COENOSES 3 (3): 121-133, 1988

MUTATIONAL AND NONMUTATIONAL SIMILARITY MEASURES: A PRELIMINARY EXAMINATION

M. Dale, CSIRO Division of Tropical Crops and Pastures, Carmody Rd., St. Lucia, 4067 Australia

Keywords. Similarity, model fitting, minimum mutation distance, standardisation, zero matching, structured data, partitioning, information, topology, synonymity, function, set intersection

Abstract. An examination of many of the indices proposed as numerical measures of pairwise similarity shows that they have strong relationships to string-to-string measures variously known as "Levenshtein distance", "longest common subsequence" or "minimal mutation distance". The variations among coefficients are created in several ways, including changing the set of operations, using a richer structural pattern, modifying weights, limiting the extent of operations and varying the basis for normalisation. In total these measures provide a very flexible means of assessing similarity and can be extended to similarities based on collections of strings. While not denying the interest to the user of other properties, such as metricity or embedding in a euclidean space, examining the coefficients as variations on the Levenshtein theme provides a common basis for their comparison and provides the user with a means of choosing between coefficients in a rational manner. But however interesting this array of coefficients might be, it remains true that only some features of similarity will be captured in a minimal mutational measure. These features may be more or less than are actually required by the user. In this paper I have made a preliminary examination of various measures, some of which are related to the Levenshtein metric, and some of which appear to capture other aspects of similarity (*i.e.* topological, functional, analogic and/or conceptual). These latter are all measures which I have been unable to relate to the Levenshtein distance, although I have not pursued this very far as yet. All measures were applied to vegetation data, classifying both plots and attributes into a two-way table. The SAHN algorithm has been used for most of the clusterings, so that differences between measures of similarity are the primary cause of differences in results. In a few cases other clustering algorithms have been used and the data has been converted to presence/absence when this was necessary with the particular coefficient.

Introduction

The calculation of some measure of similarity forms the basis of many, if not most, methods of discovering patterns, yet we know relatively little about the nature of similarity. For example, Tversky's (1977) account is titled "aspects" which suggests that he regards his treament as partial. Is similarity a single concept or several different ones? And if the latter how can these different aspects of similarity be identified? How do we choose which coefficient has the properties we desire, and indeed what are these desirable properties? Austin & Belbin (1982) extol asymmetric measures, whilst calculating a symmetric one, Blackburn (1980) needs to handle special data types with multiple values, Bykat (1979) requires measures easy to calculate when objects differ in known ways, in this case affine transforms of polygons while MacBratney (pers. comm.) is studying ways of classifying soil profiles whose horizons have been previously classified using fuzzy classification methods. It seems doubtful that a single model will fit all of these demands, yet it is highly desirable that some general approach be made available.

There are of course a very large number of similarity measures in the literature, and reviews of properties of subsets of this number, in various contexts, are available. These include: Cheetham and Hazel (1969), Sämdal (1974), Orlóci (1969, 1978), Lamont and Grant (1979), Hajdu (1981), Janson and Vegelius (1981), Burkea and Rao (1982), Rao (1982), Faith (1983), Faith (1984), Feoli, Lagonegro and Orlóci (1984) and Gower (1986). Sometimes the measures have been based on mathematical or statistical arguments, while in others ad hoc measures have been related to specific discipline-related theories, as with Mountford's (1971) use of assumed logarithmic species-area distributions in ecology, while other measures have been assumed to have general application. Sometimes distributional and sampling problems are addressed as in Wolds (1986).

As part of the review literature, there has also developed a small literature which discusses more formally the nature of similarity and the properties of various measures (c.f. Gower 1986, Rao 1982, Faith 1983). The objective in these studies is to provide a potential user with some rational means of choosing between the various alternatives without relying on empirical testing to establish the value of the measure. There is a need to be able to define similarity measures with particular properties, especially with data which do not conform to the usual property-value pairs. Sometimes we can obtain similarities directly by observing behaviour,

as in Lehman (1972), which of course avoids the observation of "properties". But similarity measures have been proposed for a great variety of situations which do not fit into such vector models.

The choice of measure must clearly interact with the properties of the data themselves. So some empirical study is likely to be needed before we can establish the efficacy of particular measures. In the present paper I shall seek to identify what seems to be a general model for a large number of coefficients of similarity, and then search for measures which do not seem to conform to the framework. Examples of applications of both kinds of coefficients, within and without the general class, will then be presented. The work is still incomplete and I am still asking questions rather than supplying answers.

Mutational measures

Lu and Fu (1978) considered the problem of classifying strings, that is ordered sequences of characters. Their solution was to employ a similarity measure based on the Levenshtein distance. In fact, an examination of the multitude of indices proposed as numerical measures of pairwise similarity shows that many have strong relationships to this string-to-string measure, known as "minimal mutation distance" and closely related to the "longest common subsequence". A general introduction to this class of measures is given by Sankoff and Kruskal (1983). The basic notion is quite simple and related to the child's game of converting one word into another; for example changing the word "amphiloby" into, say, "crocodile" The distance between two items, described in some specific manner, is the cost of the simplest series of transformations which convert (mutate) one item's description into the other's according to a set of transformation rules. In the child's game the only transformation permitted is the substitution of one letter for another, but there is also an implied semantic constraint in that all intermediate words must be meaningful. Such a transformation sequence is obviously related both to the notion of a grammar and to some kinds of expert system and seems to be a very generally applicable notion; c.f. Van Rijsbergen's (1986) comments on intensional logics.

As an example, if I restrict myself to using only insertion and deletion operations with equal weight, then the distance from the string ABBC to the string BBDD is 4; delete A and C, and insert DD. If I allow a substitution of D for C with unit weight then the distance becomes 3. The longest common subsequence, and in this case the longest common substring as well, is BB. The shortest common supersequence is ABBCDD, so I could derive a dissimilarity measure in the range 0/1 as

$$1 - \#(\text{BB}) / \#(\text{ABBCDD}) = 1\text{-}2/6 = 2/3$$

where $\#(A)$ is defined as the length of string A.

I have deliberately not identified what is meant by a description here because the notion of changing one object into another permits a much more general view of descriptions than is usually adopted in numerical classification. Instead of the usual vector of values associated with particular properties, I can use more complex structured data; all I have to do is use rules which describe how to change the structure. So problems such as classifying soil profiles or temporal sequences, classifying shapes or graphs, or classifying character state trees or logical dependency relationships, can all be handled within a single framework. Furthermore, if I can calculate a similarity measure, then I already have available algorithms to classify, ordinate, seriate or otherwise manipulate these values in a search for patterns. Note that the measure might well be asymmetric, such that the distance $A \Rightarrow B$ does not equal that $B \Rightarrow A$; some changes are easier in one direction than in the other. In most cases we would like our rules to reflect processes which could actually occur in the real world.

Given that such a distance measure can provide a useful measure of relationship between objects, then the variations among coefficients are created in various ways. Thus:

1. By modifying the *set of acceptable operations permitted for changing strings*. For example, in typewritten material I might allow transposition of symbols ($AB \Rightarrow BA$) to be regarded as a single operation additional to substitution, insertion and deletion thus reducing the costs of certain changes. Alternatively I might restrict the transformations so that only substitutions are permitted.

2. By introducing *more complex structures which have additional operations permitted*. If trees or geometric figures are used in place of strings then I have additional operations available which modify their structure. For example trees can have their branches rotated as shown in Čulik and Wood (1982). Using trees I can effectively deal with the problems of logical dependency addressed by Ben-Bassat and Zaidenberg (1984); using two-dimensional strings as does Moore (1979) I can deal with areas.

3. By altering the costs to be attached to such changes, often *using symbol-related weights or context-sensitive* weightings. This might involve asymmetric weighting of substitutions, such that $A \Rightarrow B \neq B \Rightarrow A$, or providing different weights for insertions or deletions at the ends of the strings compared to those in the body of the string. For example, Lu and Fu (1978) weight insertion of a symbol at the ends of a string differently to insertion in the middle of the string, Werman, Pelg and Rosenfeld (1985) differentially weight substitutions of pairs of symbols, whilst Feoli and Lagonegro (1983) weight both differences and similarities in the strings,

i.e. the substitution of B by B is given a (negative) weight.

