion* spp.	%	A
CERD	*Ceramiwn diaphanum*	%	A
CERR	*Ceramium rubrum*	%	A
CERX	*Ceramium* spp.	%	A
CHAP	*Champia parvula*	%	A
CHOB	*Chondria baileyana*	%	A
CHØX	*Chondria* sp.	%	A
CYSP	*Cystoclonium purpurem*	%	A
DAYP	*Daysa pedicellata*	%	A
GRAC	*Gracilaria confervoides*	%	A
GRAF	*Gracilaria foliifera*	%	A
GRAS	*Gracilaria* spp.	%	A
GRIA	*Grinellia americana*	%	A
HERT	*Herposiphonia tenella*	%	A
LØMB	*Lomentaria baileyana*	%	A
LØMX	*Lomentaria* spp.	%	A
PØLI	*Polysiphonia nigra*	%	A
PØLN	*Polysiphonia nigrescens*	%	A

TABLE A 3 Millstone Point Exposure Panels (Cont'd)

Species List			
Alpha Code	Species Name	Data in % or #	Group
	Rhodophyta (continued)		
PØLX	*Polysiphonia* spp.	%	A
PØRU	*Porphyra umbilicus*	%	A
PØRX	*Porphyra* spp.	%	A
RHOS	*Rhodomela subfusca*	%	A
RHØP	*Rhodymenia palmata*	%	A
RHØX	*Rhodymenia* spp.	%	A
	Phaeophyta		
DESV	*Desmarestia viridis*	%	A
ECTX	*Ectocarpus* sp.	%	A
ELAX	*Elachistea* sp.	%	A
FUCE	*Fucus evanescens*	%	A
FUCS	*Fucus* spp.	%	A
LAMA	*Laminaria agardhii*	%	A
LAMX	*Laminaria* spp.	%	A
PETA	*Petalonia fascia*	%	A
PHYR	*Phycodrus rubens*	%	A
PUNL	*Punctaria latifolia*	%	A
PUNX	*Punctaria* sp.	%	A
PYLL	*Pylaiella litoralis*	%	A
PYLX	*Pylaiella* sp.	%	A
	Porifera		
GRAI	*Scypha ciliata (grantia)*	%	B
GRAX	*Scypha* spp.	%	B
HALB	*Halichondria bowerbanki*	%	B
HALP	*Halichondria panicea*	%	B
HALX	*Halichondria* spp.	%	B
LEUB	*Leucosolenia botryoides*	%	B
LEUX	*Leucosolenia* spp.	%	B
RENF	Renierinae	%	B
	Cnidaria		
UNIA	Actinaria	%	B
ANTU	Unidentified Anthozoan	%	B
CAMX	*Campanularia* sp.	%	B
DIAL	*Diadumene Zeucolena*	%	B
ØBEX	*Obelia* sp.	%	B

TABLE A.3 Millstone Point Exposure Panels (Cont'd)

Species List			
Alpha Code	Species Name	Data in % or #	Group
	Cnidaria (Continued)		
	Sagartia lancolena (1976)	#	B
	Sagartia sp. (1 976)	#	B
SERP	*Sertularia pumila*	%	B
SERT	*Sertularia* spp.	%	B
TUBX	*Tubularia* sp.	%	B
METD	*Metridium dianthus*	#	B
METS	*Metridium senile*	#	B
UNIH	Unidentified Hydroid	%	B
	Platyhelminthes		
STYE	*Stylochus ellipticus*	#	C
LEPA	*Leptoplana augusta*	#	C
LEPX	*LeptopZana* spp.	#	C
LEPF	*Leptoplanidae*	#	C
	Rhynchocoela		
RHYP	Unidentified Rhynchocoela	%	C
	Ectoprocta (Bryozoa)		
	Encrusting		
CALA	*Callopora aurita*	%	B
CRIE	*Crisea eburnea*	%	B
CRYP	*Cryptosula pallasiana*	%	B
ELEC	*Electra crustulenta*	%	B
ELEP	*Electra piZosa*	%	B
ELEM	*Electra monostachys*	%	B
SCHU	*Schizoporella unicornis*	%	B
TEGU	*Tegella unicornis*	%	B
	Filamentous		
BØWG	*Bowerbankia gracilis*	%	B
BUGS	*Bugula simplex*	%	B
BUGT	*Bugula turrita*	%	B
BUGX	*Bugula* spp.	%	B

