The weakening of taxonomic inferences by homological errors

By

D. M. Jackson

and

L. J. White*

Technical Report

No. 70-66

July 1970

Department of Computer Science
Cornell University
Ithaca, New York 14850

*(Department of Computer and
Information Science
The Ohio State University
Columbus, Ohio 43210)

Abstract.

In the past decade there has been a growing concern in devising classification algorithms which are applicable to large bodies of data. Such algorithms are characterized necessarily by a sacrifice of statistical sophistication for a gain in computational simplicity. Accordingly, inferences drawn from taxonomic studies in which these algorithms have been employed may be affected by accidental and poorly understood features of such algorithms. An error analytic technique is presented which reduces this possibility. It is applicable to many of the classification algorithms currently in use. The combinatorial problems encountered in the error analysis are discussed and a computationally viable method for their solution is formulated. The technique is illustrated by an experiment with a small set of data.

Contents

THE WEAKENING OF TAXONOMIC INFERENCES BY HOMOLOGICAL ERRORS

Table of Notations.

The following table gives the principal symbols used in this analysis, their definitions and the number of the section in which they are defined.

symbol	Defined in section	Definition
t_j, t_k	4	arbitrary objects described in terms of binary attributes.
M	4	The number of objects in the population.
N	4	The number of attributes available for compiling descriptions of objects.
S	4	A scrambling of the binary vector description of an object into another binary vector description by the injection of of errors in the attribute states.
S*	4	The set of all scramblings.
P(S)	4	The probability of a particular scrambling.
r	4	The probability of an error in the assignment of an attribute to an object.
f	4	The similarity between t_j, t_k before scrambling.
f_S	4	The similarity between t_j, t_k after scrambling.
$\emptyset(f, f_S)$	4	A statistical function of f and f_S.
$\emptyset*$	4	The expectation of $\emptyset(f, f_S)$.

a	5	The <u>category</u> of attributes of the ordered pair (t_j,t_k) which have state 0 in t_j and state 0 in t_k.
b	5	The <u>category</u> of attributes of the ordered pair (t_j,t_k) which have state 1 in t_j and state 0 in t_k.
c	5	The <u>category</u> of attributes of the ordered pair (t_j,t_k) which have state 0 in t_j and state 1 in t_k.
d	5	The <u>category</u> of attributes of the ordered pair (t_j,t_k) which have state 1 in t_j and state 1 in t_k.
n_a,n_b,n_c,n_d	5	The <u>cardinals</u> of the categories a,b,c,d before scrambling of (t_j,t_k) by S.
m_a,m_b,m_c,m_d	5	The <u>cardinals</u> of the categories a,b,c,d after the scrambling of (t_j,t_k) by S.
$\underline{n},\underline{m}$	5	$(n_a,n_b,n_c,n_d),(m_a,m_b,m_c,m_c)$ respectively.
$h(\underline{m},\underline{n})$	5	A <u>generating function</u> involved in an analytic expression for $\emptyset$*.
P(i,j)	5,6	The a priori <u>transition probabilities</u> among pairs of categories a,b,c,d.
s	6	The <u>probability</u> of no error in the assignment of an attribute to an object.
t	6	r/s
I	7	A square bivariate <u>table</u> of dimension three or four, having non-negative integer entries.

k	7	N - trace(I) + $I_{13}+I_{31}$ defined only for the case of I with dimension three.
$\bar{k}$	7	The upper bound of k which permits Ø* to be computed to within a prescribed error of ϵ .
$E_{ij}^{(N)}$	7	The <u>error matrix</u> for N attributes.
K_I	9	A <u>function</u> of the table entries of I.
$T(N)$	10	The <u>number</u> of tables, I, for all row and column constraints.
$T(\underline{n})$	10	The <u>number</u> of tables, I, with prescribed row sum constraints.
$T(\underline{n},k)$	10	The <u>number</u> of tables with prescribed row sum constraints and having parameter k.
ϵ	11	The <u>permissible error</u> in the computation of Ø*.
$R(N,\bar{k})$	12	<u>Proportion</u> of the number of tables with parameter $k \leq \bar{k}$ to the number of those with $k \leq 2N$.

1. Introduction.

The general intention of cluster analysis is to reduce detailed information about the individuals or a population to generalizations about certain subsets of this population. This general statement of intent offers little guide for devising methods of constructing classifications, and indeed, taxonomists have approached the problem from a number of substantially different points of view, from the essentialism of Caesalpino, for example, to the neo-nominalism of the present day pheneticists and the evolutionism of the phylogeneticists. Classification techniques and indeed the foundations of these techniques still remain controversial as witness the discussions in recent taxonomic literature (see, for example, Colless (1) and Johnson (2)). The clustering algorithms which have been devised often lead to classifications at considerable variance with one another, even when applied to the same set of data and the same population (Minkoff (3)).

To a certain extent a classification is artificial since it depends, particularly in its form, on the purposes for which it was constructed. Classifications of the same set of data, but serving different purposes, will therefore justifiably differ from each other. Thus, for example, classifications of a set of organisms according to their ecological or morphological characteristics will doubtless differ. Nevertheless, any similarity between the two classifications, were it to be found, would be invaluable as evidence in support of a hypothesis relating the suitability of certain morphological characters in the organisms of the population to particular ecological environments. However, if a classification is to represent, in a tractable form, once this has been decided upon, the affinities which exist between subsets of the objects of the population, then it might reasonably be expected that classifications devised by different methods should be in substantial agreement with one another. Stated negatively, there would be grounds for concern if the classifications had no, or at most few, points of agreement.

Some variability, clearly, derives from the measurement of affinity between pairs of objects. There is considerable flexibility in the way in which this might be done in practice, ranging from determination by an appropriate formula from the character lists on the one hand (Sokal and Sneath (4)), to a reduction of the <u>a priori</u> "proximities" to conform to a specified accuracy with a suitable metric hypothesis on the other hand (Kruskal (5)). But the variability in the results from cluster analyses using the same set of affinities has led understandably to the questioning of the reliability and objectivity of these techniques in practice, particularly when a second analysis has been performed with the intention of corroborating the findings of the first. Before accepting the results of a cluster analysis, it is therefore appropriate to determine whether or not the results satisfy a number of necessary conditions which characterize, in part, the relationship between an adequate classification and the original population.

In this respect problems in taxonomy are analogous to those of numerical analysis in which the numerical solution of a partial differential equation, for example, is acceptable with confidence only when solutions by different methods and different step lengths agree to an appropriate number of significant figures, and when tolerable error bounds on the solution have been obtained. The purpose of this paper is to examine these conditions and to describe how, in certain cases, such conditions may be applied in practice.

2. Necessary Conditions for Classification.

Throughout this paper, we shall be concerned only with qualitative attributes and in particular with binary attributes. It will therefore be assumed that quantitative attributes and multistate attributes have been reduced to this form. A number of the better known similarity functions for binary attributes is given by Sokal and Sneath (op.cit.). The functions are divided into two sets according to the contribution of the absence of an attribute from both objects whose mutual affinity is to be computed. One set of functions is invariant under reversal of the (0-1) coding scheme. The functions of the other set are not invariant under this transformation.

The following is a set of necessary conditions which are generally applied to classification procedures:

3) The classification must be well-defined. That is, application of the algorithm must supply a single result.

4) The classification must be stable. That is, the classification must not be grossly affected by small changes in the data.

2) The algorithm must be independent of the labelling of the objects. That is, the algorithm must be unaffected by a permutation of the names of the objects.

1) The algorithm must be independent of scale. That is, the algorithm must be unaffected by multiplication of the similarity matrix by a positive, non-zero constant.

These conditions are clearly not sufficient since they leave entirely unresolved the question of the adequacy of a particular configuration of classes in summarizing the detailed inter-object affinities; that is, the adequacy of the grouping as a classification (senu strictu) of the objects of the population is not disclosed. Indeed, sufficiency would imply the "optimality" of the classification, a concept eloquently attacked by Johnson (op. cit) as premature.

Independence of scale (condition 1) is a natural requirement since the scale of similarity is essentially arbitrary. However, the stronger condition of invariance under monotonic transformation of the similarity function has been suggested, and this leads to methods in which similarities between objects are largely irrelevant. The methods rely on the inequalities between the similarities of pairs of objects. Similarity functions induce a partial ordering on the set of objects to be classified, and the classification is constructed such that the partial ordering derived from it agrees as closely as possible with the partial ordering induced on the data. The measurement of agreement between partial orderings is given by Benzécri (6). The methods are accordingly insensitive to monotonic transformations of the similarity functions, with the result that a number of the similarity functions, namely those which are mutually monotonic, may be regarded as equivalent. In addition, the agreement between the partial orderings provides a quantifiable criterion for the adequacy of a classification in summarizing the inter-object affinities. A review of these methods is given by de la Véga (7).