4. The method of calculating the cost of changes does not presuppose one-to-one matching of the descriptions. The usual algorithm employed also aligns the descriptions for best fit, introducing gaps so as to minimise the mismatching. However I can variously limit the *extent of application* of changes so that only a subset of the possible alignments is actually examined. In practice, instead of using the complete tableau associated with the dynamic programming solution I might *permit only some paths to be traversed*, or I may permit various "timewarping" techniques to allow *elastic matching*.

5. If gaps are permitted in this alignment, then the *weight assigned to a gap* may reflect its existence independent of length, or be length, and maybe also content dependent.

6. By employing various *normalisation techniques* I may bring the similarity values into some acceptable range, such as 0-1. Thus I might use the length of the shortest common supersequence as a normalising factor, the sum of the lengths of the two strings being compared, the length of the longer string or the size of the universe of comparison. Choice of normalisation factor seems to be one of the major sources of variation between coefficients.

While these 6 options are perhaps the commonest means of modifying the similarity measure to fit its appointed task, they do not by any means exhaust the possibilities. I can introduce, as does the child's game mentioned earlier, extra semantic constraints requiring that the intermediates between the two items have specified properties, or that the string be mapped to specific values, a notion akin to affixes for grammars. This, of course, involves us in specifying how the interpretation is understood or how the mapping valuation is to be made. I can introduce rules which are context sensitive, so that the permissible changes for any given symbol depend on what other symbols are present in a specified context. I might specify which other rules are permitted to follow if some change is made, or if it is not. I can apply rules sequentially, one rule at a time, in parallel, all changes of one kind being made simultaneously, or in combinations of these. The Levenshtein measures can also be extended to similarities based on collections of strings, (c.f. Lemone 1982, Kashyap and Oommen 1983b); such extension is required in many divisive algorithms which require a measure of change in fit of a model.

Examining coefficients as variations on this Levenshtein theme provides a common basis for their comparison, providing the user with a means of choosing options in a rational manner. Such considerations are especially important when developing similarity measures appropriate for mixed data, as in Lerman and Peter (1985). However, this does not deny the interest to the user of other properties of the coefficients, such as metricity or possibility of embedding in a euclidean space. Yet, however interesting this array of coefficients might be, such minimal mutational measures capture only some aspects of similarity and these aspects may be more or less than is actually required by the user. My suggestion is that there are indeed other kinds of measure, so that similarity is not a single concept. But what alternatives are there and can we identify any particular properties associated with them which might be of significance to our choice? Since I am not primarily a mathematician I have not sought to obtain a theoretical framework, which I suspect exists, but have adopted a frankly empirical evaluative approach.

Data and assessment

I started by examining a rather simple set of ecological data, due to Bowman & Wilson (1986). The data themselves were coded cover estimates for species from two areas in the Adelaide River flood plain, in Northern Australia. Each area was sampled so as to include a range of the vegetation variation. In Table 1, I give the order of plots and species from the original paper, obtained using a reciprocal averaging (correspondance analysis) seriation.

Table 1. Original gradient sequences for plots and species.

Plots

30 34 28 31 24 26 36 37 41 39 27 32 33 40...
20 6 18 12 39 19 11 23 25 22 21 35 38...
17 5 8 13 4 16 14 2 15 1 7 3 10 9

Species

1. *Cyperus rotundus*	2. *Abelmoschus ficulnens*	3. *Iponoea coptica*
4. *Cynodon arcuatus*	5. *Merremia hederacea*	6. *Alysicarpus vaginalis*
7. *Panicum cambodiencee*	8. *Abelmoschus moschatus*	9. *Melochia corchorifolia*
10. *Waltheria indica*	11. *Ludwigia octoyalys*	12. *Poaceae Sps. 1*
13. *Heliotropum crispatum*	14. *Euphorbia vachellii*	15. *Echinochloa colona*
16. *Phyla nodiflora*	17. *Paspalum scrobiculatum*	18. *Echinochloa elliptica*
19. *Poaceae Sps. 2*	20. *Phyllanthus Sps.*	21. *Goodenia purpurescens*
22. *Cardiospennum halicacabum*	23. *Sesbania Sps.*	24. *Heliotropum indicum*
25. *Dentalla dioeca*	26. *Ipomoea aquatica*	27. *Oryza Sps.*
28. *Cassia obtusifolia*	29. *Eleocharis Sps.*	30. *Pseudoraphis spinescens*
31. *Ludwigia adscendene*	32. *Polygnunon attenuatum*	33. *Aeschynomene indica*

I then examined various measures of similarity, some of which are related to the Levenshtein metric, and

some of which appear to capture *other* aspects of similarity, topological, functional, analogic and/or conceptual. These latter are all measures which I have so far been unable to relate to the Levenshtein distance, although I have not pursued this very energetically in measure space. Similarity measures were calculated for both plots and attributes, and then classified, mostly using the SAHN (Sequential Aglomerative Hierarchical Non-overlapping) algorithm, to illustrate the differences obtained. Several other methods of classification also use similarity measures directly, for example Korhonen's (1984) self-organising classification-ordination method, and Lance and Williams' (1968) general divisive method. However I am not especially concerned here with the procedure for group forming per se.

The data can be presented in a two way table to show the cross-classification, and this might be assessed in various ways, primarily forms of "nodal analysis" sensu Lambert and Williams (1962), as extended by Dale (1964). Alternatively the work of Juhász-Nagy (1984) with his associatum concepts would allow detailed assessment; these are presently in progress. Obviously using Levenshtein measures I could also compare the hierarchies directly, though there are some interesting problems about weighting internal nodes. For the moment I have stayed with the partitions into groups which can be cross-compared between different similarity measures and for this purpose I have used Rajski's (1961) metric whose utility in community studies has been shown by Orlóci (1969) and Feoli, Lagonegro and Orlóci (1984). This gives us a similarity matrix between results for each of the 16 similarity measures, Table 9, and this can be displayed as an ordination using Gower's (1966) principal coordinates analysis. To provide an alternative view, I have also used Sattath and Tversky's (1977) additive similarity tree, Figure 3, approach, and Arabie and Carroll's (1980) additive clustering, Table 10.

Table 2. Data table sorted using eident values

Group Labels	α	β	χ	δ	ε
Species	21	.3.22.1312.	13122	31...	.122.12211
	76	4073918 1049	42526	37536	8985210123
Plots					
A					
27	..	..6.....	1.1..	..2..	12.....
37	..	4.12..2...2	1...1	..1..	.....
20	14	....1.1..1.	..1..	2...2	11
23	34	..1.....21	1..3.	11..1	.....
16	3.	.2.13..12..	...21	.....	.....
34	..	4..1.3....1	1.1..	.....	....1.....
31	..	4..1.2..2.1	..1..	...1.	.....
32	..	..41..1.1.2	1....	.....	.....1....
5	4.	...32..112.	1.1..	.....	.....
10	1.	.5..2..2.1.	.....	.....	.....
9	2.	.....1.1.	1	.....	.....
B					
18	.5	....2.2..21	...2.	.....	.....
30	..	..11..4.111	.....	..11.	.....
19	15	..11..2..11	1	1...1	.1.....
26	.3	...1...4.1..	1.1..	.....	.....
15	4.	.2.....1....	.3...	.....	.....
38	4.	..1.1.1...1	2.1..	.....	.....11..
29	.3	1.11.1..411	.....	...1.	.....
11	.3	.1..1.3..11	..1.1	1....	.....
6	.4	...21.2..11	.....	1....	.....
12	.5	.1..1.1..11	...11	.....	.....
14	4.	.3..1..1.11	1	.....	.....
36	.2	4.1.1.1..11	1...1	1....	.....
7	2.	.4.12.11...	.....	.....	.....
33	.4	..11.1....2	.....	.....	1.....
41	2.	3.3...1...1	1....	.....	.....
40	15	....1.....1	1.1..	.1...	.....
13	5.	...1..1111.	..1..	.....	.....
3	2.	.1.....4...	..1.1	.....	.....
22	43	..1.....11	1.1..	.1...	.....
2	5.	.1..1..1..1	.1...	.....	.....
C					
25	34	..1.1....1.	.....	.1...	.....
1	5.	....2..1...	.....	.....	.....
17	3.	...3...1..1	1	.....	.....
8	2.	...2...111.	...1.	.....	..1.....
4	5.	...1...1.1.	.....	.....	.....
28	..	1.11.1..1.1	1.1..	...1.	.....
35	2.	..1...1..21	1.1..	.....	...1.....
24	.1	3.11.....1	1....	.1...	.....
21	..	...31....1.	.....	.....	.....
39	.2	...11....11	.....	.....	.....

