Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

Assessment of Similarity Indices for Undesirable Properties and a new Tripartite Similarity Index Based on Cost Functions

Rodham E. Tulloss[1]

Introduction

Comparison of lists is a common element of many studies including ethnomycological, ecological, and mycological investigations. The items on the lists might be species in a habitat, uses of a given organism by indigenous people, character states present in an individual fungus, or lists of unusual spellings in segments of the Dead Sea Scrolls. Often, it is desirable to express the similarity of two related lists by some formula (a similarity index). Such an index might be used in summarizing data otherwise presented or as input to further numerical processing, such as the creation of a dendrogram (Pankhurst, 1991:54).

In examining several works using formulae to provide a single number expressing the similarity of the contents of two lists, a number of difficulties with the formulae were noted. For example, for some indices the same value was generated for two or more quite different situations, e.g., one in which a pair of lists were nearly identical, and another in which one list was much larger than the other. This problem came up during review of material for the present book, thus motivating the present chapter. The purpose of this chapter is to motivate, describe, and offer an implementation for, a working similarity index that avoids the difficulties noted for the others.

Examination of indices.

The list of 20 existing and commonly used similarity indices (similarity coefficients) supplied and characterized by Hayek (1994) was examined in detail in search of indices that did not have this or a similar problem. Difficulties, some significant and well-known, were identified with all the examined similarity indices. No problem-free index was found in the list.

At least in part, the problems with the indices arise because of an

[1]*P. O. Box 57, Roosevelt, New Jersey 08555-0057, U.S.A.*

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

apparent motivation of their designers—to create an index that has as its formula an equation that can be related simply to some natural language statement about a pair of lists. The difficulty with this approach is that our intuitive concept of similarity between two lists includes a number of component requirements that may not have been made explicit in the process of developing such indices. In this chapter, I report an attempt to make these requirements explicit and derive an index based on these explicit requirements.

Anew index.

The new index was developed based on an idea familiar in manufacturing engineering—a metric based on cost functions. These cost functions are designed to express conflicting requirements mathematically. In industry, cost functions are designed to increase in value (provide a reward) when something you want to happen is happening and to decrease in value (provide a penalty or disincentive) when something you don't want to happen is happening. Usually, a group of cost functions, some providing rewards and some providing penalties, are multiplied together in order to get a single number metric that, for example, might be used to monitor or control a manufacturing process when the quality of such a process depends upon the combined states of a number of variables.

The cost functions developed for the purpose of dealing with similarity of lists were combined to generate the new similarity index, and the latter was tested on cases that had proven to demonstrate what I interpreted as limitations of the pre-existing indices. The new index behaves as it was designed to do. The new single number index proves to be satisfactorily close to being linear and invariant [two properties that Hayek (1994) states are important for measures of association of which similarity indices are one type]. A very simple computer program implementing the index in the GWBASIC language is provided in Appendix 1 of this chapter. Test input and output are displayed in Appendices 2 through 4.

Prior to explaining development of the new index, this chapter presents the relevant, pre-existing similarity indices—giving the formula and name for each and using the variable names chosen by Hayek (1994). Some problems with each index are illustrated or stated.

Methods

Examination of indices.

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders.* (Parkway Publishers, Boone, North Carolina): 122-143.

Hayek's list of 20 similarit coefficients (Hayek, 1994:table1) were examined and examples devised to demonstrate ones in which the formulae exhibited various problems.

Notations.

The notation for the expression of similarity indices is that adopted by Hayek (1994:208, table 1).

- a = the number of entries that are common to both lists.
- b = the number of entries in the first list that are not in the second.
- c = the number of entries in the second list that are not in the first.
- n = the maximum number of entries that could occur in either list.
- d = the number of entries (of the maximum n) that do not appear in either list.

The problem with n and d in contemporary mycological studies.

The formulae in Hayek's list of similarity indices that utilized the variables n and d are not reviewed in this chapter because of the impossibility at the present time of creating a list of fungi for a region in the Americas that could be called complete. We are simply comparing snapshots made through cameras with a narrow field of view and numerous "blind spots" in the lenses. The inability even to fix on a clear estimate of the fungi in consideration is treated in various chapters in this book. The indices not reviewed are those numbered 16 through 20 by Hayek (Forbes Coefficient of 1907, Gilbert and Wells Coefficient, Forbes Coefficient of 1925, Tarwid Coefficient, and Resemblance Equation Coefficient). Hayek's coefficient 4a (Dice's Asymmetric Indices) does not produce a single number as output, pairs of numbers such as these can be depicted graphically as points in space.