Independence of labelling (condition 2) requires that a classification should not be dependent on the order in which the objects are treated, and is demonstrable only for a classification in which all classes which satisfy the classificatory criterion have been enumerated. It is clear that if all classes satisfying the criterion are to be located, then all subsets of the objects of the population must have been examined. Similarly, condition 3 is demonstrable only in the case in which all sets which satisfy the classificatory criterion have been extracted. Indeed, the failure of 2 and 3 are related. Relabelling enables more classes to be found, and the examination of all permutations corresponds to inspecting all subsets of the objects to determine which satisfy the classificatory criterion. Conditions 2 and 3 are susceptible to testing only when all subsets of objects of the population have been examined. For large populations, that is, populations for which examination of all subsets on a given computer configuration cannot be contemplated, these conditions cannot be applied. Accordingly, for such populations, the remaining condition, that of stability, is decisive.

3 Stability of Classifications.

For the moment, stability will be defined informally: a classification will be said to be <u>stable</u> if small changes in the data lead to small changes in the classification. This definition will be examined more critically later and the computability and quantification of such a notion will be examined.

The need for stability arises as follows. Suppose that a certain population of objects is to be classified. Each of the objects has been examined and a set of attributes appropriate to the population has been compiled. Measurements or observations are made upon the objects to record the value or state of each of the attributes. Descriptions are therefore obtained for each of the objects of the population. Errors attributable to several factors may occur during the compilation of this data (Jackson (8)), but attention will be concentrated here on those arising from the faulty determination of homological components. A classification of the material must be obtained in spite of the errors in contains. Obviously, the original material is not recoverable and the most which can be achieved is to ensure that the erroneous parts of the data do not exert a significant effect on the classification as a whole.

Fisher and Rohlf (9) have examined the effect of errors in determing the homology of characters on the results of numerical taxonomic studies. Resulting dendrograms were compared by the Sokal and Rohlf (10) cophenetic correlation measure to determine significant disagreement between the classifications. For a particular population Fisher and Rohlf have asserted that the cophenetic correlation is largely unaffected by errors in homology in a certain proportion of the attributes. The purpose of the analysis presented here is to provide a technique for ensuring this in general for a proportion of attributes which may be specified in advance.

It is inappropriate to elaborate here on the conceptual and procedural difficulties encountered in the definition and determination of homology. Many of the definitions based on phylogenetic considerations appear to be viciously circular in that homological decisions are invoked to establish recency of common ancestry. Recently, Sneath (11) has attempted to extract some conceptual common demoninators from homological problems arising in a variety of disciplines. Jardine (12) has proposed a computational procedure for determining homological components in organisms by positional similarity, and advocates a return to Owen's (13) original criterion. However, for the purposes of the present error analysis, it is clear that errors of the type described above cbtrude, the inadequacy of current definitions of homology notwithstanding.

Goodall (14) has considered the related problem of determining the theoretical distribution of a particular similarity function, namely, the Matching Coefficient, in terms of the a priori probabilities of the binary attributes under the assumption that different attributes occur in random pairs. He remarks that a knowledge of this distribution might facilitate the use of significance tests in numerical taxonomy. It is clear that dependencies between attributes will reduce the number of degrees of freedom below that for independent attributes, and that this will have an effect on the estimation of the significance of an affinity.

The need for such tests, while reluctantly admitted, has seldom been attempted.

The case for stability may be argued on purely practical grounds. Suppose that a few additional observations were made on the population of objects already classified. It would be expected that the classification constructed on the basis of the additional information would differ only marginally from the previous classification. Indeed, a large change in the classification would indicate extreme sensitivity to the additional set of observations and would suggest, in the first instance, the insufficient information had been gathered to provide adequate descriptions of the objects.

A connection may be seen between stability and continuity. Suppose that a classification is produced by a transformation of the data. Small changes in the data necessarily lead to small changes in the classification, provided that the transformation is continuous. The classification is therefore stable under these circumstances and stability can be inferred by verifying the continuity of the transformation. In general, however, classification algorithms cannot be expressed in these terms. This is particularly true of classifications constructed from categorical data where the similarity function although continuous, has a discrete domain and therefore a discrete range. It is also true of classifications in which class membership is not susceptible to quantification, for example, by a probablistic weight, but in which class membership is a two state relationship (present/absent).

Small changes in the object descriptions are measured quite straightforwardly. Suppose that two arrays are given, representing different descriptions of the same population. The arrays may be regarded as preference tables and the difference between them may be computed by Kendall's (15) Coefficient of Agreement. Changes in the classification, however, are less readily measured, particularly in the most common case in which the population is too large to permit examination of all subsets of objects; that is, in the case in which the classification is incomplete.

The problem reduces to one of comparing classifications. Sokal and Rohlf (10) have devised a method which applies to dendrogram classifications. For non-hierarchical classifications, a method of low order for comparing a classification with the original data, itself a primitive classification, has been discussed by Jackson (16).

For the case in which a classification algorithm cannot be represented by a continuous transformation of the data, stability is not an analytic property of the algorithm. Rather, it is dependent both on the algorithm <u>and</u> the population which is to be classified. Stability must therefore be demonstrated each time a classification algorithm is applied to a set of data, by investigating the action of the algorithm on the particular set of data.

Alternatively, stability must be ensured by an appropriate modification of algorithms currently in use. This latter possibility will now be examined. The analysis will be applicable both to hierarchical techniques and to clustering techniques utilizing a matrix of inter-object affinities.

4 Modification of Classification Algorithms.

Suppose that there are M objects in the population and N attributes available for providing descriptions of the objects. The attributes are binary. The i-th. row of the binary array D gives the objects j which are described in terms of the i-th. attribute. If $D_{ij} = 1$ then object j is described by attribute i, while if $D_{ij} = 0$ then object j is not described by attribute i. Errors of the type described in Section 3 have the effect of changing a zero into a one or vice versa, for a number of the D_{ij}'s. The errors are assumed to be independent and equiprobable. The effect of errors is to change the array D into an array D'. Let r be the probability of finding an error in a particular D_{ij}. Thus rMN is the expected number of errors in the array. The number of errors which will be tolerated may be decided upon in advance, and from this a value of r may be obtained. Typically r will be set to 0.05, corresponding to a 5% error rate, although this rate may be changed as appropriate. Consider two arbitrary objects t_j and t_k represented by the binary vectors A and B, both of length N, and suppose that these vectors are scrambled by errors into the vectors A' and B', the vector representations of the same objects in D'. In terms of the data array D, A and B are given by $D_{.j}$ and $D_{.k}$ respectively, where $D_{.j}$ is the j-th. column of D. Then (A,B) is transformed into (A',B') by a <u>scrambling S</u>. Let f and f_S be the values of the similarity between t_j and t_k before and after A and B have been transformed into A' and B'. For example, suppose:

$$\begin{array}{lcl} t_j = A = (1\ 0\ 1\ 1\ 0\ 1\ 0\ 1) & \xrightarrow{S} & A' = (\underline{0}\ 0\ 1\ \underline{0}\ 0\ 1\ \underline{1}\ \underline{0}) \\ t_k = B = (0\ 1\ 0\ 1\ 0\ 1\ 1\ 1) & & B' = (0\ \underline{0}\ 0\ 1\ 0\ 1\ \underline{0}\ 1) \end{array}$$

The erroneous object-attribute states have been underlined. Suppose that $\emptyset$ is a function of f and f_S, and $\emptyset^*$ is the expectation of $\emptyset$. Let S* be the set of all possible scramblings for pairs of vectors of length N. Then:

$$\emptyset^*(t_j,t_k) = \sum_{S \text{ in } S^*} P(S).\emptyset(f_S,f)$$

where S denotes a particular scrambling in S*, and P(S) is the probability of a particular scrambling. Let $\emptyset(f_S,f) = |f - f_S$ Accordingly, $\emptyset^*(t_j,t_k)$ gives the expectation of the absolute error in $f(t_j,t_k)$ on the assumption that errors occur independently and equiprobably with probability r. Let E be the matrix such that: $E_{jk} = \emptyset^*(t_j,t_k)$. E will be called the error matrix and is of dimension M by M. It will be seen in Section 6 that this matrix may be stored in a more economical form.

Classification algorithms which may be expressed as hill-climbing algorithms are now considered. These algorithms involve the iterative optimization, generally by a local strategy, of prescribed function, g say, by shifting objects from class to class to form new partitions of the population. Clearly, clustering by the methods of Rubin and Friedman (17), the methods of Parker-Rhodes and Needham (18) and Lerman's method (19), for example, may be expressed in this form. Rubin and Friedman's method involves the optimizing of "criterion functions"; Parker-Rhodes' and Needham's methods involve optimizing "cohesion functions" and Lerman's method involves optimizing "comparison criteria."