Similarity measures

In all 15 dissimilarity measures were employed to generate groups, with a sixteenth analysis which does not strictly involve a dissimilarity measure directly. I shall briefly describe the coefficients in the following but I shall not include the specific formulae permitting their calculation, which can obtained from the original publications. The serial order in which the coefficients are presented is used in some diagrams, and I shall also list the abbreviations used for each coefficient in other analyses.

The *Eid* results were based on a method of selecting "interesting" species using the eident values of Dale and Williams (1978) derived from the two-parameter method of Dale and Anderson (1973). This is the "odd method out" in that it does not rely on a dissimilarity measure directly; eident values are sums of deviations from expected values calculated for the entire data matrix using the two-parameter model. Since the species and plots are ordered in terms of their interest, I simply chopped these orderings wherever there seemed to be a significant break. Thus the groups formed do not necessarily represent homogeneous collections. The present groups should not be expected to resemble the groups found by any other analysis.

The next coefficient, *Syn*, is based on the notion of synonymity, or functional similarity, first treated by Lewis, Baxendale and Bennet (1967) and applied by Dale, Clifford and Ross (1984). Linguistically the notion of a synonym involves the ability of one word to replace another in a similar context, i.e. to function equivalently. Ecologically this is interesting because if two species fulfil the same functional role then they should be equivalently synonymous, although there is also a possibility that one may be a more specialised version

of the other rather than a true replacement. However in the case of synonymy the two words will not co-occur except in special cases such as dictionaries. The coefficient therefore identifies items as similar if they occur in a similar context but do NOT occur with each other.

As a measure of topological distance I have used Calhoun's coefficient, *Cal*, as described by Bartels, Bahr, Calhoun and Wied (1970). For two items, A and B, we simply count the number of other items which lie between them. For multidimensional data the original definition accepted "betweenness" defined as follows:

for three individuals x, y, z, y is "between" x and z *iff*

$\exists j, x_j < y_j < z_j \mid z_j < y_j < x_j$ where j is some variable.

However, tied values, where $(x_j = y_j)$ and $(y_j \neq z_j)$ or $(y_j = z_j)$ and $(x_j \neq y_j)$, and doubly tied values where $x_j = y_j = z_j$, still contributed although downweighted, so the "betweenness" was not strict.

Minimum discrimination information statistics (Kullback 1959) and other related information measures (*e.g.* Orlóci 1978, Feoli *et al.* 1984) have provided several useful measures of dissimilarity, for example the diversity information, *Div*, of Dale and Anderson (1972). Here we measure the change in diversity induced by fusing two items, and while this can be related to Levenshtein measures, the constraints and the complexity of weighting are rather extreme. As an alternative we have also employed the partitionable information measure of Williams (1973) which combines quantitative and qualitative elements of the values. This is referenced as the *Tinf* coefficient but since some coefficients can only be applied to qualitative data it seemed sensible to also use the qualitative component alone, here called *Qinf* as a basis for comparison.

Not all approaches to clustering emphasise the homogeneity of the clusters over all other aspects. There is much to be said for clusters whose definition is simple; in particular Micalaski and Stepp (1985) have argued for the conceptual import of simplicity in the context of expert systems. Monothetic methods provide typical examples of these but in the present study I have used a polythetic approach. Le Quense (1974) in examining cladistic compatibility approaches noted that, if homeoplasy was to be avoided then, for two binary characters, at least one possible combination of states should not exist. However a more general approach using predicate calculus was developed by Vesely (1981). Dealing with binary data he sought to be able to describe clusters using simple formulae in the predicate calculus. He noted that for two binary characters, there are obviously only 4 possible combinations; 00, 01, 10, 11. If in a group of items only one of these combinations occurs, irrespective of which, we have a very simple description in terms of intersections; if 2 combinations of them occur then we must include difference sets, for 3 combinations a union operation

Table 3. Data table sorted using synonymity measure.

Group Labels	α	β	χ	δ	ε	φ
Species	11111222223	1	11	11222	2	2333
	467904568346793	123581	23	79015	8	2012
Plots						
A						
15	.....1..	.....	..	.....	.	.111
B						
21	.....1.11.	.....	..	.1...	.	....
10	.....1.11.	.....	..	.....	.	.11.
9	.....111..	.....	..	.....	.	..11
1	.....11.	.....	..	.....	.	..1.
3	.....1....11..	.....	..	.....	.	.11.
C						
4	....1....1..1..	.....	..	.....	.	..1.
36	1.11.1.11111..1	.....	..	.....	.	....
41	1.11.1..1...1..	.....	..	.....	.	....
40	...1.111....11.	.....	..	1....	.	....
20	.1....111.1.111	.....	11	.....	.	....
12	...1...11.11.1.	.....	..	.....	.	11..
19	.111...11111111	.....	..	.1...	.	....
11	...1..111.11.11	.....	..	.....	.	.1..
23	.111.1.1..1.1.1	.....	..	1....	.	1...
25	..1....1..1.11.	.....	..	1....	.	....
22	..11.111..1.1..	.....	..	1....	.	....
35	..11.11.1.1.1..	.....	..	1	.	....
38	..11.11.1...11.	.....	..	..11.	.	....
17	...1.....1.11..	.....	..	.....	.	..1.
5	....1.1..11.11.	.....	..	.....	.	..1.
8	....1....11.1..	.....	..	.....	1	1.1.
13	....1.1.111.1..	.....	..	.....	.	..1.
16	....1....1.111.	.....	..	.....	.	111.
14	...1.....1111.	.....	..	.....	.	.11.
2	...1.....11.	.....	..	.....	.	.111
7	.....11..11.	.....	..	.....	.	.11.
D						
27	..11.11.....	...11.	..	.....	.	....
30	..1111...1.....	1.11..	..	.....	.	....
34	1..1.11..1.....	11....	..	.....	.	....
28	1.11111..1.....	1.1...	..	.....	.	....
31	1..11.1..1.....	1.1...	..	.....	.	....
24	1.11.1.1.1.....	.....	..	1....	.	....
26	....1111.1.....	1.....	..	.....	.	....
37	1.11.1..11.1...	...1..	..	.....	.	....
29	1.111..1.11....	1.1...	..	.....	.	....
32	..1111..11.....	1	..	.....	.	....
33	..11...1.1.....	1...1.	..	.....	.	....
6	...1...1111..11	.....	..	.....	.	....
18	...1...11.1..1.	.....	..	.....	.	1...
39	...1...1.11..1.	.....	..	.....	.	....

is needed, and finally if all 4 occur no simple description is possible. It is fairly easy to convert this into a similarity measure, *Pred*, which can be used in the standard SAHN algorithm so that we fuse items to keep maximal simplicity. This need not be equivalent to homogeneity of any single attribute.

Because of its convenient relationship to variance, the Standardised Euclidean distance, *SED*, is of some interest. A recent study involving this measure is that of Gambarov, Mandel and Rybina (1980), but Gower (1986) has given a broad review of the advantages of metricity and of a euclidean space. However ecologists have also long been interested in metrics of various kinds, for example the local set intersection Canberra metric with, or without, double-zero suppression (Lance and Williams 1967), coefficients *COO* and *Can* respectively or the earlier Bray-Curtis set intersection measure, *BC*, apparently first used by Czekanowski (1909). This is also related to proposals for "fuzzy intersection" *(FI)* measures (c.f. Levandowsky 1972, Miyamoto and Nakayama 1986). In fact both Bray-Curtis and Fuzzy intersection are simple Levenshtein distances if the data are coded in a particular way; this is NOT the only way possible but is reasonably intuitive. Werman, Pelg and

Rosenfeld (1985) suggest that if we have frequency data then we can convert it to a string form. The method is simply to label each attribute with some symbol, traditionally alphabetic, and then simply generate a string by taking the attributes in a fixed order and, for a frequency of n, simply repeating the symbol n times. They further point out that this transformation is also applicable if the attributes are related in some way. Their example is a (spatial) transition matrix. The coding into a string is performed in the same way as before with each cell of the transition matrix being denoted by a different symbol, and a specific ordering of cells being predetermined to allow a string representation. However the manner of calculating the Levenshtein distance is now changed to reflect the inter-relationships of the attributes, weighting transitions between adjacent cells less costly than transitions between more distant cells.

Table 4. Data table sorted using Calhoun's measure of topological similarity.