The Simpson Coefficient—an Example and the Primary Component Requirements

Simpson Coefficient (1).

$$I = \frac{a}{a + min(b, c)}$$

As many others have noted, this index is independent of the number of entries on the larger of two lists to be compared. If the smaller list has x percent of its entries appearing also in the larger list, then the value of the

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders.* (Parkway Publishers, Boone, North Carolina): 122-143.

index is $\frac{x}{100}$ no matter how extreme the difference in list sizes may be. This seems highly undesirable if several pairs of lists are to be compared via the generated Simpson Coefficient values. Two lists containing 1000 and 10 entries respectively and sharing 5 entries will have the same index (0.5) as will two lists containing 10 entries each and sharing 5 entries. This situation, a formula's providing the same index for a number of different input data sets is sometimes called *aliasing* in engineering contexts.

Primary component REQUIREMENTS and RECOMMENDATIONS.

From this example, we can see that there are at least three potentially conflicting requirements for a similarity index. The first two requirements are stated negatively and suggest penalty cost funtions, and the third requirement suggests a reward cost function:

REQUIREMENT 1: A similarity index shall be sensitive to the relative size of the two lists to be compared; and great difference in size shall be interpreted to reduce the value of the similarity index.

REQUIREMENT 2: A similarity index shall be sensitive to the size of the sublist shared by a pair of lists; and an increase in difference in size between the smaller of the two lists and the sublist of common entries shall be interpreted to reduce the value of the similarity index.

The first two requirements are stated negatively and suggest penalty cost functions. The following requirement suggests a reward cost function:

REQUIREMENT 3: A similarity index shall be sensitive to the percentage of entries in the larger list that are in common between the lists and to the percentage of entries in the smaller list that are in common between the two lists and shall increase as these two percentages increase.

For logical completeness, we add the following definition:

DEFINITION 1: When two lists to be compared by means of a similarity index are of the same size (cardinality), one shall arbitrarily be selected to be called "the larger." The remaining list shall

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

be “the smaller.”)

It is also desirable, as noted by Hayek, that a similarity index be bounded above and below and that the index achieve its upper limit in the case of identical lists and its lower limit in the case of disjoint lists.

REQUIREMENT 4: A similarity index shall yield values having fixed upper and lower bounds.

REQUIREMENT 5: A similarity index shall have the property that when two lists are identical, the similarity index for the two lists shall be equal to the upper bound of the index.

REQUIREMENT 6: A similarity index shall have the property that when two lists have no entries in common, the similarity index for the lists shall be equal to the lower bound of the index.

RECOMMENDATION 1: The upper bound of a similarity index should be one; the lower bound of a similarity index should be zero.

REQUIREMENT 7: Distribution of values of a similarity index between zero and one shall be such that (a) if the size of two input lists is fixed, then the output shall vary roughly directly as the number of entries shared between the lists; and (b) if the smaller list is a subset of the larger list, then the value of the similarity index shall vary roughly inversely as the size of the larger list.

REQUIREMENT 7 part (a) is a variation of the definition of “linearity” of Hayek (1994), which is discussed further, below.

Experience with other similarity indices (also below) shows that an additional requirement must be added to the list. It relates to convenience in using a program that implements a similarity index.

REQUIREMENT 8: A similarity index program shall check its input data to verify that the following relationships hold:

$$a + b > 0$$

$$a + c > 0$$

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

Review of Other Similarity Indices

Hayek (1994:table 20 and accompanying text) provides valuable analyses of the similarity indices she discusses. The approach taken in this chapter is intended to augment her analyses by producing illustrative examples of problems with the similarity indices. In some cases, the behavior of an index is such that it seems unsuited for use. With others, the difficulties are more subtle. The examples devised are related, in each case, to one or more requirements that are not met. Hayek's number is provided for each index.

Second Kulczynski Coefficient (2).

$$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$$

One of the problems with this index is that if the two lists are very disparate in size and if all the entries on the smaller list appear in the larger, then the minimum value of the coefficient is 0.5. An example can be constructed easily in which the coefficient's value is very unsatisfactory (e.g., <a, b, c> = <5,100,0>.). Or compare the values of the coefficient for the input data triples <2,1,26> and <3,0,25>—cases in which there is a list of three items and a list of 28 items, on the one hand sharing two entries in common, on the other sharing three entries. The value of the coefficient in the first case is 0.37; and in the second, 0.55. The change seems very large given the small difference in the input data, and the values seem inappropriately high. Compare the situation in which there are two lists of equal size sharing 55% of their entries. In this case, surely one of greater similarity than either of the previous examples, the value of the second Kulczynski Coefficient is also 0.55. The coefficient violates two primary component requirements—REQUIREMENTS 1 and 3.