Let g_k be the value of g after the k-th step. Step k+1 is acceptable if g has become smaller between these two steps. The stability condition requires that the reduction in g at step k+1 is attributable to a redistribution of the objects of the population among the classes of the classification rather than to the effect of errors in the data array D. Let g'_k be the smallest value of g_k which can be obtained when the similarities between the relevant objects are subjected to the appropriate fluctuations $\pm E_{ij}$ and let e'_k be the decrease in g_k. Similarly, let g''_{k+1} be the largest value of g_{k+1} under the appropriate fluctuations $\pm E_{ij}$ and let e''_{k+1} be the increase in g_{k+1} . The requirement $g''_{k+1} - g'_k < 0$, yields $g_{k+1} - g_k < -(e'_k + e''_{k+1})$. This is the revised condition for minimization necessary to ensure stability of the classification with respect to errors occurring with probability r in the data array, and is stronger than the original condition $g_{k+1} - g_k < 0$.

5. Estimation of the Expectation of a Statistical Function of Similarity.

Consider two objects and suppose that their vector descriptions are A and B. Suppose that (A,B) is scrambled into (A',B') by a scrambling S. Let A_i and B_i be the i-th. components of A and B. The following combinations may occur for (A_i,B_i): (0,0),(1,0),(0,1) and (1,1). For simplicity, denote these by a, b, c, and d respectively. When A and B change into A' and B', any one of a, b, c and d may be changed into any other. a, b, c and d will be called categories and changes from one category to another will be called transitions. Let S_{ij} be the number of i's which are converted from j's. Then the following relations hold among the S_{ij} since the categories are mutually exclusive and account for all transitions:

$\sum_i S_{ij} = n_j$ where n_j is the number j's in (A,B)

$\sum_j S_{ij} = m_i$ where m_i is the number of i's in (A',B')

$\sum_j n_j = \sum_i m_i = N$ the number of attributes.

Let $P(i,j)$ be the probability of the transitions $j \longrightarrow i$.

There are n_a a's which must be distributed among the categories a, b, c, d in such a way that there are S_{aa} in a, S_{ba} in b, S_{ca} in c and S_{da} in d. Let T_a be the probability of one of these. Then:

$$T_a = \frac{n_a!}{S_{aa}!\, S_{ba}!\, S_{ca}!\, S_{da}!} \cdot (Q(a,a).Q(b,a).Q(c,a).Q(d,a))$$

$$= \frac{n_a!}{\prod_i S_{ia}!} \cdot \prod_i Q(i,a)$$

where $Q(i,j) = P^{S_{ij}}(i,j)$

Now $P(S) = \prod_i T_i$, so substituting the above expression for T_a, we have:

$$P(S) = \frac{\prod_j n_j!}{\prod_{i,j} S_{ij}!} \cdot \prod_{i,j} Q(i,j)$$

Now, from Section 4:

$$\emptyset^* = \sum_{S \text{ in } S^*} P(S).\emptyset(f_s,f)$$

$$= \sum_{\underline{m}}{}^* \sum_{S}{}^{**} P(S).\ \emptyset(f_S,f)$$

by splitting the first sum into a sum over all scramblings which reach a particular $\underline{m}$, and then a sum over all permitted values of the components of $\underline{m}$.

where $\sum\limits_{\underline{m}}^{*}$ is the restricted summation subject to

$\sum\limits_{i} m_i = N$ and $\sum\limits_{S}^{**}$ is the restricted summation subject t

$\sum\limits_{i} S_{ij} = n_j$ and $\sum\limits_{j} S_{ij} = m_i$.

We may therefore write:

$$\emptyset^* = \left(\prod_j n_j!\right) . \sum_{\underline{m}}^{*} \emptyset(f,f_S) . h(\underline{m},\underline{n})$$

where:

$$h(\underline{m},\underline{n}) = \sum_{S}^{**} \left(\frac{\prod\limits_{i,j} Q(i,j)}{\prod\limits_{i,j} S_{ij}!} \right)$$

By inspection:

$$h(\underline{m},\underline{n}) = \underline{\text{coeff of}} \prod_{i} q_i^{m_i} \ \underline{\text{in}} \ \frac{\quad}{\prod\limits_j n_j!} . \prod_j \left(\ _i \ q_i P(i,j) \right)$$

where q_i are dummy variables.

The analysis may be carried further on the assumption that f is a polynomial in $\underline{m}$. A generating function containing formal differential operators may be derived in this case. The details are given by Jackson (20).

6. Application of the Analysis to a Special Case.

The analysis is now applied to a particular class of similarity functions, of which the functions of Jaccard, Russel and Rao, Dice and Kulczynski (Sokal and Sneath (óp.cit.)) are special cases. Each of these functions involves the quantities n_a, (n_b+n_c) and n_d. Since the distinction between the categories b and c is not utilized, the set of categories a,b,c,d may be replaced by the smaller set a,x, d where $x = b \vee c$. These categories are again exclusive and complete. Table 1 gives expressions for the similarity between two objects before and after scrambling, for each of the four functions

Name of function	Value before scrambling	Value after scrambling
Jaccard:	$f = n_d/(N-n_a)$	$f_S = m_d/(N-m_a)$
Russel and Rao:	$f = n_d/N$	$f_S = m_d/N$
Dice:	$f = 2n_d/(N+n_d-n_a)$	$f_S = 2m_d/(N+m_d-m_a)$
Kulczynski:	$f = n_d/(N-n_a-n_d)$	$f_S = m_d/(N-m_a-m_d)$

Table 1

Since r is the probability of an error in the assignment of an attribute to an object, s=1-r is the probability that no error has occurred. Let P(i,j) be the *a priori* transition probability between i and j; that is, P(i,j) is the *a priori* probability that j will change into i.

Then $P(a,c)$, for example, is equal to the probability that $(0,0)$ changes into $(0,1)$, and this is equal to rs, the probability that the first component is unaffected by the scrambling and the second component is in error. Thus the complete probability matrix is given by:

$$[P(i,j)] = \begin{bmatrix} s^2 & rs & rs & r^2 \\ rs & s^2 & r^2 & rs \\ rs & r^2 & s^2 & rs \\ r^2 & rs & rs & s^2 \end{bmatrix}$$

The transition probabilities for the three category case a,x,d may be derived straightforwardly from the more general case a,b,c,d. Thus, for example,

$$\begin{aligned} P(a,x) &= P(a,b \vee c) \\ &= \tfrac{1}{2}(P(a,b) + P(a,c)) \\ &= \tfrac{1}{2}(rs + rs) = rs \end{aligned}$$

and

$$\begin{aligned} P(x,x) &= P(b \vee c, b \vee c) = \tfrac{1}{2}(P(b \vee c, b) + P(b \vee c, c)) \\ &= \tfrac{1}{2}(P(b,b) + P(c,b)) + \tfrac{1}{2}(P(b,c) + P(c,c)) \\ &= \tfrac{1}{2}(s^2 + r^2) + \tfrac{1}{2}(r^2 + s^2) = r^2 + s^2 \end{aligned}$$

For the three category case a,x,d, the transition matrix $P(i,j)$ is therefore given by:

$$[P(i,j)] = \begin{bmatrix} s^2 & rs & r^2 \\ 2rs & r^2 + s^2 & 2rs \\ r^2 & rs & s^2 \end{bmatrix}$$

Therefore substituting these values in the general expression for $h(\underline{m},\underline{n})$, it follows that:

$$h(\underline{m},\underline{n}) = s^{2N} \cdot \sum_{I}^{**} C_I \, (1 + t^2)^{I(x,x)} \cdot t^k$$

and where $C_I = \dfrac{2^{m_x - I(x,x)}}{\prod_{i,j} I(i,j)!}$; $k=N-\Sigma I(i,i) + I(a,d) + I(d,a)$; and $t=r/s$

$\emptyset^*$ is therefore given by:

$$\emptyset^* = \left(\prod_j n_j!\right) \cdot \sum_{\underline{m}}^{*} \emptyset(f_S,f).h(\underline{m},\underline{n})$$

The same polynomial $h(\underline{m},\underline{n})$ in t is applicable in the case of each of the four similarity functions given above. Some attention will therefore be given in the next section to the problems involved in computing $h(\underline{m},\underline{n})$.

Clearly, $\emptyset^*$ is a function of the initial n_a, n_x, n_d which are determined from the attribute lists for a pair of objects t_j,t_k say. Accordingly, the values of $\emptyset^*$ for all pairs of objects may be tabulated either in terms of (j,k) or in terms of (n_a,n_d) or indeed, any pair of n_a, n_x, n_d. Tabulation of $\emptyset^*$ in terms of (j,k) yields a square matrix of dimension M while tabulation in terms of (n_a,n_d) yields a matrix of dimension N. The choice between the two tabulations may be resolved by selecting the scheme which gives the smaller dimensional matrix. Further reduction may be obtained for (n_a,n_d) tabulation by determining the values of $\emptyset^*$ for only those (n_a,n_d) which actually occur in a given population.