Group Labels	α	β	χ	δ	ε	φ
Species	2333	1111111222223	.	...11222	1	2
	9012	23568123457901 5683	4	17908234	6	7
Plots						
A						
41	....	.....1.....	3	.31.1...	.	2
27	....	..2.1...11.....	.	.62.....	.	.
38	1...	.....21..11....	.	.11.1...	.	4
30	....	.11.....1.....	.	4111..1.	.	.
34	....	1.....11.....	4	3.1...1.	.	.
28	....	.1.....11.....	1	1111..1.	.	.
31	....	.1.....1.....	4	2.12..1.	.	.
24	....	.....1.1.....	3	.11...1.	1	.
26	....	.....11.....	.	4..1..1.	3	.
36	....	.....1.....1.1	4	.11.1.11	2	.
37	....	..1.....1.....1..	4	.12.2.2.	.	.
29	....	.1.....	1	1114..11	3	.
32	....	.....1..1.....	.	.4212.1.	.	.
6	1...	1	.	..1.2.21	4	.
39	1...	.....	.	..1...11	2	.
17	..1.	.....1..	.	..1...3.	.	3
5	2.1.	.....1.....	.	...1..32	.	4
8	..1.	.....1.	.	...1.121	.	2
B						
33	....	....1.....	.	112...1.	4	.
40	1...	.....111.....	.	..1.....	5	1
20	1..	...2..11.1.....2	.	....1..1	4	1
18	1...	.....	.	..1.22.1	5	.
12	11..	.....1..	.	..1.11.1	5	.
19	1...	...1.....1...1.1	.	.11.2.11	5	1
11	11..	.....1.....1.1	.	..1.3..1	3	.
C						
23	....	...1....1.1.....1	.	.11..3.2	4	3
25	1...	.....1.....	.	.1.....1	4	3
22	....	.....111.....	.	.11....1	3	4
21	1...	.....1.....	.	1	.	3
35	....	.....11....1...	.	.11.1..2	.	2
9	..16	.....1..		1	.	2
16	321.	.....1..	.	...2.21.	.	3
7	241.	.....	.	....1.1.	.	2
10	252.	.....	.	1	.	1
13	..1.	.....1.....	.	...11.11	.	5
4	..1.	.....	.	...1..1.	.	5
14	131.	.....1..	.	..1....1	.	4
2	1111	.....	.	..1.....	.	5
15	.213	.....	.	.....	.	4
1	2.1.	.....	.	.....	.	5
3	.14.	.....1.....1..	.	.....	.	2

I have made this transformation and used a basic insertion/deletion form of the Levenshtein measure, *Lev*. However the Levenshtein measure, as noted earlier, is very versatile. It can be used for the logical dependency problem addressed by Ben-Bassat and Zaidenberg (1984). Čulik and Wood (1982) have proposed a measure suitable for trees, and there are others which might have relevance to the problems of consensus tree formation. Kashyap and Oommen (1983a) give a simple measure for 2 strings, but later extended it to cover several strings (Kashyap and Oommen 1983b). Klopman and Macina (1985) have used this approach with chemical structure data in an attempt to relate structure with functional activity, while Littleand Ross (1985) have applied it to with soil profiles. Probably the best general introduction is that of Sankoff and Kruskal (1983) but even this is by no means exhaustive.

At present one of the favourite measures of similarity among ecologists is certainly the chi-square metric used by Hill, Bunce and Shaw (1975). Rather than calculate the chi-square metric directly I have employed the TWINSPAN program itself to obtain the results for

Table 5. Data table sorted using total information.

Group Label	α	β	χ	δ	ε	φ
Species	111112222233	1122	23	23	..1	...112
	2568123790125823	6849	60	71	794	134053
Plots						
A						
17	.....	....	1.	31	.1.	3
14	.....	..11	13	41	.1.	.....
2	.....1.	...1	.1	51	.1.	.....
15	.....3.	....	.2	41	...	.....
3	.....	....	11	24	...	....1.
10	.....	..12	.5	12	...	.....
9	.....6.	..1.	1.	21	...	.....
5	.....	..22	..	41	...	...113
8	.....1.1..	..1.	..	21	...	...1.2
13	.....	.11.	..	51	...	...111
4	.....	....	..	51	...	...1.1
16	.....2....	...3	12	31	...	...2.1
1	.....	...2	..	51	...	.....
7	.....	.1.2	.4	21	...	1
B						
20	..2..11.....2	4111	..	1.	...	....1.
6	1	4211	..	..	.1.	2
18	.....2....	5211	..	..	.1.	.....
12	.....1....	5111	11	..	.1.	.....
39	.....	2.11	..	..	.1.	1
19	..1.....1.....1	5211	1.	1.	11.	1
11	1	3311	11	..	.1.	....1.
41	.....	.1..	..	2.	311	..3...
35	.....1....	.12.	..	2.	111	....1.
38	.....11.....	.1.1	..	4.	112	....1.
40	.....1.....	5..1	..	1.	.11	....1.
23	..1....1...3...1	4.2.	..	3.	111	.....
25	.....1.....	4.11	..	3.	1..	.....
22	.....1.....	3.1.	..	4.	111	....1.
21	.....1.....	..11	..	3.	...	.....
C						
34	1.....	....	..	..	.11	3.4.11
28	.....	....	..	..	111	111111
31	.....	....	..	..	.1.	214211
26	.....	3...	..	..	..1	4..111
29	.....	3.1.	..	..	11.	1114.1
33	...1.....	4...	..	..	12.	1....1
30	.1.....	....	..	..	111	41.1.1
24	.....1.....	1...	..	..	111	..3..1
36	1	211.	1.	..	111	..4..1
37	.1.....	.2..	1.	..	121	..4..2
27	.2.1.....	....	..	..	621	....1.
32	....1.....	.2..	..	..	421	...1.1

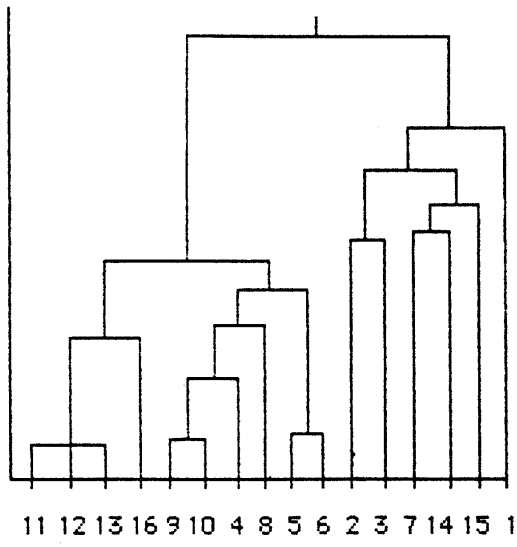

Fig. 1. Classification of measures. SAHN, flexible sorting. Key to Items: 1. Eid; 2. Syn; 3. Cal; 4. Div; 5. Tinf; 6. Qinf; 7. Pred; 8. SED; 9. COD; 10. Can; 11. BC; 12. FI; 13. Lev; 14. Chi; 15. FC1; 16. CCn.

measure *Chi*. However the measure is clearly related to a euclidean distance if the data are normalised in the right way, but the normalisation is global and depends on row and column totals in the data matrix. I have assumed that, like the *SED* this measure is a very constrained version of the Levenshtein measure, with only substitution permitted and with restrictions on the comparisons permitted as well.

I noted earlier that Calhoun's distance counted an item as "between" if any attribute was acceptable. In a modification which I call "fuzzy calhoun", *FC1*, I have used the proportion of attributes which indicated betweenness rather than solely the existence of such an attribute. Obviously some strong relationship to original Calhoun measure is to be expected, but I am unclear as to the exact nature of the resulting fuzzy coefficient.

Finally there have been various proposals which seek to combine similarity and dissimilarity as distinct components in a single measure, as in Tversky (1977). Feoli and Lagonegro (1983) weighted various contributions from a 2×2 contingency table differentially, while Faith(1984) proposed his Ccont coefficient to combine aspects of other coefficients and further work is reported by Day and Faith (1986). I have used Faith's original suggestion for coefficient *Ccn*. These proposals are related to the simplicity criteria used in the predicate calculus approach discussed earlier due to Vesely (1981), but use the information in a different way.

Table 6. Data Table sorted using Predicate Calculus simplicity.