The coefficients numbered 14 (McConnaughey) and 15 (Johnson Coefficient) in Hayek's list are variations on the second Kulczynski Coefficient by scale transformation and by multiplication by 2 respectively. Therefore, they have the problems of coefficient number 2 and are not considered further.

Ochiai/Otsuka Coefficient (3).

$$\frac{a}{\sqrt{(a+b)(a+c)}}$$

This index has a somewhat subtle aliasing problem, but the more seri-

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

ous problem is caused by the square root function of the denominator. Consider the product ($a + b$) ($a + c$). Suppose it is the product of a number of small prime integers, say, (2^6)(3^4). Then consider some of the possible cases that lead to the value 24/72 (≈0.33):

<24,0,192> - lists of 24 and 216 entries, the set of entries in the smaller forming a subset of the entries of the larger (equivalent to the data triple <*n*,0,8*n*>)

<24,3,168> - lists of 27 and 192 entries with 24 entries in common

<24,24,84> - lists of 48 and 108 entries, with 24 entries in common

<24,30,72> - lists of 54 and 96 entries with 24 entries in common

<24,40,57> - lists of 64 and 81 entries with 24 entries in common

<24,48,48> - two lists of 72 entries with 24 entries in common (equivalent to the data triple <*n*,2*n*,2*n*>).

One might argue reasonably that REQUIREMENT 1 is not well-satisfied here. However, also consider the triples <1,0,(*n* - 1)>, where $n \geq 1$. These triples describe the situation in which there are two lists, one with a single entry that is also in the larger list and one with cardinality of *n*. The value of the Ochiai/Otsuka Coefficient for these triples is

$$\frac{1}{\sqrt{n}}$$

If this value is computed for several values of *n*, not only will the value be observed to be unsatisfactorily "high," but the drop off in value as *n* increases is, of course, governed by the square root function; so the unsatisfactory nature of the index becomes more pronounced as *n* increases. For example, <1,0,1> yields the value 0.71; <1,0,69> yields the value 0.12; etc. This is a failure to satisfy REQUIREMENT 7(b).

Coefficient number 9 on Hayek's List (Correlation Ratio), is the square of the index presently under discussion; hence, the Correlation Ratio eliminates the problem with REQUIREMENT 7(b) just noted. However, the problem with REQUIREMENT 1 is unresolved; and the Correlation Ratio is still not linear (Hayek, 1994:213) because, with fixed list sizes, the value now varies as the square of the variable *a* [i.e., violates REQUIREMENT 7(a)].

Dice Coefficient (4).

$$\frac{a}{a + \frac{b + c}{2}}$$

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

Because of the use of the mean of b and c in this coefficient's formula, it is very easy to demonstrate aliasing—a single number may be the mean of many pairs of numbers. This means that, since b and c reflect the sizes of the two lists under comparison, the Dice Coefficient suffers from some insensitivity to the difference in size of the two lists, a problem with REQUIREMENT 1. The Nonmetric Coefficient (13 in Hayek's list) is the additive inverse of The Dice Coefficient; and, hence, has the same difficulties besides being designed to reverse the scale of REQUIREMENTS 5 and 6.

Jaccard Coefficient (5).

$$\frac{a}{a + b + c}$$

The Jaccard Coefficient experiences aliasing for the same reason that the Dice Coefficient does—for a given sum of b and c, many pairs of values can produce the same sum. Hence, problems vis-a-vis REQUIREMENT 1 occur. Moreover, the absence of the averaging function of Dice's Coefficient means that the values of the Jaccard Coefficient may be undesirably depressed. Hayek (1994:211) points out that this metric is not linear.

Sokal and Sneath Coefficient (6).

$$\frac{a}{a + 2b + 2c}$$

The problem is the same as with formulae numbers 4 and 5; moreover, the undesirable depression of values is exacerbated. The metric is not linear (Hayek, 1994:213).

First Kulczynski Coefficient (7).

$$\frac{a}{b + c}$$

The same problem is experienced again. There is the added disadvantage that instead of two identical lists getting a similarity index of one, a divide-by-zero problem arises—REQUIREMENTS 4 and 5 are violated. The metric is not linear (Hayek, 1994:213).

Mountford Coefficient (8).

$$\frac{2a}{2bc + ab + ac}$$

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

This coefficient is immediately seen to have a divide by zero problem and, therefore, violates REQUIREMENTS 4 and 5. Moreover, the behavior of the formula can be very erratic and, hence, produce counter intuitive values. Consider the following cases:

<100,1,1> - 2 lists of 101 entries, with each list containing only one entry not on the other list

<101,0,0> - 2 identical lists of 101 entries each

<1000,1,1> - 2 lists of 1001 entries, with each list containing only one entry not on the other list

<100,2,2> - 2 lists of 102 entries, with each list containing exactly two entries not on the other list

<100,5,5> - 2 lists of 105 entries, with each list containing exactly five entries not on the other list

<100,5,0> - a list of 105 entries containing all of the entries in a list of 100 entries

<5,5,0> - a list of 10 entries containing all of the entries in a list of 5 entries.