It has already been remarked that the same polynomial $h(\underline{m},\underline{n})$ is applicable to each of the four well known similarity functions of Jaccard, Russell and Rao, Dice and Kulczynski. Providing a technique can be found for computing $h(\underline{m},\underline{n})$ rapidly, considerable generality has been achieved. Since taxonomic studies frequently involve using a number of such functions exploratively, this generality is valuable in practice and will be retained throughout the subsequent analysis. The computation of $h(\underline{n},\underline{n})$, however, is not straightforward, for it involves summation over all 3 by 3 integer non-negative tables, I, with row and column sum constraints. In this connection, O'Neil (21) has determined table frequencies for given row and column sum constraints. However, the tables considered by him are of larger dimension and the table entries themselves are constrained to be zero or one. Once all feasible 3 by 3 tables have been determined, the value of $h(\underline{m},\underline{n})$ may be computed by substituting directly a suitable numerical value of t. If the number of feasible tables is large, not all of the tables will result in substantial contributions to $\emptyset^*$. Provided the polynomial expansion of $h(\underline{m},\underline{n})$ in terms of t may be truncated after a specified number of terms with tolerable error, and provided those tables which contribute to the retained terms may be isolated, then some economy has been achieved. A number of results will now be given which are useful for this purpose.

7. Specification of the Computational Problem.

The computational problem is to construct the $M \times M$ error matrix E, where $E_{jk} = \emptyset^*(t_j, t_k)$ and t_j, t_k ranger over all M objects of the population. There are N binary attributes associated with each object. Consider the two binary N-dimensional attribute vectors corresponding to the two objects, and suppose that in n_1 components both vectors have zeros in common, n_2 components of the two vectors disagree, and in n_3 components both vectors have ones in common. It is clear that $E_{jk} = \emptyset^*(\underline{n})$ where $\underline{n}$ specify the parameters for t_j and t_k. To find all elements of E, $\emptyset^*(\underline{n})$ must be evaluated for all possible non-negative integers n_1, n_2, n_3 which sum to N. The expectation $\emptyset^*$ of an arbitrary function $\emptyset(f, f_S)$ is given by

$$\emptyset^*(\underline{n}) = \sum_{\underline{m}} \emptyset(f, f_S) . h(\underline{m}, \underline{n})$$

where the summation is taken over all feasible vectors $\underline{m}$. The probability weight $h(\underline{m}, \underline{n})$ is evaluated by considering 3 by 3 tables I_{ij} of non-negative entries with row sums n_1, n_2, n_3 and solumn sums m_1, m_2, m_3. Define $k = N - \text{trace}\ (I) + I_{13} + I_{31}$, an off-diagonal sum. Now:

$$h(\underline{m}, \underline{n}) = s^{2N} \cdot \sum_{I} K_I . t^k \text{ , } (1 + t^2)^{I_{22}}$$ where $s = 1-r$, $t = r/s$. The summation is over all tables I_{ij} with row and solumn sums $\underline{n}$ and $\underline{m}$ respectively and:

$$K_I = \frac{2^{(m_2 - I_{22})} \cdot \prod_i (n_i)!}{\prod_{i,j} (I_{ij})!}$$

Minor nctational changes have been made to simplify the following analysis. $n_a, n_x, n_d, m_a, m_x, m_d$ have been replaced by

$n_1, n_2, n_3, m_1, m_2, m_3$ respectively to facilitate the manipulation of the table I. This notation will be maintained throughout the remaining sections.

Suppose that N, r and the functions f and $\emptyset$ are given. With s and t determined, $\emptyset^*(\underline{n})$ must be computed for all $\underline{n}$. In subsequent sections this problem is investigated.

Two sets of constraints are described, each of which completely characterize valid tables, and these constraints are used to reduce the time required to compute $\emptyset^*$. It is proposed that the amount of computation can be substantially reduced if only the few most important tables are considered. It will be shown that these tables may be characterized by $k \leq \bar{k}$ for some $\bar{k}$. The method for determining $\bar{k}$ requires the following quantities: the range of K_I and the conditions under which the extrema are attained, and the number of tables with specified parameter k. In order to describe the method explicitly, f is chosen to be the well-known Jaccard function and $\emptyset$ is chosen to be the modulus (ie. absolute value) of the difference between f and f_s. Further discussion and a complete documentation of the computer program for computing error matrices are given by Mathis _et al_.(22)

8. The Construction of Feasible Tables.

Given non-negative integers $n_1, n_2, n_3, \bar{k}$, and N such that $n_1 + n_2 + n_3 = N$ the problem is to find feasible tables I_{ij}, $i,j = 1,2,3$ such that:

$$\sum_j I_{ij} = n_i, \quad i = 1,2,3$$

and $k \leqslant \bar{k}$, where k is defined in section 7.

Feasible tables may be found by considering the following constraints. $\lfloor q \rfloor$ denotes the integer part of q.

$$n_2 \leqslant N$$

$$n_1 \leqslant N - n_2$$

$$0 \leqslant I_{13} \leqslant \min(n_1, \left\lfloor \frac{\bar{k}}{2} \right\rfloor)$$

$$0 \leqslant I_{12} \leqslant \min(n_1 - I_{13}, \bar{k} - 2I_{13})$$

These determine $I_{11} = n_1 - I_{12} - I_{13}$

$$0 \leqslant I_{31} \leqslant \min(n_3, \left\lfloor \frac{\bar{k} - 2I_{13} - I_{12}}{2} \right\rfloor)$$

$$0 \leqslant I_{32} \leqslant \min(n_3 - I_{31}, \bar{k} - 2I_{13} - 2I_{31} - I_{12})$$

These determine $I_{33} = n_3 - I_{31} - I_{32}$

$$0 \leqslant I_{21} \leqslant \min(n_3, \bar{k} - 2I_{13} - 2I_{31} - I_{12} - I_{32})$$

$$0 \leqslant I_{23} \leqslant \min(n_2 - I_{21}, \bar{k} - 2I_{13} - 2I_{31} - I_{12} - I_{32} - I_{21})$$

These determine $I_{22} = n_2 - I_{21} - I_{23}$

The conditions are necessary and sufficient and are generated by considering $\bar{k}$ in the form:

$$\bar{k} = I_{12} + I_{21} + I_{23} + I_{32} + 2I_{13} + 2I_{31}$$

Thus, for example, it follows that $I_{13} \leqslant \left\lfloor \frac{\bar{k}}{2} \right\rfloor$, which forms part of the third inequality.

These constraints are particularly simple since the lower bounds are zero. Once the table entries have been determined the parameters m_1, m_2, m_3 are specified. The latter are necessary for the calculation of $\emptyset(f, f_S)$. The constraint on $\bar{k}$ is

$$1 \leq \bar{k} \leq 2N$$

The upper bound corresponds to an absence of constraint on k. In practice, however, $\bar{k}$ is usually relatively small, as shown in Section 11.

If $n_1, n_2, n_3, m_1, m_2, m_3$ and N are given such that $n_1 + n_2 + n_3 = N$ and $m_1 + m_2 + m_3 = N$ then the following constraints may be used to construct feasible tables:

$$\mathrm{Max}(0, n_1-m_2-m_3, m_1-n_2-n_3) \leq I_{11} \leq \mathrm{Min}(n_1, m_1)$$

$$\mathrm{Max}(0, m_2-n_2-n_3, n_1-I_{11}-m_3) \leq I_{12} \leq \mathrm{Min}(m_2, n_1-I_{11})$$

$$\mathrm{Max}(0, m_1-n_3-I_{11}, n_1+n_2-m_2-m_3-I_{11}) \leq I_{21} \leq \mathrm{Min}(n_2, m_1-I_{11})$$

$$\mathrm{Max}(0, n_1+n_2-m_3-I_{11}-I_{12}-I_{21}, m_1+m_2-n_3-I_{11}-I_{12}-I_{21}) \leq I_{22} \leq \mathrm{Min}(n_2-I_{21}, m_2-I_{12})$$

The other table entries may be expressed in terms of these:

$$I_{13} = n_1 - I_{11} - I_{12}$$

$$I_{31} = m_1 - I_{11} - I_{21}$$

$$I_{23} = n_2 - I_{21} - I_{22}$$

$$I_{32} = m_2 - I_{12} - I_{22}$$

$$I_{33} = m_3 - n_1 - n_2 + I_{11} + I_{12} + I_{21} + I_{22}$$

These conditions are again necessary and sufficient.

9. Optimization of the Table Coefficient.

An analysis of the coefficient

$$K_I = \left(\prod_{i=i}^{3} n_i!\right) C_I = \frac{2^{(m_2 - I_{22})} \cdot \prod_{i=i}^{3} n_i!}{\prod_{i,j} I_{ij}!}$$

is required to describe a method for calculating the power $\bar{k}$ of t at which the polynomial $h(\underline{m},\underline{n})$ may be curtailed with specified truncation error. This method is discussed in Section 11 and results in substantially fewer tables to evaluate and gives a more efficient technique for computing $\emptyset^*(\underline{n})$. Three quantities are required:

(i) Given $\underline{n} = (n_1, n_2, n_3)$, find the table I^* which results in a maximum coefficient K_I^* for this vector.

(ii) A vector $\underline{n}$ and a table resulting in the maximum coefficient $(K_I)_{max}$.