Group Label	α	β	χ	δ	ε	φ
Species	1222	..1	112	.11112233	2233	1122
	1234581015	794	053	623792823	6701	6849
Plots						
A						
30	1.1.1.....	111	1.1	.....	....	....
34	11.1.....	.11	.11	.....	....	....
28	1.11.....	111	111	.....	....	....
31	1.11.....	.1.	111	.....	....	....
26	1.....	..1	111	.....	....	1...
29	1.11.....	11.	1.1	.....	....	1.1.
24	...1.....	111	..1	...1.....	....	1...
36	...1.....	111	..1	1	1...	111.
37	...11.....	111	..1	.....	1...	.1..
33	1....1....	11.	..1	.....	....	1...
B						
41	...1.....	111	...	.....	.1..	.1..
27	....11....	111	.1.	.....	....	....
32	.....1...	111	1.1	.....	....	.1..
35	1	111	.1.	.....	.1..	.11.
38	.....11.	111	.1.	.....	.1..	.1.1
C						
40	.....	.11	.1.	...1.....	.1..	1..1
23	.....	111	...	1..1.1..1	.1..	1.1.
25	.....	1..	...	...1.....	.1..	1.11
22	.....	111	.1.	...1.....	.1..	1.1.
20	.....	...	.1.	111.....1	.1..	1111
21	.....	...	...	....1....	.1..	..11
1	.....	...	...	.....	.1.1	...1
D						
6	.....	.1.	..1	1	....	1111
18	.....	.1.	...	.....1...	....	1111
12	.....	.1.	...	.....1...	1.1.	1111
39	.....	.1.	..1	.....	....	1.11
19	.....	11.	..1	1...1...1	11..	1111
11	.....	.1.	.1.	1	1.1.	1111
17	.....	.1.	..1	.....	11.1	....
5	.....	...	111	.....	.1.1	..11
8	.....	...	1.1	.....11..	.1.1	..1.
13	.....	...	111	.....	.1.1	.11.
4	.....	...	1.1	.....	.1.1	....
16	.....	...	1.1	.....1...	1111	...1
14	.....	.1.	...	.....	1111	..11
2	.....	.1.	...	.....1.	.111	...1
15	.....	...	...	.....1.	.111	....
7	.....	...	..1	.....	.111	.1.1
3	.....	...	.1.	.....	1111	....
10	.....	...	...	.....	.111	..11
9	.....	...	...	.....1.	11.1	..1.

Assessment of results

To examine the results I have used two approaches. First, within a single table generated by plot and species classification using one coefficient, I have a cross-classification in which various features can be observed. This could be examined using the nodal analysis techniques of Lambert and Williams (1962), (see also Dale 1964, for some extensions) or Juhász-Nagy (1984), to examine how far the plot and species classifications are coincident for a given coefficient. Most such methods are unfortunately restricted to binary data, which is inapplicable unless I adopt some such coding of the frequency data as is implicit in the Levenshtein approach anyway. The approach is really concerned with the interpretability of the data, and rests on the ecological significance of species as a means of assigning environmental meanings to plot clusters. This requires information which is sadly lacking for most species in

this study, although the existence of coincidences between plot and species classifications is still noteworthy. I have in fact looked at some of these assessment methods, Dale's (1964) extensions of Lambert and Williams (1962) nodal analysis, Buser's (1983) homogeneities and Juhász-Nagy's (1984) associatum but it is not clear that they provide much more information than a simple visual inspection and in any case most of the methods are restricted to presence data only. For these reasons they have been omitted here.

The second alternative approach is to examine the coincidence of plot or species classifications between the 16 coefficients used here. This would involve something like the Fowlkes-Mallowes (1983) statistic to test for significant agreement. Since some sort of comparison of results is the intention here, this is the approach I have emphasised. I have used Rajski's (1961) distance between partitions in a contingency table to organise the methods because I do not think the significance testing is really very helpful given the biased nature of the Fowlkes-Mallowes and the competing Rand statistics, and their somewhat peculiar properties.

Results

The ordering of plots and species from the original analysis is given in Table 1 to allow crossreference. The Table 2-8 show the two-way tables for a selection of the coefficients. These tables show 6 (5 for *Eid*) attribute groups, labelled α, β, χ, δ, ϵ, ϕ, as columns and various numbers of plot groups, labelled A, B, C... Absence of an attribute value is indicated by a period, otherwise the coded cover value is recorded, except for the *Syn* and *Pred* measures where only presence/absence information is given. From the tables the cross relationships between attribute clusters and item clu-

Table 7. Data table sorted using Levenshtein/Bray-Curtis/Fuzzy intersection measure.

Group Labels	α	β	χ	δ	ε	φ
Species	22 01	11 23	.13 693	.....1112 13457890453	11222233 68467901	.112223 2172582
Plots						
A						
17	..	..	...	.....1...3	...13..1	.....
5	..	..	...	.....1.13	..2.42.1	.....
8	..	..	...	.....1..2	..1.2..1	...2.1.
13	..	..	...	.....1.11	..1.5..1	.....
4	..	..	...	.....1..1	....5..1	.....
21	..	..	.1.	.....	..1.31..	.....
16	..	..	...	.....2..1	...13321	...2...
14	..	..	...	.....1....	..114131	.....
2	..	..	...	.....1....	5111	1
15	..	..	...	.....	....4.21	3
1	..	..	...	.....	....52.1	.....
7	..	..	...	1	.1..2241	.....
3	..	..	...	.....1.	...12.14	.....
10	..	..	...	.....	..1.1252	.....
9	..	..	...	.....	..112..1	6
B						
29	..	..	...	111.1.14..1	3.1.....	.....
33	..	..	...	1...112...1	4.....	.....
39	..	..	...	.....1...1	2.1..1..	.....
40	..	..	...	.....1.11.	5...11..	..1....
20	..	11	2.2	.....1.	411.11..	.....
6	..	..	..1	.....1...2	421..1..	.....
18	..	..	...	.....1....	521..1..	...2...
12	..	..	...	.....1....	5111.11.	...1...
19	..	..	111	....1.1...1	521111..	.....
11	..	..	..1	.....1..1.	3311.11.	.....
23	..	..	1.1	....1.1.1..	4.2..3..	..13...
25	..	..	...	....1.....	4.1.31..	..1....
22	..	..	...	....1.1.11.	3.1.41..	..1....
C						
30	..	..	...	41.11.1.111	.....	1.....
34	..	..	...	3.4...1.111	.....	.....
28	..	..	...	111.1.11111	.....	.....
31	..	..	...	214...12.11	.....	.....
26	..	..	...	4.....1111	3.....	.....
35	..	..	...	....1.1.11.	.12.2...	....1...
38	11	..	...	....1.1.21.	.1..41..	.....
24	..	..	...	..3.11..1.1	1.....	..1....
36	..	..	..1	..411.1.1.1	2111....	.....
37	..	..	...	..4112..1.2	.2.1....	.....
41	..	..	...	..3.3.1.1..	.1..2.....	
27	..	..	...	...2612.11.	.....	.....
32	..	..	...	....4.211.1	.2.....	.1.....

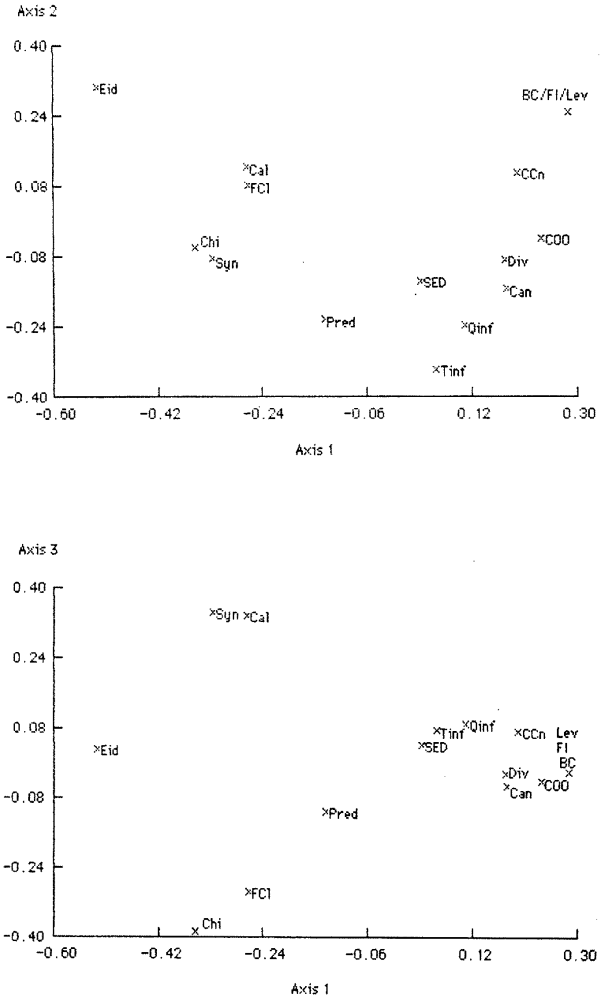

Fig. 2. Principal coordinates analysis: axes pair 1/2 and 1/3 plotted.

sters are visible, and also the tendency of any particular analysis to produce groups with one or two members only. However we should not be misled into believing that a "minimal variance" structure with each item/attribute intersection block filled with identical values, is necessarily the ideal result, however much it has attracted statisticians in the past . Other structure can be more valuable.