We have already noted that the second case leads to division by zero. Contrast this with the first and third cases (intuitively, very nearly the same states of affairs—nearly perfect matches between two lists) that yield values slightly less than 1.0. The point is that when two lists are nearly identical, the coefficient has a value very close to one, but when they are identical, we get division by zero. On the other hand, once a little more difference exists between the two lists, the value of the coefficient crashes. For the fourth case, the value is 0.49. For the fifth, it is 0.19. The sixth case, which appears to be very close to the fifth yields the value 0.4. That the coefficient violates REQUIREMENTS 1 and 3 can be seen from the fact that the seventh case is aliased with the sixth—it also yields the value 0.4.

Hayek (1994:213) also notes that this coefficient is both nonlinear and not invariant. To see the latter in a single example, compare a case in which (like the seventh) one list's entries comprise exactly 50% of the other's—<100,100,0>. However, here the coefficient yields the value 0.02.

The Mountford Coefficient should not be used as a similarity index.

Braun-Blanquet Coefficient (10).

This index is very similar to the Simpson Coefficient (see above) except that the denominator is the cardinality of the larger list instead of

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

the smaller one. The same set of problems arises, especially because of the complete insensitivity to the size, in this case, of the smaller of the two lists. The Savage Coefficient (Hayek's number 12), is really the mild-mannered additive inverse of the Braun-Blanquet Coefficient. Hence, it runs into the same difficulties and also inverts the scale of REQUIREMENTS 5 and 6.

Fager and McGowan Coefficient (11).

$$\frac{a}{\sqrt{(a+b)(a+c)}} - \frac{max(b, c)}{2}$$

Hayek observes that the Fager and McGowan coefficient is the same as the Ochiai/Otsuka Coefficient less a "correction factor." One price paid for this correction is the violation of REQUIREMENTS 4 and 6; for the resulting formula produces values that are unbounded below. Another price is paid when a relatively large number is subtracted from a small fraction—the correction factor takes over the process and obliterates the important sensitivities that were present in the Ochiai/Otsuka Coefficient. For example, compare the values for these two triples:

<50,50,50> two lists each having 100 entries, with 50 entries in common

<2,50,50> two lists each having 52 entries, with only 2 entries in common.

Both cases yield values ≈ -50. Hence, the coefficient fails to meet REQUIREMENTS 2 and 3. [The coefficient is nonlinear (Hayek, 1994:213).] The fact that the correction factor is based on the max function means that this dominant factor will remove any sensitivity to the size of the smaller of the two lists being compared—a violation of REQUIREMENT 1.

The Fager and McGowan Coefficient should not be used as a similarity index.

Three Cost Functions and the Tripartite Similarity Index

Having explained reasons for concern regarding use of the similarity indices reviewed by Hayek, the next step is to try to generate a similarity index that will satisfy the requirements developed above. This function is composed of three pieces. It is tripartite so that there will be a factor (a cost function) representing each of the first three (conflicting) require-

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

ments. When the three factors are multiplied together, a similarity index is generated that has the required sensitivities to input data.

Cost Function 1.

The first cost function is designed to provide a penalty for pairs of lists according to REQUIREMENT 1:

$$\frac{\log\left(1 + \frac{min(b,c) + a}{max(b,c) + a}\right)}{\log 2} = U$$

This function will always have a value greater than zero and less than or equal to one because the formula is based on logarithms base 2. [Division by log 2 converts the log function to a $\log_2$ function. The function is expressed in this slightly more complicated way in order to have it relate directly to what is programmable in simple implementations of BASIC (i.e., with $\log_2$ not available).] The function takes on the value one when the two lists being compared are of the same size. It satisfies REQUIREMENTS 4 and 5 as well as REQUIREMENT 1. The logarithmic expression of the numerator will always have value greater than zero because it makes no sense to perform a comparison between lists one or both of which have no members. (See REQUIREMENT 8.)

Cost Function 2.

The second cost function is designed to provide a penalty for pairs of lists according to REQUIREMENT 2:

$$\frac{1}{\sqrt[4]{\frac{\log\left(2 + \frac{min(b,c)}{a+1}\right)}{\log 2}}} = S$$

The value of the second cost function will always be less than one. I selected the square root in the denominator based on trial runs of the function and the particular root was selected simply to give results that are intuitively pleasing. The value of *a* is increased by one in order to avoid division by zero when the two lists being compared have no entries in common. If the number of elements in common between the two lists under consideration is small, then *a* is small relative to min(*b*, *c*); hence, the second cost function will have a value less than one (will act as a penalty function). The second cost function takes the value one when the two lists being compared are identical. Hence, the second cost function is

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

designed to meet REQUIREMENT 2 and REQUIREMENTS 5 and 6.