(iii) A vector $\underline{n}$ and table which correspond to the minimum coefficient $(K_I)_{min}$.

(iii) is straightforward. It is clear that when $I_{11} = n_1$, $I_{22} = n_2$, $I_{33} = n_3$ and all other table entries zero, that $K_I = 1$ and this is valid independent of the vector $\underline{n}$. When any other assignment of table elements is made, however, the denominator decreases, and thus K_I must increase.

(ii) is given by the solution of:

$$\max_{n_1,n_2,n_3} \left\{ \frac{2^{(I_{12})} \cdot (n_1)!}{\prod_j (I_{1j})!} \cdot \frac{(n_2)!}{\prod_j (I_{2j})!} \cdot \frac{2^{(I_{32})} \cdot (n_3)!}{\prod_j (I_{3j})!} \right\}$$

subject to $n_1 + n_2 + n_3 = N$ and $\sum_j I_{ij} = n_i$.

Each of these three factors may be optimized separately, regardless of the choice of $\underline{n}$, so that the solution of (i) is embedded in (ii). It may be shown that if $\underline{n} = (n_1, n_2, n_3)$ is given, the optimum value K_I^* of K_I is determined by the table I^*, where:

$$I_{ij}^* = n_i/(3 - (-1)^j |i - 2|) ,$$

in which the entries must be integers as close to these values as possible. This is the solution to (i).

For (ii), K_I^* must be optimised over all $\underline{n}$. It may be shown that this maximum, $(K_I^*)_{max}$, is attained when either $n_1 = N$ or $n_3 = N$. In either case, $(K_I^*)_{max}$ is given by the tables:

$$I_{ij}^{(1)} = N\delta_{i1}/ (2 + 2 |i - 2|)$$

$$\text{or} \quad I_{ij}^{(2)} = N\delta_{i3}/ (2 + 2 |i - 2|)$$

where δ_{ij} is the Kronecker delta. Again, the entries of these tables must be integers as close as possible to the indicated values.

10. Table Frequencies.

Since the computation time required to construct the error matrix is proportional to the number of valid tables which are to be considered (see Section 12), the number of tables T under various conditions will be determined. For details of these derivations, see Mathis et al. (op. cit.)

For any given N, let T(N) be the number of tables which must be evaluated to calculate $\emptyset*$ for all $\underline{n}$. There are $\binom{N+2}{2}$ such vectors and $T(N) = \binom{N+8}{8}$, the latter being the number of ways of distributing N indistinguishable objects among nine cells. T(N) is of the order of N^8, which anticipates the time consuming nature of this enumeration for computational purposes.

Given $k = N - \text{trace}(I) + I_{13} + I_{31}$, let $T(\underline{n},k)$ be the number of tables with this parameter. To determine $T(\underline{n},k)$, the dependence of k on $\underline{n}$ must be taken into account, and this leads to a large number of cases to be considered. Some simplification may be obtained, however, for the special case in which:

i) $k \leqslant \min(n_1,n_2,n_3)$

Define $j = I_{13} + I_{31}$

whence $k - 2j = I_{12} + I_{21} + I_{32} + I_{23}$

There are j + 1 non-negative integer solutions to the first equation and $\binom{k-2j+3}{3}$ non-negative integer solutions to the second, the latter estimate being the number of ways of distributing $(k-2j)$ indistinguishable objects among four cells. Then:

$$T(\underline{n},k) = \sum_{j=0}^{\lfloor k/2 \rfloor} (j+1) \binom{k-2j+3}{3}$$

$$= \begin{cases} (k+2)(k+4)(k+6)(k^2+8k+10)/(2.4.6.10) & \text{for even } k \\ (k+1)(k+3)(k+4)(k+5)(k+7)/(2.4.6.10) & \text{for odd } k \end{cases}$$

There remain a further four cases to consider. These are:

ii) $n_1 < \bar{k}$ (and $n_2, n_3 \geqslant \bar{k}$)

iii) $n_2 < \bar{k}$ (and $n_1, n_3 \geqslant \bar{k}$)

iv) $n_1, n_3 < \bar{k}$ (and $n_2 \geqslant \bar{k}$)

v) $n_1, n_2 < \bar{k}$ (and $n_3 \geqslant \bar{k}$)

It is required that $\bar{k} \leqslant N$, for $\bar{k}$ to be an effective parameter for computational purposes. The five cases are mutually exclusive and include all possible cases, n_1 and n_3 having been treated symmetrically. Referring to the first set of necessary and sufficient conditions for feasible tables given in Section 8, case (i) has the effect of making the constraints on n_1, n_2, n_3 larger than those on $\bar{k}$, since the latter is so small. The row constraints are therefore made redundant. For the remaining cases this does not happen and both the row constraints and the k-constraint are applied. Since these conditions are more binding on the construction of tables, the table counts $T(\underline{n},k)$ in each of the cases (ii), (iii), (iv) and (v) will be smaller than the count for case (i). The latter may therefore be used as an upper bound for the counts for cases (ii),(iii),(iv) and (v).

11. Approximation of the Polynomial by Table Truncation.

The results concerning coefficient K_I from Section 9 and $T(\underline{n},k)$ from Section 10 are now applied in order to develop a method for specifying the tables which determine $\emptyset^*(n)$ to within an assigned error. It is now necessary to select a specific similarity function $f(\underline{n})$ and statistical function $\emptyset(f,f_S)$. The well-known Jaccard function $f(\underline{n}) = n_3/(n_2 + n_3)$ will be used. Let $\emptyset(f,f_S)$ be the modulus of the difference between $f(\underline{n})$ and $f_S(\underline{m})$. The motivation for this choice of $\emptyset$ is given in Section 4.

Clearly, the smallest non-zero value of $\emptyset$ is $1/N$ and occurs when $n_2 + n_3 = m_2 + m_3 = N$, $m_3 \geq n_3 - 1$. The largest value of $\emptyset$ is unity which occurs when $n_3 = m_2 = 0$ or when $m_3 = n_2 = 0$. The $(1 + t^2)^{I_{22}}$ factors may be ignored since $t \ll 1$, and although I_{22} may be as large as N, even in this case $(1 + t^2)^N \cong 1 + Nt^2$ and will not be an important factor provided that $N < 1/t^2$. Moreover, few tables will possess an entry I_{22} of this magnitude.

Suppose now that an error $\epsilon > 0$ due to table truncation is given where ϵ is typically in the range $0.01 \leq \epsilon \leq 0.1$. The truncation point is defined as the minimum $\bar{k}$ such that:

$$\sum_{\substack{I \\ k \leq \bar{k}}} K_I \emptyset(\underline{n},\underline{m})\ t^k\ (1+t^2)^{I_{22}} \geq \frac{1}{\epsilon} \sum_{\substack{I \\ k > \bar{k}}} K_I \emptyset(\underline{n},\underline{m})\ t^k\ (1+t^2)^{I_{22}}$$

where $k = N - \text{trace}\ (I) + I_{13} + I_{31}$ and the summation is over those tables with parameter k less than or equal to $\bar{k}$. For a conservative estimate of $\bar{k}$ assume that minimum values of K_I and $\emptyset \neq 0$ occur on the left hand side of the inequality, and that maximum values of these two factors occur on the right hand side. Now $\emptyset_{min} = 1/N$, $\emptyset_{max} = 1$, $(K_I)_{min} = 1$ and the maximum value K_I^* occurs, as shown in Section 9, for the table

$$I^*_{ij} = n_i/(3 - (-1)^j\ |i - 2|\ )$$

This corresponds to $k = n_1 + \frac{2}{3}n_2 + n_3$, which is nearly equal to N. Applying these approximation, we obtain:

$$\emptyset_{min}\ (K_I)_{min} \sum_{\substack{I \\ k \leq \bar{k}}} t^k \geq \frac{1}{\epsilon} \cdot \emptyset_{max}\ (K_I)_{max} \sum_{\substack{I \\ k > \bar{k}}} t^k$$

whence, on substitution:

$$\sum_{\substack{I \\ k \leq \bar{k}}} t^k \geq \frac{N}{\epsilon} \cdot (K_I)_{max} \sum_{\substack{I \\ k > \bar{k}}} t^k$$

The summations are governed by the number of tables with parameter k and $\underline{n}$ fixed, and $T(\underline{n},k)$, discussed in Section 10. For the limits of k, k = 0 may be ignored since $\emptyset = 0$. The range of k is therefore $1 \leq k \leq 2N$, the upper bound being attained when $k = 2(I_{13} + I_{31})$, which occurs for tables in which $I_{13} + I_{31} = N$. The inequality reduces to:

find the minimum $\bar{k}$ such that:

$$\sum_{k=1}^{\bar{k}} T(\underline{n},k)\ t^k \geq \frac{N}{\epsilon} \cdot (K_I)_{max} \cdot \sum_{k=\bar{k}+1}^{2N} T(\underline{n},k)\ t^k$$

This, together with the results of Section 10, provides an algorithm for computing an estimate, albeit conservative, for the value of $\bar{k}$. The method was used in the case $N = 10$ and $\epsilon = 10^{-3}$. The estimate arrived at for $\bar{k}$ was $\bar{k} = 6$. This is seen to be conservative, as anticipated, from the values given in Table 2.

