Hunt Institute for Botanical Documentation
5th Floor, Hunt Library
Carnegie Mellon University
4909 Frew Street
Pittsburgh, PA 15213-3890
Telephone: 412-268-2434
Email: huntinst@andrew.cmu.edu
Web site: www.huntbotanical.org

The Hunt Institute is committed to making its collections accessible for research. We are pleased to offer this digitized item.

Usage guidelines
We have provided this low-resolution, digitized version for research purposes. To inquire about publishing any images from this item, please contact the Institute.

Statement on harmful and offensive content
The Hunt Institute Archives contains hundreds of thousands of pages of historical content, writing and images, created by thousands of individuals connected to the botanical sciences. Due to the wide range of time and social context in which these materials were created, some of the collections contain material that reflect outdated, biased, offensive and possibly violent views, opinions and actions. The Hunt Institute for Botanical Documentation does not endorse the views expressed in these materials, which are inconsistent with our dedication to creating an inclusive, accessible and anti-discriminatory research environment. Archival records are historical documents, and the Hunt Institute keeps such records unaltered to maintain their integrity and to foster accountability for the actions and views of the collections' creators.

Many of the historical collections in the Hunt Institute Archives contain personal correspondence, notes, recollections and opinions, which may contain language, ideas or stereotypes that are offensive or harmful to others. These collections are maintained as records of the individuals involved and do not reflect the views or values of the Hunt Institute for Botanical Documentation or those of Carnegie Mellon University.

About the Institute
The Hunt Institute for Botanical Documentation, a research division of Carnegie Mellon University, specializes in the history of botany and all aspects of plant science and serves the international scientific community through research and documentation. To this end, the Institute acquires and maintains authoritative collections of books, plant images, manuscripts, portraits and data files, and provides publications and other modes of information service. The Institute meets the reference needs of botanists, biologists, historians, conservationists, librarians, bibliographers and the public at large, especially those concerned with any aspect of the North American flora.

Hunt Institute was dedicated in 1961 as the Rachel McMasters Miller Hunt Botanical Library, an international center for bibliographical research and service in the interests of botany and horticulture, as well as a center for the study of all aspects of the history of the plant sciences. By 1971 the Library's activities had so diversified that the name was changed to Hunt Institute for Botanical Documentation. Growth in collections and research projects led to the establishment of four programmatic departments: Archives, Art, Bibliography and the Library.

An application of electronic computation to studies of variation in Manihot esculenta.

David J. Rogers and Taffee T. Tanimoto

One of the greatest difficulties for the plant taxonomist in the course of his endeavors is the correlation of his data. Because of the almost insurmountable problem of correlation of as many factors as even the most superficial practitioners would like, there have been very few monographs which have been really thorough. Almost all monographers feel this problem, no matter how clear-cut their particular species may be. All have felt the doubts which assail us when we realize that we do not know how several independent characters are correlated or how, if these correlations could be achieved, these factors would very likely influence decisions as to our division into species, genera, and families. Problems of correlation are perhaps most difficult when dealing with the rather tenuous differences in sub-specific categories, and the problem is magnified where man has been interested in the plants for his own use.

In the cultivated species Manihot esculenta, which is a crop of fundamental importance to millions of tropical people, many varieties have been recorded. These are largely "artificial" varieties in the sense that they are maintained almost exclusively by vegetative means and probably would have little stability if reproduced from seed. If one wishes to make a classification of these varieties, one must largely depend upon characters which ordinarily seem very tenuous and unstable.

One way of testing stability of any one character over a number of generations is by growth in different environmental conditions for a number of years. This is obviously impractical as very few

stations or organizations would be willing to support such a project for many different crops.

Furthermore, it is a laborious task to make correlations of characters by hand techniques, because we can at best test six to eight characters simultaneously (perhaps a few more by Andersonian techniques). If judging relationship by ordinary methods, we cannot correlate (at least, I cannot) more than three items at a time. The alternative to this, as I see it, is to follow the procedures which Taffee Tanimoto and I have been trying with IBM.

The obvious advantage of the electronic computer is its capacity to handle large quantities of data, which in turn allows the analysis of the stability of and correlation among many characters simultaneously.

It is important to know, of course, that an electronic computer is no more useful than the program which is prepared for it. Some of the ground work for a program for taxonomy was laid by P. H. A. Sneath*, working with strains of the bacterial genus Chromobacterium, but his program did not have sufficient flexibility to allow for prediction, nor did it give any clue as to what should be done in case of failure of the selected characters to differentiate. We hope that some of these shortcomings have been overcome in our program.

Population samples of Manihot esculenta collected over a period of four years have given us specimens collected in warm, dry regions, in cool dry areas, and in several intermediate zones, in soils of volcanic origin, in alluvial soils, and in marine clays. With these population samples as a basis, we have attempted to

* Jour. Gen. Microbiol. 17: 184-226. 1957.

evaluate the various characters singly and together.