Cost Function 3.

The third cost function is designed to provide a reward to pairs of lists according to REQUIREMENT 3:

$$\frac{\log\left(1 + \frac{a}{a+b}\right) \cdot \log\left(1 + \frac{a}{a+c}\right)}{(\log 2)^2} = R$$

Each logarithmic factor of the numerator is divided by log 2. This is equivalent to having used logarithms base 2 instead of the log function. Since logarithms base 2 may not easily be available to a person programming this formula, the equivalent (but longer) form is provided. The third cost function not only satisfies REQUIREMENT 3; but, because of the use of logarithms base 2, both the factors of the numerator can be seen to approach zero as a decreases relative to b or c and to approach one as b or c, as the case may be, approaches zero. The limit values are achieved. Hence, REQUIREMENTS 4 through 6 are met as well.

Tripartite Similarity Index.

We can then form the Tripartite Similarity Index (T) by multiplying the three component cost functions and scaling “to taste”:

$$\sqrt{U \times S \times R} = T$$

By creating an index from the product of the cost functions and implementing it in a GWBASIC program satisfying REQUIREMENT 8, all the primary component requirements previously developed are satisfied.

Invariance

Invariance is a property of a function f (in this case a similarity index) that assures that, for any input data $<a,b,c>$ and any factor n, $f(a,b,c) = f(na,nb,nc)$. Hayek (1994:230) states, “Seemingly, no cogent biological reason argues for the use of a measure that is not invariant in this sense.” The Tripartite Similarity Index is very close to invariant. Since the equation was an attempt at an engineering solution, we can ask if the “near invariance” is satisfactory for application purposes. Some of the test data applied to the GWBASIC implementation of the index provided results listed in Appendix 2. From such experiments, it appears that the “near invariance” of T is satisfactory for our purposes.

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

Linearity

The form of linearity strongly urged by Hayek (1994:230) for a similarity index is described by her as follows: "Linearity in measures of association means equal amounts of change in the value of the coefficient when values of joint occurrence change by a factor of one." To demonstrate linearity or "near linearity" experimentally, I selected several pairs of list sizes and, for each such pair, created data triples running from the case in which no entries were shared between the lists to the case in which the smaller list was a subset of the larger one. As in the case of invariance, the Tripartite Similarity Index is not precisely linear, but is extremely close to linear in its behavior on the trial data. Some of the trial data with the corresponding values for *T* are given in Appendix 3 to this chapter.

Variation as the Inverse of the Size of the Larger List of a Pair

It was also a goal for *T* to vary roughly as the inverse of the size of the larger of the two lists compared—in order to avoid the problem noted with the Ochiai/Otsuka Coefficient. A set of data triples of the form $<1,0,n - 1>$ was input to theTripartite Similarity Index. The input and output values are supplied in Appendix 4 to this chapter.

Manipulation of *T* Values

It is important to remember that, while the Tripartite Similarity Index appears to have the desirable property of creating an intuitively satisfying scale on which degrees of similarity can be assigned, this scale may be perceived as abstract, i.e., there is not a simple, compact phrase giving a meaning to the values computed—in contrast to values produced by the Simpson Coefficient. For example, it would be meaningless to convert *T* to percent. Nevertheless, there appears to be no reason not to manipulate the Tripartite Similarity Index values in post processing such as the generation of dendrograms depicting supposed relationships between a set of lists compared using the index. The property of linearity is cited by Hayek (1994:230) as one which supports such post processing.

What does "accuracy" mean in the case of similarity indices? Given correct computation, there is no absolute right answer (see the plethora of attempts at such indices). Three variables are condensed to a single value

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

with concomitant loss of information. Three dimensions are compressed to a line. Our primary hope is that our intuitions about a loosely defined property of points in the three dimensional space (similarity) is reflected in the position of a corresponding point on a line. It is a matter of the sort of distinction one wishes to draw and of the behavior of the index used, of course; but it seems to me that an index that behaves in such a way as to make two digits of accuracy insufficient is flawed. This is certainly a problem with the Fager and McGowan coefficient—along with other difficulties. I would round off the output of the tripartite similarity coefficient to two decimal places except, perhaps, when comparing a set of very dissimilar lists (i.e., in cases in which all the computed values of T have two or more leading zeros); and then I would not suggest using more than one nonzero digit.