To aid in visualising the relationships between the various coefficients, I have analysed the Rajski distances, given in Table 9 classifying using SAHN, Fig. 1, and also providing a principal coordinates analysis (Gower 1966), Fig. 2. On a %-trace basis the three ordination axes account for about 46%, which is not too bad.

I also identified an additive similarity tree, Fig. 3, using the methods of Sattath and Tversky (1977). The Kruskal's stress measure for the fit of the additive tree is about 32%, a rather bad fit, which is not really unexpected, since it is unrealistic to expect the coefficients to fall into a simple hierarchy. Finally, I have also employed Arabie and Carroll's (1980) additive clustering. Table 10, because the notion of parts of similarity being shared in different ways among the different coefficients seems very sensible. The main problem here, apart from computational cost which will become negligible when our Cyber 205 version is working, is choosing the number of clusters. The method allows more clusters than items, and as yet I have had insufficient experience with the method to have developed any heuristics or hunches.

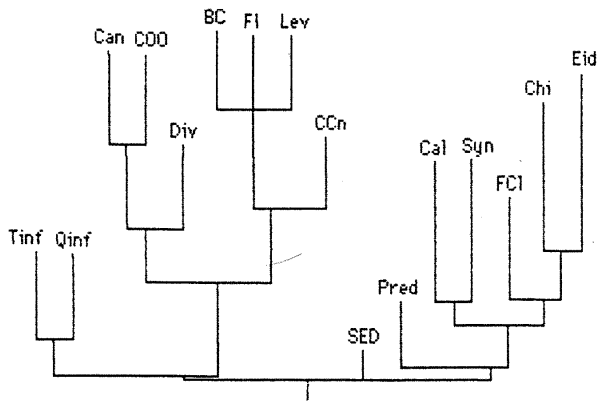

Fig. 3. Additive similarity tree (only vertical distances are significant).

The two-way tables can be briefly dealt with, with interest focussed on the existence of clear "blocking" where plot and species groups coincide, and the tendency to form small groups of 1 or 2 members. The *Eid* and *Chi* results seem to lack strong blocking, the seriation tendency being obvious in the latter, but blocking is apparent in the other analyses. However this is confounded with the tendency to form small groups; such groups are found more among the species than the plots partly because there are more groups permitted and partly because species are inherently unique items. Such small groups, then, seem to be associated either with the very rare species, notably 12 and 13 and 20 and 21, or with the very common, and potentially important species such as 27 (which would be the division species in a two-parameter analysis), 16 or 4. Obviously many of the measures closely linked to the Levenshtein measure give groups which are very similar. However the *Tinf* and *Syn* measures both show

Table 8. Data table sorted using Chi-square metric.

Group Label	α	β	χ	δ	ε	φ
Species	11	.111112223	12	122	1222	2333
	1234578914	6236780153	54	023	9679	8012
Plots						
A						
8	...411.2.1	.....2....	..	..2	.1..	....
9	...3.3.1.1	.....1....	..	...	..2.	....
11	....2612.1	....	1.	...	....	....
12	.....4.211	.....2....	..	1.1	....	....
1	4.1.11.1.1	....	..	1.1	....	....
2	31.4...1.1	....	1.	1.1	....	....
3	1.11.1.1.1	....	1.	1.1	....	....
4	2.14...1..	....	1.	2.1	....	....
5	...3.1.1.1	...11.....	..	..1	....	....
6	4.......1	...3.....	1.	1.1	....	....
10	1.11.1.1..	...3.....	.1	4.1	....	....
13	1....112..	...4.....	..	..1	....	....
B						
14	.....1.1	...51.....	1.	...	..11	....
22	.....1.1.1	1..41.....	.2	.3.	..3.	....
23	.....1.....	...41.....	11	...	..31	....
24	.....1.1.1	...31.....	11	...	.4.	....
26	.....1.1.1	.....1..1.	12	...	.2.	....
27	.....1.1.2	.....111..	1.	...	..41	....
7	...4.1.1.1	...2.1...1	.1	..1	.1..	....
15	....	2114.1...2	11	...	..11	....
16	.....1..	...4.2...1	.1	..2	...1	....
17	.....1..	...5.2....	.1	.2.	...1	.1..
18	.....1..	...5.2....	.1	.1.	.1.1	....
19	.....1..	...2.....	.1	..1	...1	....
20	.....1.1..	1..5.2...1	.1	..1	1111	....
21	.....1..	...3.3...1	11	...	.1.1	.1..
C						
25	....	....	.1	...	1.31	....
28	.....1..	....	..	..3	.13.	..1.
29	....	....	12	1.3	..42	..1.
30	....	....	.1	112	..2.	1.1.
31	....	.....1....	11	1.1	..5.	..1.
32	....	....	..	1.1	..5.	..1.
37	....	....	..	...	..52	..1.
33	....	....	..	221	.133	.21.
38	....	.....1....	..	..1	..22	.41.
D						
34	.....1..	....	.1	...	.141	.31.
35	.....1..	....	..	...	..51	.111
36	....	....	..	...	..4.	.213
41	....	....	.1	...	.12.	..16
39	....	....	1.	...	.12.	.14.
40	....	....	.1	...	..12	.52.

blocking, albeit of rather different kinds while *Pred* is the only measure which has neither single nor two membered groups. The *SED* result is atypical in its production of single membered groups, presumably due to the effects of standardisation to zero mean and unit variance. *Syn* is the only other measure which produces single membered plot groups and the apparent blocking is clearly unrelated to any minimal variance. *COO* seems more blocked than *Can, Fcl* more than *Cal* while overall *Tinf* is possibly the most strongly blocked, minimal variance result, but the core Levenshtein measures produce reasonably strong blocking. I have not attempted a detailed interpretation of the groups formed in environmental terms, since I am not an expert on the vegetation being studied.

Turning now to the SAHN results, Fig. 1, and their comparison, what is most obvious is that *BC, FI* and *Lev* are all identical. Further the *Ccn* coefficient is very close to them. An adjacent grouping of the various information measures, *Div, Tinf, Qinf*, and Canberra metrics, *COO, Can*, shows that these coefficients all of which require very similar constraints in a Levenshtein context, are all somewhat similar and not so far from the more relaxed. Levenshtein distances of *BC* and *Ccn*. The other coefficients do not really form a coherent cluster, and I suspect they are grouped more on differences from the Levenshtein measures than on any real similarity among themselves. The inclusion of the *Eid* value result as an outlier is expected of course.

The coordinate analysis, Fig. 2, provides a similar kind of result though one might interpret the result as a horseshoe, representing a single axis running from *Eid* to *BC*. It is interesting, though that the third axis separates the two Calhoun variants, *Cal, Fcl*, and puts the *Syn* measure near the original topological one while the fuzzy version is placed near the *Chi* metric. *Eid* is as always an outlier. Whether these placements have any significance remains to be seen.

The additive clustering, Table 10, also stresses the strong relationships of the Levenshtein based measures, while the *Eid* result is not even included in a cluster! Of course with this method I could request more clusters, and presumably at some stage the eident result would enter some cluster, though the weight attached to that cluster would necessarily be small. In summary there does not seem to be a strongly weighted cluster without the *BC-FI-Lev* core, and the most heavily weighted cluster contains the Levenshtein related measures. Ccn is associated with this triple. The *Div, COO* and *Can* measures group together several times, often accompanied by *Pred. Tinf, Qinf* and *SED* also group with them most of the time. The remaining measures appear in a single cluster only, with *Chi* forming an oddity in the last cluster. We appear then to have three major centres, with a number of peripheral measures largely unrelated. Since the three centres as composed of Levenshtein related measures, presumably the outliers are either very odd Levenshtein forms, or reflect aspects of similarity not reflected by Levenshtein measures.