The approximation to the polynomial $\emptyset^*$ by table truncation is not equivalent to truncation of the polynomial at a specific power of t, although the latter is, of course, a valid approach. The development of an analytic form of a suitable truncation of $\emptyset^*$ as a polynomial in t, depends on a specific choice of the similarity function f. The development of polynomials of more generality would entail individual treatment of $h(\underline{n},\underline{m})$ for dominant $\underline{m}$. For large N, the number of polynomials would proliferate to unmanageable proportions.

However, the technique which has been presented retains considerable generality. Since most of the similarity functions on two-by-two contingency tables are normalized between zero and unity, the values of $\emptyset_{max}$ and $\emptyset_{min}$ are valid, given $\emptyset$, for all of these similarity functions. Consequently in the analysis given, $\bar{k}$ is unaffected by a specific choice of similarity function. Moreover, a specific choice of similarity function does not affect which tables are to be constructed in the determination of $\emptyset^*$, nor their dominance. The procedure which has been developed utilizes f only in the evaluation of the weights $\emptyset(f,f_s)$ of $h(\underline{n},\underline{m})$, and f may be substituted for another quite simply. Indeed, the error matrices for different choices of f may be computed simultaneously with little additional cost since the bulk of the computation is absorbed by the isolation and construction of dominant tables. Since the technique drastically reduces the number of tables entailed in the computation of $\emptyset^*$, is efficient and retains sufficient generality to apply to all similarity functions of the specified type, the necessity for approximating individual $h(\underline{n},\underline{m})$'s is avoided.

12. Computational Characteristics of the Method.

A program implementing the technique for generating error matrices $E_{ij}^{(N)}$ of dimension N for different values of $\bar{k}$ has been written in Fortran IV. An IBM 360/75 was used for all the computations. Details of the computation of the three error matrices $E^{(10)}$, $E^{(20)}$, and $E^{(30)}$ will be given to demonstrate the economies which may be achieved by a judicious choice of $\bar{k}$.

The matrices are generated for the Jaccard function with an error probability of 0.05.

Table 2 shows the cells of the error matrix $E^{(10)}$ generated for several values of $\bar{k}$. Only the even numbered rows and columns are given for brevity. Comparing the matrices for $\bar{k}=5$ and $\bar{k}=6$, it is seen that sufficient accuracy has been obtained with $\bar{k}=5$.

The reduction in the number of tables needed to be generated for a given value of $k \leq \bar{k}$ is estimated by the following expression:

$$R(N,\bar{k}) = \frac{\sum_{k=1}^{\bar{k}} \sum_{\underline{n}} T(\underline{n},k)}{\sum_{\underline{n}} T(\underline{n})}$$

This expression gives the proportion of the number of tables employed in the computation with $k \leq \bar{k}$ to the number of tables with k unrestricted. The above expression may be reduced to:

$$\text{Estimated } R(N,\bar{k}) = \frac{\binom{N+2}{2}}{\binom{N+8}{8}} \cdot \sum_{k=1}^{\bar{k}} T^*(\underline{n},k)$$

since, from Section 10, $T^*(\underline{n},k)$ is the upper bound for $T(\underline{n},k)$ and is given as a polynomial in k, $\binom{N+2}{2}$ is the number of $\underline{n}$'s, and the denominator is the expression for $T(N)$. This expression for $R(N,\bar{k})$ is asymptotic in N for $\bar{k}$ small compared with N.

n_1	n_3	$\bar{k}=1$	$\bar{k}=3$	$\bar{k}=5$	$\bar{k}=6$	$\bar{k}=20$ (exact)
0	10	0.038	0.090	0.095	0.095	0.095
0	8	0.038	0.073	0.076	0.076	0.076
0	6	0.036	0.062	0.064	0.064	0.064
0	4	0.032	0.056	0.058	0.058	0.058
0	2	0.027	0.052	0.054	0.054	0.054
0	0	0.019	0.047	0.050	0.050	0.050
2	8	0.046	0.108	0.114	0.114	0.114
2	6	0.044	0.083	0.087	0.087	0.087
2	4	0.038	0.067	0.069	0.069	0.069
2	2	0.030	0.056	0.058	0.058	0.058
2	0	0.019	0.047	0.049	0.049	0.049
4	6	0.059	0.135	0.143	0.143	0.143
4	4	0.051	0.095	0.098	0.098	0.098
4	2	0.038	0.067	0.069	0.069	0.069
4	0	0.019	0.046	0.048	0.048	0.048
6	4	0.083	0.181	0.190	0.190	0.190
6	2	0.058	0.102	0.105	0.105	0.105
6	0	0.019	0.045	0.047	0.047	0.047
8	2	0.138	0.278	0.289	0.289	0.289
8	0	0.019	0.042	0.044	0.044	0.044
10	0	0.000	0.015	0.016	0.016	0.016

Table 2

Table 3 gives the predicted reduction tables generated, together with actual running times for the error matrix $E^{(10)}$ with several values of $\bar{k}$. The running time for k unrestricted was 10.6 secs. The accuracy obtained with a parameter value of $\bar{k}$ is specified by the largest error in the third significant figure which occurs in the error matrix.

1 N	2 $\bar{k}$	3 Estimated $R(N,\bar{k})$	4 Actual no of tables	5 Actual no. of tables/total no. of tables	6 Actual time saving ratio.	7 Actual running time (secs).	8 Accuracy
10	2	0.02	846	0.02	0.04	0.43	57
	3	0.07	1,926	0.04	0.06	0.65	12
	4	0.15	3,761	0.09	0.12	1.25	2
	5	0.32	6,489	0.15	0.18	1.95	Converged
	6	0.60	10,203	0.23	0.28	2.91	Converged
	20	—	43,758	1.00	1.00	10.57	$\emptyset^*=0.397$

Table 3

It is clear that no element of the matrix for $\bar{k}=6$ is larger than the same element for $\bar{k}=5$ by more than 10^{-3}. Furthermore, comparison with the last column of Table 2, showing the matrix computed without constraint on k, indicates that sufficient accuracy has been attained at $\bar{k}=5$.

Greater reductions of time were obtained in the case of the matrix $E^{(20)}$. The complete generation of the matrix with k unrestricted took 732 secs. Predicted reductions and actual running times are given in Table 4. Comparison of columns 3 and 5 of Table 4 shows that the predicted reduction of tables agrees with the actual reduction. In addition, comparison of columns 5 and 6 shows that the fraction of tables utilized for $k = \bar{k}$ is a good indicator of the time reduction factor. It is therefore evident, at least for $\bar{k}$ small compared with N, that $R(N,\bar{k})$ is a reliable indicator of the time reduction factor.

1 N	2 $\bar{k}$	3 Estimated $R(N,\bar{k})$	4 Actual no. of tables.	5 Actual no. of tables/ total no. of tables	6 Actual time Saving ratio.	7 Actual running time. (secs)	8 Accuracy
20	2	0.002	3,391	0.002	0.002	1.81	250
	3	0.003	8,331	0.003	0.004	2.96	110
	4	0.006	17,676	0.006	0.006	5.34	39
	5	0.011	33,464	0.011	0.013	9.25	11
	6	0.019	58,338	0.019	0.021	15.04	2
	7	0.031	94,986	0.031	0.032	23.38	Converged
	40	___	3,108,105	1.00	1.00	732 (estimated)	$\emptyset^* = 0.582$

Table 4

Running times for generating $E^{(30)}$ for several values of $\bar{k}$

are given in Table 5. The empirical fact, verified in the case of $E^{(20)}$, that $R(N,\bar{k})$ is a good estimate for the time reduction factor is confirmed by the running times for generating $E^{(30)}$ with several values of E. This fact may be utilized in estimating the total time required to compute $E^{(30)}$ with k unrestricted. The estimated time is approximately 1.14×10^4 secs. or 3.18 hrs.

1 N	2 $\bar{k}$	3 Estimated $R(N,\bar{k})$	4 Actual no. of tables	5 Actual no. of tables/ Total no. of tables.	6 Predicted time saving ratio.	7 Actual running time, (secs)	8 Accuracy
30	2	0.0002	7,636	0.0002	0.0003	3.93	460
	3	0.0004	19,236	0.0004	0.0006	6.64	286
	4	0.0010	41,891	0.0009	0.0011	12.68	149
	5	0.0021	81,539	0.0017	0.0019	22.02	66
	6	0.0040	146,373	0.0030	0.0033	37.32	25
	7	0.0072	245,881	0.0050	0.0053	60.10	8
	8	0.0121	391,876	0.0080	0.0084	96.11	2
	9	0.0195	597,196	0.0122	0.0123	140.24	Converged
	60	___	48,903,492	1.00	1.00	3.18hrs (estimated)	$\emptyset^* = 0.688$

Table 5

For N=100, a single element of the matrix has been computed. This element is given by $n_1 = 1$, $n_2 = 0$, $n_3 = 99$. For $\bar{k} = 21$ and 22, the estimated values of $E^{(100)}$ (1,99) were 0.883 and 0.884 respectively. Sufficient accuracy is therefore obtained with $\bar{k} = 21$. The time taken to compute this element was approximately 0.15 secs, so that the time taken to compute the complete error matrix $E^{(100)}$ is approximately 15 mins. Without table truncation the time is estimated as $\left(\binom{108}{8} \Big/ \binom{38}{8}\right)$ .3 hrs, or $2 \cdot 4 \times 10^4$ hrs (approx 3 years). The drastic reduction obtained with the aid of table truncation demonstrates the value of the method.