As was recommended by Hayek (1994), when publishing data summarized by means of a similarity index, it is valuable to provide a matrix of the input data for the set of lists involved. In addition, other forms of graphical display of the data or values computed from it may enhance understanding by readers (e.g., use of Dice's Asymmetric Indices as defining points in 2-space). Evaluation of such a publication and potential for reproduction of results is greatly facilitated by following such recommendations.

Summary

The purpose of this chapter was to provide a workable similarity index for use in the comparison of pairs of lists. Pre-existing indices of this type are not recommended because of a number of mathematical flaws. A similarity index is widely applicable. In particular, it can be applied for purposes of anaylsis to comparison of pairs of presence-absence lists for the following (all of interest to this readers of this book): agarics in inventoried habitats, fungi available in markets, uses of fungi by groups of indigenous peoples, industrial organizations purchasing wild mushrooms from different geographic zones or different groups of people, etc.

The new similarity index is recommended to be used in place of all pre-existing similarity indices that have been examined.

Acknowledgments

I am grateful to Dr. Lee-Ann C. Hayek, Smithsonian Institution, Wash-

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

ington, for her careful and expeditious review of this chapter; to Dr. Mary Palm, U. S. National Fungus Collections, U. S. Department of Agriculture, Beltsville, Maryland; for encouraging me to write the chapter; and to Ms. Mary A. Tulloss, Roosevelt, New Jersey, for assistance in preparing the work for publication.

Literature Cited

Hayek, L.-A. C. 1994. Analysis of amphibian biodiversity data. Pp. 207-269. *In: Measuring and monitoring miological diversity. Standard methods for amphibians*. W. R. Heyer et al., eds. (Smithsonian Institution, Washington, D. C.).

Pankhurst, R. J. 1991. *Practical taxonomic computing*. Cambridge Univ. Press. xii+202 pp.

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

Appendix 1 Listing of BASIC Program Implementing Tripartite Similarity Index

```
 100   REM------------------------------------------------
 200   REM            TRIAL - TRIPARTITE SIMILARITY INDEX PROGRAM - TRIAL
 300   REM            Rodham E. Tulloss
 304   REM            P. O. Box 57, Roosevelt, NJ 08555-0057, U.S.A.
 310   REM            email: ret@njcc.com
 320   REM            fax: +1 609 426-4164
 400   REM            Original code: 18 June 1996
 500   REM            Most recent change: 06 September 1996, 9:00 p.m.
 550   REM            Vers. 0.6 [REMEMBER TO CHANGE VERS. NO. IN PRINT STMT.
 600   REM------------------------------------------------
 700   REM
 800           DEFSNG R-U
 900           DEFINT A-C, I-K, M-N, Y-Z
1000   REM
1100   REM------------------------------------------------
1200   REM  A is the number of spp. in common between populations 1 & 2.
1300   REM  C is the number of spp. in pop. 1 that are NOT in pop. 2.
1400   REM  B is the number of spp. in pop. 2 that are NOT in pop. 1.
1500   REM  The desired properties of the metric include the following:
1600   REM  A relatively high value of A should be rewarded.
1700   REM  If the unshared part of the smaller population is large relative
1800   REM    to A, there should be a punitive effect.
1900   REM  When the larger population is much greater than the smaller,
1950   REM    there should be a punitive effect.
2000   REM
2100   REM  The reward factor will be computed as the value R.
2200   REM  The first punitive factor will be computed as the value S.
2300   REM  The second punitive factor will be computed as the value U.
2400   REM
2500   REM  It is assumed that it is nonsense for either population
2600   REM  to have zero species in it.
2700   REM
2800   REM  It will be noted that the formula creating the "tripartite
2900   REM  similarity index (T)" has the following properties that seem
3000   REM  desirable:
3100   REM
3200   REM      When no species are shared, the value is zero.
3300   REM      If and only if the sets of spp. in the two populations are
3400   REM        identical, the value of the index is one.
3500   REM      All values of the metric lie between zero and one.
3600   REM      I believe that much of the undesirable aliasing seen
3700   REM        in other indices (very different situations generating
3800   REM        the same index value) has been avoided in the present
3900   REM        metric.
4000   REM
4100   REM  This metric is entirely heuristic.  The expression of the index
4200   REM  value as a percentage would be meaningless.
4300   REM
4400   REM------------------------------------------------
4500   REM
4600    PRINT "THIS PROGRAM GENERATES AN INDEX OF SIMILARITY FOR TWO
        POPULATIONS"
4700    PRINT ""
4800    PRINT "ENTER DATA ON TWO POPULATIONS AS THREE NUMBERS SEPARATED BY
        COMMAS"
4900    PRINT "PER LINE.  THERE SHOULD BE NO PUNCTUATION AT THE END OF THE
        LINE."
5000    PRINT "THE DATA ITEMS ON A SINGLE LINE ARE VALUES OF THE VARIABLES
        A, C, & B:"
5050    PRINT ""
5100    PRINT "     A (NO. OF ITEMS COMMON TO BOTH POPULATIONS)"
```