TABLE A.3 Millstone Point Exposure Panels (Cont'd)

Alpha Code	Species Name	Data in % or #	Group
	Species List		
	Annelida		
AMPF	Ampharetidae	#	C
AMPH	*Amphitrite* sp.	#	C
CAPI	Capitellidae	#	C
CIRF	Cirratulidae	#	C
CIRG	*Cirratulus grandis*	#	C
EUGR	*Euchone rubrocinta*	#	C
EULV	*Eulalia viridis*	#	C
EULX	*Eulalia* spp.	#	C
EUMX	*Eumida* sp.	#	C
GLYF	Glyceridae	#	C
HARI	*Harmothoe imbricata*	#	C
HYDD	*Hydroides dianthus*	#	C
HYDX	*Hydroides* sp.	#	C
LEPS	*Lepidonotus squamatus*	#	C
MARS	*Marphysa sanguinea*	#	C
MARX	*Marphysa* sp.	#	C
NERP	*Nereis pelagica*	#	C
NERS	*Nereis succinea*	#	C
NERV	*Nereis virens*	#	C
NERX	*Nereis* spp.	#	C
NEPX	*Nephtys* sp.	#	C
NØTL	*Notomastus latericeus*	#	C
PHYA	*Phyllodoce arenae*	#	C
PHYX	*Phyllodoce* sp.	#	C
PHYF	Phyllodicidae	#	C
PØDØ	*Podarke obscura*	#	C
PØLC	*Polydora ciliata*	#	C
PØLF	Polynoidae	#	C
SABM	*Sabella microphthalma*	#	C
SABF	Sabellidae	#	C
SERV	*Serpula vermicularis*	#	C
SERW	Serpulid tubes	#	C
TERL	*Terebella lapidaria*	#	C
TERF	Terebellidae	#	C
SPIW	Spirorbis tubes	#	
MUDW	Mudworm tubes	%	
NERM	*Platynereis megalops*	#	C
SIPX	Sipuncula	#	C

TABLE A.3 Millstone Point Exposure Panels (Cont'd)

Species List			
Alpha Code	Species Name	Data in % or #	Group
	Mollusca		
	Gastropoda		
CØLA	*Anachus avara*	#	C
CERG	*Cerithiopsis greenii*	#	C
CREF	*Crepidula fornicata*	#	C
CREP	*Crepidula plana*	#	C
CREX	*Crepidula* sp.	#	C
HERX	*Hermaea* sp.	#	C
ILYØ	*Ilyanassa obsoleta*	#	C
ILYX	*Ilyanassa* spp.	#	C
LITØ	*Littorina obtusata*	#	C
LITS	*Littorina saxatilis*	#	C
LITX	*Littorina* spp.	#	C
CØLL	*Mitrella lunata*	#	C
THAL	*Nucella lapilla*	#	C
URØC	*Urosalpinx cinerea*	#	C
UNIN	Unidentified nudibranch	#	C
	Pelecypoda		
ANØS	*Anomia simplex*	#	B
ANØX	*Anomia* spp.	#	B
CRAV	*Crassostrea virginica*	#	B
MØDM	*Modiolus modiolus*	#	B
MYTE	*Mytilus edulis*	%	B
SAXA	*Saxicava artica (Hiatella artica)*	#	B
TERB	*Teredo bartschi*	%	B
TERN	*Teredo navalis*	%	B
TERE	Teredinidae	%	B
	Arthropoda		
	Amphipoda		
AEGL	*Aeginella longicornis*	#	C
AMPE	*Amphitoidae*	#	C
AMPR	*Amphithoe rubricata*	#	C
AMPX	*Amphithoe* spp.	#	C
CAPG	*Caprella geometrica*	#	C
CAPX	*Caprella* sp.	#	C
CAPF	Caprellidae	#	C