Table 9. Partition dissimilarities: Rajski's metrix × 1000.

	Eid	Syn	Cal	Div	Tinf	Qinf	Pred	SED	COO	Can	BC	FI	Lev	Chi	FC	CCn
Eid	*	973	960	986	982	979	986	954	988	992	978	978	978	996	958	972
Syn		*	798	768	756	710	841	808	788	825	814	814	914	895	929	782
Cal			*	774	788	772	840	730	765	783	738	738	738	947	883	723
Div				*	439	344	555	369	188	298	311	311	311	817	754	445
Tinf					*	151	616	421	396	287	533	533	533	828	791	445
Qinf						*	641	498	283	392	439	439	439	828	761	339
Pred							*	646	577	546	671	671	671	821	832	734
SED								*	485	405	497	497	497	831	773	424
COO									*	147	227	227	227	808	729	326
Can										*	343	343	343	805	757	430
BC											*	0	0	840	749	227
FI												*	0	840	749	227
Lev													*	840	749	227
Chi														*	950	854
FC															*	770
CCn																*

Table 10. Additive similarity clustering: group membership and group weights.

Weight = .308
Members: BC, FI, Lev, CCn, Div, COO, Can, Tinf, Qinf, SED,

Weight = 0.284
Members: BC, FI, Lev, CCn

Weight = 0.166
Members: BC, FI, Lev, Pred, Div, COO, Can

Weight = 0.155
Members: BC, FI, Lev, CCn, Pred, Syn, Cal, FCl, Div, COO, Can, Tinf, Qinf, SED

Weight = 0.103
Members: CCn, Pred, Div, COO, Can, Tinf, Qinf, SED, Chi

Finally the additive similarity tree, Fig. 3, again isolates a branch structure relating the Levenshtein measures. (Note that on this diagram only vertical distances are significant in estimating the dissimilarities). It does seem successful in discriminating the related partitioned information measures, and in maintaining separation between the non-Bray-Curtis coefficients. The isolation of the *Chi* result is notable, while *Ccn* is again close to the Levenshtein measures.

Discussion

From these results it would seem that the Levenshtein related measures do give the same kind of result, provided they are employing similar transformation rules and other constraints. The dual coefficient of Faith seems to give largely similar results, so that, on these data, the differences seem very small.

In contrast, the Euclidean distance and Chi-square measures, which are both very constrained and limited to substitution, differ markedly. What is perhaps surprising is that the information measures do not differ at least as much. True, the diversity measure gives the same kind of result, but none of the various information measures seems to be quite so clearly separated as might be expected, given that the rules and constraints necessary for a Levenshtein formulation of them would seem to be very close to those of the euclidean and chi-square metrics. As far as the other measures are concerned, they do seem to be capturing aspects of similarity which are not well represented in the Levenshtein-based measures commonly employed. In particular, the Predicate calculus coefficient gives a very interesting two-way table, unrelated to the patterns seen in the majority of tables. It does seem that this coefficient is recording aspects of similarity not present in the Levenshtein measures used here.

I do not, of course, claim that these exhaust all the possible measures by any means. In some cases the similarities are directly observed rather than measured, as in Lehmann (1972). What properties should be expected under these circumstances is not clear; metricity or euclidean properties do not seem likely. Other obvious examples which have been excluded from this preliminary survey are Goodall's (1964) probabilistic measure, Mountford's (1971) coefficient is based on assumptions of the form of the species area curve, Levandowsky and Winter's (1971) geodesic distance, and Wahl's (1983) or Vašiček and Jičin's (1976) shape distances. Largely these were omitted simply because programs for their calculation were not immediately available, but they will need examination in the future. Correlation coefficients seem to be Levenshtein metrics with "time-warping" permitted but certainly require further investigation.

There is, however, one other class of distances which does seem of more than passing interest and these are the Hausdorf distances. Bednarek and Ulam (1979) present a method of calculating distances in the following manner. Give an item x, define an operation f(x) which returns a set of items associated with, and including, x. Now define f(f(x)) as the set resulting if the operation is applied to all members of the set f(x), and more generally $f^n(x)$ as the operation applied n times. Now consider two items A, B and determine the value of n where $f^n(A) = f^n(B)$? This value is the, integer, Hausdorf distance between A and B. Obviously the result depends on the function used; one Levenshteinian possibility would be that f(x) returns all items one "rule-application" away from x.

I have been toying recently with the possibility of using the taxonomic hierarchy in an analogous manner, the f(x) now being a change from species to genus, or genus to family. Nakamura and Iwai (1982) propose something similar when they admit "properties of high order". If x is like y then the properties of y are regarded as properties of x, but at the second order, and so on for higher orders. The problem is one of relating, if it be possible, the Hausdorf distance to a Levenshtein distance; this would provide a very general model of similarity measures. For example, Calhoun's distance seems to have what I might call a Hausdorfian flavour. A taste of the future perhaps?

Acknowledgements. To Franta, Susan and Tomas for the non-monotone walk in the park, and to Alessandra for the swans' dialogue.

REFERENCES

ARABIE, P. and J.D. CARROLL. 1980. MAPCLUS: a mathematical programming approach to fitting the ADCLUS model. Psychometrika 45: 211-235.

AUSTIN, M.P. and L. BELBIN. 1982. A new approach to the species classification problem in floristic analysis. Aust. J. Ecol. 7: 75-89.

BARTELS, P.H., G.F. BAHR, D.W. CALHOUN and G.L. WIED. 1970. Cell recognition by neighbourhood grouping techniques in Ticas. Acta Cytol. 14: 313-324.

BEDNAREK, A.R. and S.M. ULAM. 1979. An integer valued metric for patterns. In: *Fundameantals of Computation Theory*, pps 52-57. Academie-Verlag, Berlin.

BEN-BASSAT, M. and L. ZAIDENBERG. 1984. Contextual template matching: a distance measure for patterns with hierarchically dependent features. IEEE Trans Patt. Anal. Machine Intell. PAMI-6: 201-211.

BLACKBORN, D.T. 1980. A generalized distance metric for the analysis of variable taxa. Bot. Gaz. 141: 325-335.

BOWMAN, D.M.J.S. and B.A. WILSON. 1986. Wetland vegetation pattern on the Adelaide River flood plain, Northern Territory, Australia. Proc. Roy. Soc. Qld. 97: 69-77.

BURKEA, J. and C.R. RAO. 1982. Entropy differential metric distance and divergence measures in probability spaces: a unified approach. J. Multivar. Anal. 17: 575-596.

BYKAT, A. 1979. On polygon similarity. Inform. Process. Lett. 9: 23-25.

ČULIK, K. and D. WOOD. 1982. A note on some tree similarity measures. Inform. Process. Lett. 15: 39-42.

CHEETHAM, A.H. and J.E. HAZEL. 1969. Binary (presence-absence) similarity coefficients. J. Paleont. 43: 1130-1136.

CZEKANOWSKI. J. 1909. Zur differential Diagnose der Neanderthalgruppe. Korrespbl. dt. Ges. Anthrop. 40: 44-47.

DALE, M.B. 1964. "The application of multivariate methods to heterogenous data." Ph. D. Thesis, University of Southampton.

DALE, M.B. and D.J. ANDERSON. 1972. Qualitative and quantitative information analysis. J. Ecol. 60: 639-653.

DALE, M.B. and D.J. ANDERSON. 1973. Inosculate analysis of vegetation data. Austral. J. Bot. 21: 253-276.

DALE, M.B., H.T. CLIFFORD and D.R. ROSS. 1984. Species, equivalence and morphological redescription: a Stradbroke Island vegetation study. In: R.J. Coleman, J. Covacevich and P. Davie (eds.), *Focus on Stradbroke: New Information on North Stradbroke Island and surrounding areas*, 1974-1984. Boolarong Publ., Brisbane and Stradbroke Island

Management Organization, Amity Point.

DALE, M.B. and W.T. WILLIAMS. 1978. A new method of species reduction for ecological data. Austral. J. Ecol. 3: 1-5.

DAY, W.H.E. and D.P. FAITH. 1986. A model in partial orders for comparing objects by dualistic measures. Math. Bio. Sci. 78: 179-192.

FAITH, D.P. 1983. Asymmetric binary similarity measures. Oecologia (Berlin) 57: 287-290.