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

```
5200    PRINT "     C (NO. OF ITEMS IN POP. 1, NOT IN POP. 2)"
5300    PRINT "     B (NO. OF ITEMS IN POP. 2, NOT IN POP. 1)"
5325    PRINT ""
5350    PRINT "THE FIRST LINE OF THE INPUT FILE SHALL CONSIST OF AN INTEGER"
5375    PRINT "EQUAL TO THE NUMBER OF TRIPLES IN THE REMAINDER OF THE FILE."
5400    PRINT ""
5500    PRINT "THE INPUT FILE TO THIS PROGRAM MUST BE NAMED SIMINDEX.IN."
5600    PRINT "THE OUTPUT FILE OF THIS PROGRAM WILL BE NAMED SIMINDEX.OUT."
5700    PRINT ""
5720  REM----------------------------------------------------
5721  REM
5725  REM  CHECK FOR ZERO DIVIDE PROBLEMS IN INPUT DATA
5726  REM  Y IS A FLAG RECORDING THE DETECTING OF ANY SUCH PROBLEM.
5727  REM  Z IS A SIMILAR FLAG BUT IT IS RESET FOR EACH DATA TRIPLE
5728  REM  AND HAS ONLY LOCAL EFFECT.
5730  REM
5735  REM----------------------------------------------------
5750    OPEN "O", 2, "SIMINDEX.OUT"
5800    OPEN "I", 1, "SIMINDEX.IN"
5803    Y = 0
5805    INPUT #1, N
5810    FOR M = 1 TO N
5813      Z = 0
5815      INPUT #1, A, C, B
5820      IF A + C <= 0 THEN Z = 1 ELSE
5825      IF A + B <= 0 THEN Z = 1
5830      IF Z = 0 THEN GOTO 5890
5835      PRINT #2, "ERROR: LIST WITH LESS THAN ONE ELEMENT?"
5837      PRINT #2, "PROCESSING OF THE INPUT DATA SET WILL NOT OCCUR."
5840      PRINT #2, "INPUT TRIPLE #"; M; ", INPUT VALUES "; A; C; B
5850      Y = 1
5890    NEXT
5895    CLOSE #1
5900    IF Y = 1 THEN GOTO 9998
5910   REM---------------------------------------------------
5911   REM
5912   REM  ERROR CHECKING COMPLETED
5914   REM
5915   REM---------------------------------------------------
5920    OPEN "I", 1, "SIMINDEX.IN"
5930    PRINT #2, "TRIPARTITE SIMILARITY INDEX V. 0.6, "; DATE$; ", "; TIME$
5960    PRINT #2, ""
6000    A = 0
6100    C = 0
6200    B = 0
6225    INPUT #1, N
6300    FOR M = 1 TO N
6400      INPUT #1, A, C, B
6500   REM---------------------------------------------------
6600   REM      COMPUTATION OF THE REWARD FACTOR, R.
6700   REM---------------------------------------------------
6800      R = (LOG(1 + A / (C + A)) * LOG(1 + A / (A + B))) /
          (LOG(2) * LOG(2))
6900   REM---------------------------------------------------
7000   REM      COMPUTATION OF THE PUNITIVE FACTOR, S.
7030   REM
7040   REM              J = MAXIMUM(C,B)
7050   REM              K = MINIMUM(C,B)
7100   REM---------------------------------------------------
7200      IF C >= B THEN J = C ELSE J = B
7300      IF C >= B THEN K = B ELSE K = C
7400      S = 1 / (SQR(LOG(2 + K / (A + 1)) / LOG(2)))
7500   REM---------------------------------------------------
7600   REM      COMPUTATION OF THE PUNITIVE FACTOR, U.
7700   REM---------------------------------------------------
```

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

```
7800      U = LOG(1 + (K+A)/(J+A)) / LOG(2)
7900   REM-----------------------------------------------
8000   REM    COMPUTATION OF TRIPARTITE SIMILARITY INDEX (T)
8100   REM-----------------------------------------------
8200      T = SQR(R * S * U)
8300      PRINT #2, "TRIPARTITE SIMILARITY (T) FOR"
8400      PRINT #2, "SPECIES COMMON TO TWO POPULATIONS          = "; A
8500      PRINT #2, "SPECIES IN FIRST POPULATION, NOT IN SECOND = "; C
8600      PRINT #2, "SPECIES IN SECOND POPULATION, NOT IN FIRST = "; B
8700      PRINT #2, ""
8800      PRINT #2, "T = "; T
8900      PRINT #2, ""
9000    NEXT
9998    CLOSE
9999    STOP
```