13. Taxonomic Results.

An experiment has been carried out with the Single Linkage Method (Sneath (23)), although one of the related methods of Complete Linkage (Sørensen (24)) or Average Linkage (Sokal and Michener (25)), might equally well have been chosen. Although work on hill-climbing methods is in progress, results for the Single Linkage method are presented because this method is one of the more widely used in the initial stages of taxonomic investigations. The techniques developed in this paper and the effect exposed by their application in a simple case will be of particular interest to practising taxonomists.

As usually stated, none of the above three methods involve hill-climbing, nor is a well defined hierarchy induced on the population by minimization of a prescribed function. Before the technique described in Section 4 could be applied to these methods, substantial restatement of them would be necessary.

To avoid this, a simple alternative to the approach adopted in Section 4 for hill-climbing algorithms is employed as follows. Since there is a probability of error in the attribute states of any of the objects, the similarity $f(t_j,t_k)$ between objects t_j and t_k is not reliable. In the notation of the previous sections, the lowest estimate for it is:

$$f(t_j,t_k) - E^{(N)}(n_1,n_3)$$

It is expected that the use of the modified matrix of similarities will result in a general weakening or removal of structural details and an overall, but not uniform, lowering of the levels at which fusions take place. In addition, the range over which fusion levels occur will be contracted, corresponding to a diminution in the discrimination between objects.

The dendrograms given in Figures 1 and 2 show these effects quite markedly. These dendrograms are classifications of a set of twently objects described in terms of twenty attributes. The data is given in Table 6. Between-object similarity was computed by the Jaccard function. The dendrogram of Figure 1 was derived by Single Linkage without application of the error matrix, while the dendrogram of Figure 2 was produced by the same classification algorithm applied to the similarity matrix modified by the error estimates. The error estimates were contained in the matrix $E^{(20)}$, the error matrix for twenty attributes with an error rate of the standard 5%.

In Figure 2, the comparative closeness of the three fusion lines at A and the two fusion lines at B clearly suggest the conflation of each set into a single line. This correction corresponds in Figure 1 to the elimination of an apparently significant distinction between objects 8, 15 and objects 3, 5. Similarly, the apparently significant distinction between objects 14, 19 and 12 vanishes. It is therefore seen that if classification is carried out with this modified matrix of similarities, the taxonomist will be in less danger of soliciting unwarrented inferences from the structural relationships between classes, given the unreliability of the data. This restraint is particularly valuable when taxonomic procedures are used exploratively to detect the existence of hitherto unknown groups in established genera, and when the traditional safeguards of accumulated experience with the population is absent.

The error matrix $E^{(20)}$ was computed with $\bar{k} = 6$ and with k unrestricted. No difference was seen between corresponding elements in the first three places of decimals. The generation of $E^{(20)}$ with $\bar{k} = 6$ consumed a little less than half 1/30th of the time required for computation with k unrestricted. Further economy may be obtained by computing only those elements of the error matrix for which (n_1,n_3) pairs arise in this specific set of data (see Section 6). There are only 49 such pairs in this case. The expected total reduction over the complete generation of $E^{(20)}$ with k unrestricted obtained by this means, using the estimated value of R(20, 6) in Table 4, is therefore:

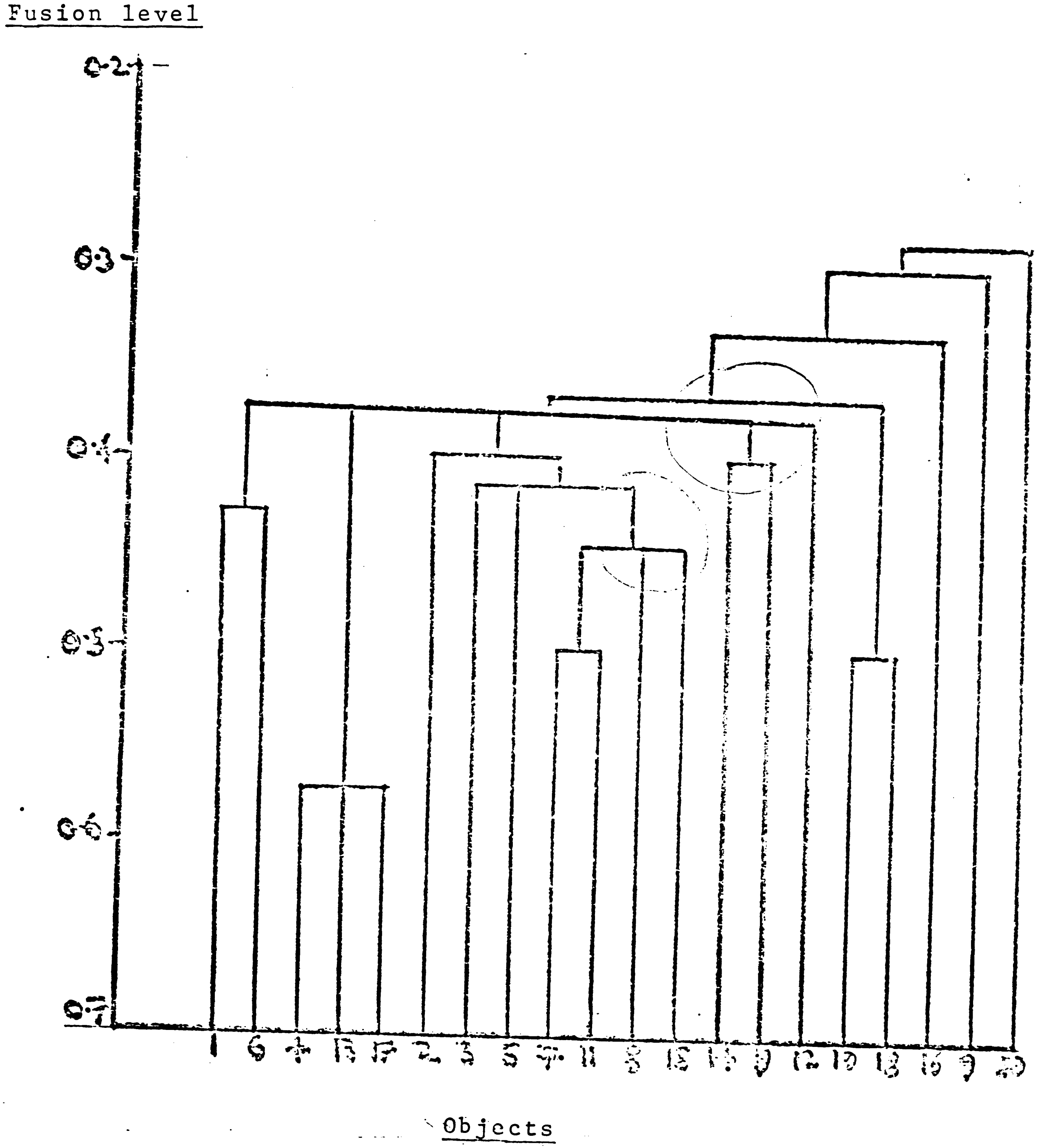
$$\frac{(0.032) \times 49}{T(N)} = (0.032) \times \frac{49}{231} \doteq 0.007$$

14. Conclusions.

A computer program has been developed which will serve as an instrument for constructing error matrices for a large number of similarity functions. If the similarity functions to be used in a taxonomic study are decided upon in advance, error matrices corresponding to each of these may be computed simultaneously at a cost comparable with that of a single error matrix. The program is efficient enough to be applicable to realistic classification problems.

Experiments with methods which do not involve hill-climbing have shown that precarious regions of the dendrograms produced are isolated and are structurally reduced. Provided the relevant error matrix is employed, taxonomic inferences made from the hierarchies are less susceptible to idiosyncracies both of the classification algorithms and other data.