FAITH, D.P. 1984. Patterns of sensitivity of association measures in numerical taxonomy. Math. Bio. Sci. 69: 199-207.

FEOLI, E. and M. LAGONEGRO. 1983. A resemblance function based on probability: applications to field and simulated data. Vegetatio 53: 3-9.

FEOLI, E., M. LAGONEGRO and L. ORLÓCI. 1984. *Information Analysis of Vegetation Data*. Dr. W. Junk, the Hague, pps. 143.

FOWLKES, E.B. and C.L. MALLOWES. 1983. A method for comparing two hierarchical clusterings. J. Amer. Statist. Assoc. 78: 553-569.

GAMBAROV, G.M., I.D. MANDEL and I.A. RYBINA. 1980. Some metrics arising in data analysis. Automat. Remote Control 41: 1717-1723.

GOODALL, D.W. 1964. A probabilistic similarity index. Nature 203: 1098.

GOWER, J.C. 1966. Some distance propreties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325-338.

GOWER, J.C. 1986. Metric and euclidean properties of dissimilarity coefficients. J. Classif. 3: 5-48.

HAJDU, L.J. 1981. Graphical comparison of resemblance coefficients in phytosociology. Vegetatio 48: 47-59.

HILL, M.O., R.G.H. BUNCE and M.W. SHAW. 1975. Indicator species analysis, a divisive polythetic method of classification and its application to a survey of native pine-woods in Scotland. J.: Ecol. 63: 597-613.

JANSON, S. and J. VEGELIUS. 1981. Measures of ecological association. Oecologia 49: 371-376.

JUHÁSZ-NAGY P. 1984. Spatial dependence in plant populations 2. A family of new models. Acta Bot, Hung. 30: 363-402.

KASHYAP, R.L. and B.J. OOMMEN. 1983a. A common basis for similarity measures involving two strings. Int. J. Comput math. 13:17-40.

KASHYAP, R.L. and B.J. OOMMEN. 1983b. Similarity measures for sets of strings. Intern. J. Comput Math. 13: 95-104.

KLOPMAN, G. and O.T. MACINA. 1985. Use of the computer automated structure evaluation program in determining quantitative structure-activity relationships with hallucinogenic phenylalkylamines. J. theor. Biol. 113: 637-648.

KORHONEN, T. 1984. "Self-Organization and Associative Memory". Springer-Verlag, Berlin. pps 125-188.

KULLBACK, S. 1959. Information Theory and Statistics. Wiley, New York.

LAMBERT, J.M. and W.T. WILLIAMS. 1962. Multivariate methods in plant ecology IV. Nodal analysis. J. Ecol. 50: 775-802.

LAMONT, B.B. and K.J. GRANT. 1979. A comparison of twenty-one measures of site dissimilarity in: L. Orlòci, C,R,Rao and W.M, Stiteler (eds.), *Multivariate Methods in Ecological Work*. pp. 101-126. International Coop. Publ. House, Fairland, Maryland.

LANCE, G.N. and W.T. WILLIAMS. 1967. Mixed data classificatory programs Agglomerative systems. Aust. Comput. J. 1: 82-85.

LANCE, G.N. and W.T. WILLIAMS. 1968. Mixed data classificatory programs II. divisive systems. Austral. Comput. J. 1: 82-85.

LE QUENSE, W.J. 1974. The uniquely derived character concept and its cladistic application. Syst. Zool. 23: 513-517.

LEHMANN, D.R. 1972. Judged similarity and brand-switching data as sililarity measures. J. Marketing Res. 9: 331-334.

LEMONE, K.A. 1982. Similarity measures between strings extended to sets of strings. IEEE Trans. Patt. Anal. Mach. Intel. PAMI - 4; 345-347.

LERMAN, I-C and P. PETER. 1985. Elaboration et logiciel d'un indice de similarité entre objets d'un type quelconque. Application au probleme de consensus en classification. IRISA, Rennes. Publ. Intern. 262. pps 72.

LEVANDOWSKY, M. 1972. An ordination of phytoplankton populations of ponds of varying salinity and temperature. Ecology 53: 398-407.

LEVANDOWSKY, M. and D. WINTER. 1971. Distance between sets. Nature 234: 34-35.

LEWIS, P.A.W., BAXENDALE, P.B. and J.L. BENNET. 1967. Statistical discrimination of the Synonymy/Antonymy relationship between words. Assoc. Comput. Mach. J. 14: 20-44.

LITTLEM, I.P. and D.R. ROSS. 1985. The Levenshtein metric, a new means for soil classification tested by data from a sand-podzol chronosequence and evaluated by discriminant analysis. Aust. J. Soil Res. 23: 115-130.

LU, S-Y. and K.S. FU. 1978. A sentence-to-sentence clustering procedure for pattern analysis. IEEE Trans. Systems, Man and Cybernetics SMC-8: 381-389.

MICALASKI, S. and R.E. STEPP. 1985. Automated construction of classifications: conceptual clustering versus numerical texonomy. IEEE Trans. Patt. Anal. Mach Intel. PAMI-5: 396-410.

MIYAMOTO, S. and K. NAKAYAMA. 1986. Similarity measures based on a fuzzy set model and application to hierarchical clustering IEE. Trans. Syst. Man Cyber. SMC-16: 479-482.

MOORE, R.K. 1979. A dynamic programming algorithm for the distance between two finite areas. IEE Trans. Patt. Anal. Machine Intell. PAMI-1: 86-88.

MOUNTFORT, M.S. 1971. A test of the difference between two clusters. In: Patil G.P.,Pielou, E.C. and W.E. Waters. *Statistical Ecology 3*. pp 237-251. Penn State Univ. Press.

NAKAMURA, K. and S. IWAI. 1982. A rapresentation of analogical inference by fuzzy sets and its application to information retrieval system. In. M.M. Grupta and E. Sanchew (eds.), Fuzzy Information and Decision Processes pp 373-386. North Holland.

ORLÓCI, L. 1978. Multivariate Analysis in Vegetation Research. Dr. W. Junk, the Hague. pps. 451.

ORLÓCI, L. 1969. Information theory models for hierarchic and non hierarchic classification. In. A.J. Cole (ed.), Numerical Taxo nomy. pp. 148-164. Academic Press. London.

RAO, C.R. 1982. Diversity and dissimilarity coefficients: unified approach. Theor. Popultn. Biol. 21: 24-43.

RAJSKI, C. 1961 Entropy and metric spaces, In: C. Cherry (ed.) Information Theory pp. 41-45. Butterworth, London.

SANKOFF, D. and J.B. KRUSKAL. 1983. *Time Warps, String Edits and Macromolecules: the Theory and Pratice of Sequence Comparison*. Addision Wesley, London.

SAMDAL, C.E.A. 1974. A comparative study of association measures. Psychometrika 39: 165-187.

SATTATH, S. and A. TVERSKY. 1977. Additive similarity trees. Psychometrika 42: 319-345

SEPOLSKY, J.J. 1974. Quantified coefficients of association and measurement of similarity. Math. Geol. 6: 135-152.

TVERSKY, A. 1977. Features of similarity. Psychol Rev. 84: 327-352.

VAN RIJSBERGEN, C.J. 1986. A non-classical logic for information retrieval. Comput. J. 29: 481-485.

VAŠIČEK, Z. and R. JIČIN. 1976. The problem of similarity of shape. Syst. Zool. 21: 91-96.

VENOT, A. LEUBRUCHEC, J.F. and J.C. ROUCAYROL. 1984. A new class of similarity measures for robust image registration. Comput. Vision, Graphics. Image Process. 28: 176-184.

VESELY, A. 1981. Logically oriented cluster analysis. Kybernetika 17: 82-92.

WAHL, F.M. 1983. A new distance mapping and its use for shape measurement of binary patterns. Comput. Vision, Graph. Image Process. 23: 218-226.

WALLBRECHER, E. 1976. Ein-Cluster-Verfahren wur richtungsstatistichen Analyse tektonischer Daten. Geol. Rdsch. 67: 840-857.

WERMAN, M., PELG, S. and A. ROSENFELD. 1985. A distance metric for multidimensional histograms. Comput. Vision, Graphics and Image Process. 32: 328-336.

WILLIAMS, W.T. 1973. Partition of information. Austral. J. Bot. 21: 277-281

WOLDS, H. 1986. Similarity indices, sample size and diversity. Oecologia 50: 296-302.