TABLE A.3 Millstone Point Exposure Panels (Cont'd)

Species List			
Alpha Code	Species Name	Data in % or #	Group
	Amphipoda (Continued)		
CHET	*Chelura terebrans*	#	C
CORC	*Corophium cylindricum*	#	C
ELAL	*Elasmopus laevis*	%-#	C
GAMA	*Gammarus annulatus*	#	C
GAML	*Gammarus locusta*	#	C
GAMX	*Gammarus* sp.	#	C
GAMF	Gammaridae	#	C
GRUC	*Grubia compta*	#	C
JASA	*Jassa falcata*	#	C
MELD	*Melita dentata*	#	C
MELN	*Melita nitida*	#	C
MICX	*Microdeutopus* sp.	#	C
UNIC	Unidentified copepods	#	C
UNCI	*Unciola irrorata*	#	C
	Cirripedia		
BALA	*Balanus amphitrite niveus*	%	B
BALB	*Balanus balanoides*	%	B
BALC	*Balanus crenatus*	%	B
BALE	*Balanus eburneus*	%	B
BALI	*Balanus improvisus*	%	B
BALX	*Balanus* spp.	%	B
	Limnoridae		
LIML	*Limnoria lignorum*	#	C
LIMT	*Limnoria tripunctata*	#	C
LIMU	*Limnoria tuberculata*	#	C
LIMG	*Limnoria* tunnels	#	C
	Decapoda		
BRAX	Brachyura		
CARM	*Carcinus maenas*	#	C
EURD	*Eurypanopeus depressus*	#	C
PANH	*Panopeus herbstii*	#	C
DECF	Unidentified crabs	#	C

TABLE A.3 Millstone Point Exposure Panels (Cont'd)

Species List			
Alpha Code	Species Name	Data in % or #	Group
	Arthropoda (Continued)		
	Isopoda		
IDØB	*Idotea baltica*	#	C
IDØP	*Idotea phosphorea*	#	C
JAEM	*Jaera marina*	#	C
TANC	*Tanais cavolini*	#	C
	Echinodermata		
ASTF	Asteriidae	#	C
ASTØ	*Asterias forbesii*	#	C
	Chordata		
ASCX	*Ascidia* sp.	%	B
AMAX	*Amaroucium* sp.	%	B
BØTS	*Botryllus schlosseri*	%	B
CIØI	*Ciona intestinalis*	%	B
MØLC	*Molgula citrina*	%	B
MØLM	*Molgula manhuttensis*	%	B
MØLX	*Molgula* spp.	%	B
STYP	*Styella partita*	%	B

A = Algae; B = Invertebrate, sessile; C = Invertebrate, motile.

List updated through May 11, 1976.

DISTRIBUTION

No. of Copies

OFFSITE

A. A. Churm
ERDA Chicago Patent Group
9800 South Cass Avenue
Argonne, IL 60439

3 M. Jinks, Chief Mail and Files
USNRC Central Files
Washington, DC 20555

245 ERDA Technical Information Center
For Basic Distribution Under NRC-1

P. G. Voilleque
ERDA Health Services Laboratory
Idaho Falls, ID 83401

H. T. Peterson
USNRC Office of Standards Development
Washington, DC 20555

No. of Copies

ONSITE

ERDA Richland Operations Office

P. G. Holsted

Atlantic Richfield Hanford Company

G. E. Backman

United Nuclear Industries, Inc.

A. E. Engler

Hanford Engineering Development Laboratory

G. D. Carpenter

40 Battelle-Northwest

K. L. Gore (5)
J. L. Helbling
C. Hudges
J. Johnston (10)
L. D. Kannberg (5)
J. Mahaffey
J. A. Strand
J. M. Thomas
D. G. Watson (2)
B. E. Vaughan
Technical Information Files (5)
Technical Publications (3)