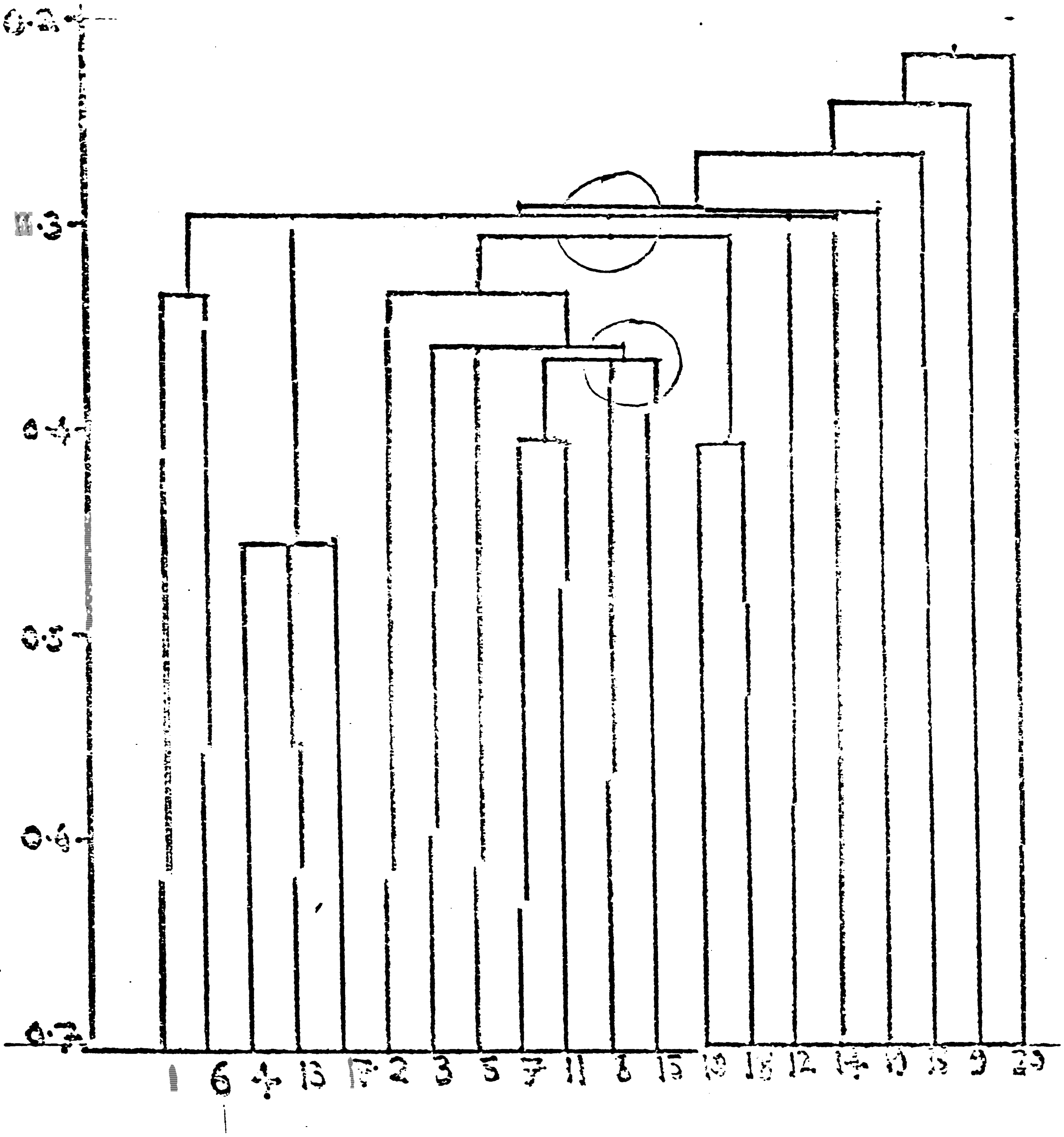
A more quantitative study of the stability of hill-climbing methods may now be undertaken. This will involve an investigation of the assignment rules in general for these algorithms along the lines proposed in Section 4.



(Analysis shows that fusion lines through each of the encircled regions should be conflated.)

Figure 1

Fusion level



Objects

(Anaylysis shows that fusion lines through each of the encircled regions should be conflated.)

Figure 2

OBJECT NO.	ATTRIBUTES															
1	2	4	6	8	10											
2	1	3	5	7	9	17										
3	1	2	5	6	8	9	10	11	12	13	14	15	16	17	19	20
4	5	6	7	8	9	10										
5	1	2	7	9	11	13	15	17								
6	2	3	4	5	6											
7	12	14	16	18	20											
8	1	15	18	19	20											
9	2	3	11	13	16	18										
10	3	9	11	12	15	16	19									
11	3	12	13	14	18	19	20									
12	17	18	19													
13	1	5	6	9	10											
14	11	13	17													
15	11	14	15	16	17	18	19	20								
16	10	11	12	13	14											
17	7	8	9	10	11											
18	3	8	12	15	19											
19	1	8	13	17												
20	4	9	11	18												

Table 6

Acknowledgements.

We should like to thank Dr. R. M. Needham (Cambridge University Mathematical Laboratory, Cambridge, England) for the original suggestion of this problem, and Miss B. A. Mathis (Department of Computer and Information Science, The Ohio State University) for her valuable contributions to the solution of a number of theoretical problems emerging from her program for generating the error matrices. We are also grateful to Mr. R. Dattola (Department of Computer Science, Cornell University) for his helpful suggestions and programming assistance. The classification programs were adapted from the CLUSTAN System provided by Mr. D. Wishart (Mathematical Institute, University of St. Andrews, Scotland).

The work was supported, at separate times, by grants from the Office of Scientific and Technical Information (United Kingdom) and by a National Science Foundation Grant (GN-534.1) to the Computer and Information Science Research Center, The Ohio State University.

References.

1. Colless, D. H. The phylogenetic fallacy. Syst. Zool., 16, 1967, pp. 289-295.

2. Johnson, L. A. S. Presidential Address. Proc. Linn. Soc. of New South Wales, 93(1), 1968, pp. 8-45.

3. Minkoff, E. C. The effects on classification of slight changes in numerical technique. Syst. Zool. 14, 1965, pp. 196-213.

4. Sokal, R. R. and Sneath, P. H. A. Principles of Numerical Taxonomy. Freeman, 1963.

5. Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a non-metric hypotheses. Psychometrika, 29, 1964, pp. 1-27

6. Benzécri, J. P. Sur les algorithmes de classification 1965, Texte multigraphié. pp. 8

7. de la Vega, W. F. Techniques de classification automatique utilisant un indice de ressemblance. Revue française de al Sociologie, 8, 1967, pp. 506-520.

8. Jackson, D. M. The stability of classifications of binary attribute data. Bulletin of the Classification Society, 1970, (in press).

9. Fisher, D. R. and Rohlf, F. J. Robustness of numerical taxonomic methods and errors in homology. Syst. Zool., 18, 1969, pp. 33-36.

10. Sokal, R. R. and Rohlf, F. J. The comparison of dendrograms by objective methods. Taxon, 11, 1962, pp.33-4

11. Sneath, P. H. A. Problems of homology in Geology and related fields. Bulletin of the Classification Society, 2(1), 1969, pp. 5-11.

12. Jardine, N. The concept of homology in Biology. Brit. J. Phil. Sci., 18, 1967, pp. 125-139.

13. Owen, R. Rep. Brit. Assoc. Adv. Sci., 16, 1847, pp. 169.

14. Goodall, D. W. The distribution of the Matching Coefficient. Biometrics, 22, 1966, pp. 882-907.

15. Kendall, M. G. The Advanced Theory of Statistics. Vol. I. C. Griffin and Co. Ltd., London, 1943.

16. Jackson, D. M. Comparison of classifications, In: Numerical Taxonomy. (Ed. Cole, A. J.), Academic Press, 1969, pp. 91-111.

17. Friedman, H. P. and Rubin, J. On some invariant criteria for grouping data. Journal of the American Statistical Assoc., 62, 1967, pp. 1159-1178.

18. Parker-Rhodes, A. F. and Needham, R. M. The Theory of Clumps. Report No. 126, Cambridge Language Research Unit, Cambridge, England, 1960, 31pp.

19. Lerman, I. C. On two criteria of classification. In: Numerical Taxonomy, (Ed. Cole, A. J.) Academic Press, 1969, pp. 114-127.

20. Jackson, D. M. An error analysis for functions of qualitative attributes with application to information retrieval. Proc. Third Conference on Computer and Information Science, 3, 1969, Academic Press (in press).

21. O'Neil, P. E. Asymptotic and random matrices with row-sum and column-sum restrictions. Bulletin of the American Mathematical Society, 75, (6), 1969, pp.1276-1282.

22. Mathis, B. A., White, L. J. and Jackson, D. M. Stability analysis of term similarities for Information Classification Theory. Technical Report 70-4, Computer and Information Science Research Center, The Ohio State University, May 1970.

23. Sneath, P. H. A. The application of computers to taxonomy. J. Gen. Microbiol., 17, 1957, pp. 201-226.

24. Sørensen, T. A method of establishing groups of amplitude in plant sociology based on similarity of species content and it application to analyses of the vegetation of Danish Commons. Biol. Skr., 5, 1948, pp. 1-34.

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