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

Appendix 2: Input/Output Data Illustrating Near Invariance of the Tripartite Similarity Index

$T(1,1,1) = 0.55$
$T(2,2,2) = 0.54$
$T(5,5,5) = 0.53$
$T(10,10,10) = 0.53$
$T(50,50,50) = 0.52$
$T(100,100,100) = 0.52$
$T(1000,1000,1000) = 0.52$

$T(1,2,3) = 0.29$
$T(2,4,6) = 0.29$
$T(5,10,15) = 0.28$
$T(10,20,30) = 0.28$
$T(50,100,150) = 0.28$
$T(100,200,300) = 0.28$
$T(1000,2000,3000) = 0.28$

$T(7,5,11) = 0.44$
$T(35,25,55) = 0.44$
$T(70,50,110) = 0.44$
$T(140,100,220) = 0.44$
$T(700,500,1100) = 0.44$

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

Appendix 3: Input/Output Data Illustrating Near Linearity of the Tripartite Similarity Index with Regard to Variation in Size of the Set of Shared Entries

T(0,20,20)	= 0
T(1,19,19)	= 0.05
T(2,18,18)	= 0.10
T(3,17,17)	= 0.16
T(4,16,16)	= 0.21
T(5,15,15)	= 0.27
T(6,14,14)	= 0.32
T(7,13,13)	= 0.37
T(8,12,12)	= 0.42
T(9,11,11)	= 0.47
T(10,10,10)	= 0.53
T(11,9,9)	= 0.58
T(12,8,8)	= 0.62
T(13,7,7)	= 0.67
T(14,6,6)	= 0.72
T(15,5,5)	= 0.77
T(16,4,4)	= 0.82
T(17,3,3)	= 0.86
T(18,2,2)	= 0.91
T(19,1,1)	= 0.96

T(0,70,30)	= 0
T(5,65,25)	= 0.08
T(10,60,20)	= 0.17
T(15,55,15)	= 0.26
T(20,50,10)	= 0.35
T(25,45,5)	= 0.43
T(30,40,0)	= 0.51

T(0,125,66)	= 0
T(5,120,61)	= 0.04

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

$T(10,115,56)$	= 0.09
$T(15,110,51)$	= 0.14
$T(20,105,46)$	= 0.19
$T(25,100,41)$	= 0.23
$T(30,95,36)$	= 0.28
$T(35,90,31)$	= 0.33
$T(40,85,26)$	= 0.38
$T(45,80,21)$	= 0.42
$T(50,75,16)$	= 0.47
$T(55,70,11)$	= 0.51
$T(60,65,6)$	= 0.56
$T(65,60,1)$	= 0.60
$T(66,59,0)$	= 0.61

Offprint from Palm, M. E. and I. H. Chapela, eds. 1997. *Mycology in Sustainable Development: Expanding Concepts, Vanishing Borders*. (Parkway Publishers, Boone, North Carolina): 122-143.

Appendix 4: Input/Output Data Illustrating Variation of the Tripartite Similarity Index with Regard to Variation in Size of the Larger of Two Lists

$T(1,0,1)$	= 0.58	$\frac{1}{n}$	= 0.50	$\frac{1}{\sqrt{n}}$	= 0.71
$T(1,0,2)$	= 0.42		= 0.33		= 0.58
$T(1,0,3)$	= 0.32		= 0.25		= 0.50
$T(1,0,4)$	= 0.26		= 0.20		= 0.45
$T(1,0,5)$	= 0.22		= 0.17		= 0.41
$T(1,0,6)$	= 0.19		= 0.14		= 0.38
$T(1,0,7)$	= 0.17		= 0.12		= 0.35
$T(1,0,8)$	= 0.15		= 0.11		= 0.33
$T(1,0,9)$	= 0.14		= 0.10		= 0.32
$T(1,0,10)$	= 0.13		= 0.09		= 0.30
$T(1,0,20)$	= 0.07		= 0.05		= 0.22
$T(1,0,30)$	= 0.05		= 0.03		= 0.18
$T(1,0,40)$	= 0.03		= 0.02		= 0.16
$T(1,0,50)$	= 0.03		= 0.02		= 0.14
$T(1,0,60)$	= 0.02		= 0.02		= 0.13
$T(1,0,70)$	= 0.02		= 0.01		= 0.12
$T(1,0,80)$	= 0.02		= 0.01		= 0.11
$T(1,0,90)$	= 0.02		= 0.01		= 0.10
$T(1,0,100)$	= 0.01		= 0.01		= 0.10