The actual preparation of material for use with the machine is similar to the normal collection of data which one would use in preparation of a species or varietal description. It is obviously necessary to carry such preparati n a step farther to allow for translation of the data to the binary system which the machine can use.

Preparation of the data in this manner is merely a coding device familiar to most of us and in frequent use in such works as those of Anderson on introgressive hybrids. Any character such as pigmentation, for example, may be recorded as a number, and each occurrence of this particular pigmentation recorded as that number. This may be readily converted to binary language. If one wishes to examine the shape of leaves as a differential character, it is found that the process is easily accomplished within a group of the size of a species, wherein the number of different leaf shapes will readily fall into a small number of categories.

In this particular case, some 50 characters were chosen, although the machine could handle many more than this if needed. The practicality of the classification scheme was considered here, and it was felt that not more than 20 or 25 characters should be used if the keys and definitions of the variants were to be useful to other workers. We are in a much better position to select the most significant characters after the machine manipulation of the 50 characteristics with which we started.

As pointed out, the program assists in determination of the value of certain characteristics both as "key" characters and as indications of relationship among the variants. For example, the examination of M. esculenta plants growing in museum plots in

Jamaica and Costa Rica seemed to demonstrate that the color of the stem and the color of the root were correlated factors of value both as to key characters and as indicators of relationship. Using these, it was a simple matter to divide the samples into two approximately equal stacks of plants. However, we are still left with a tremendous amount of heterogeneity in the two stacks. Inasmuch as other characters which seemed useful for classification really gave little information on how further to divide the variants, it was necessary to make some sort of analysis which would provide satisfactory divisions. How to do this again turned toward what seemed the only sensible solution--electronic computation.

Another problem arises in the process of giving weight to characters. Weighting becomes one of the major problems for the taxonomist. Which character or set of characters is most significant for classification? Which combination of characteristics will provide us with as natural a combination as is possible? The problem, I think, is controlled in setting up a program for the computers in that: (1) the taxonomist himself selects many characteristics which his experience tells him may have value, and (2) the program for the machine correlates all of those selected, simultaneously assessing their value for the purposes required of the selected characters. In assessment of weight by techniques which are most commonly used by classifiers today, one or two characters at a time are evaluated by the slow and tedious process of inspection. With the machine, we feed all characters to the computer without weight, and by the correlative abilities inherent in the program we can evaluate the weight of characters, not one at a time, but many simultaneously. After correlations are made by the machine we can see that the program has given us insights which would have

been much longer in coming to conscious levels by the usual taxonomic routine.

Even before evaluation of the weight which a character should have, we tested the program by ordinary calculating machine--a process which took us some 40 man hours of work--to determine the value of our techniques. It was heartening to note that we achieved an orderly arrangement, in most aspects similar to that which I had worked out rather arduously over the past few years. It is even more heartening to note that, when the program is finally prepared for the machine, the same operation would require about one minute!

Now, if I can just get the machine to write the keys and descriptions for me--! From other studies it looks as though we may eventually be able to get Latin diagnoses prepared for us!

The characters employed were all of a morphological nature: pigmentation, and branching pattern (frequency of branching); leaf lobe-number, characteristic lobe shape (three separate shapes), pigmentation of the petioles and young foliage; root surface (the epidermal layer either roughened or smooth) pigmentation of the cortical zone; etc.

I hope eventually to include biochemical data to assist in evaluation of the cultivars (varieties) for agricultural purposes. We already have accumulated data for nearly 100 cultivars on HCN content of the root, starch and sugar concentrations, plus crude protein concentration of the younger foliage.

In the actual classification of the cultivars, the rank of the taxa and considered evolutionary pattern are beyond the machine's level of discernment. These items are strictly dependent upon the individual taxonomist's judgment. The machine data assist in discerning break-off points between units, but the machine cannot make

decisions as to the values to be assigned any one group, nor how they are to be ordered, unless we can give the appropriate directions in the program. We found, for example, that one of our samples had more characteristics in common with all other samples than any other specimen had in common with all others. This one sample, then, is a sort of central point of the variation to be found within the total sample. Unless we, as taxonomists, can decide what to do with this sample, there is obviously little value to the whole program. Actually, we will not say that the center sample of this study has any significance until all samples of all our population samples have been correlated. When this is done, we will have some basis for a taxonomic decision, but still, the decision rests upon judgment. This judgment can only be accomplished by a person of intimate acquaintance with a particular group, and the knowledge for the judgment comes as we all know through thorough field and herbarium study.

Again, it must be said that if we have this judgment, we can ask appropriate questions of the computer and get assistance in our decisions.

In recapitulation, we find that the knowledge and skills of the taxonomist are still the most significant thing. We do find, however, that our skills and knowledge are brought to a much higher level, that we cut through tremendous drudgery, and we gain insights at a speed never before realizable through this work with the computer.

I want to take this opportunity to thank Taffee Tanimoto for the fact that this work ever saw the light of day, and IBM, who gave Dr. Tanimoto the opportunity to work on this program and provided the necessary machine time to carry it out.

AN ELEMENTARY MATHEMATICAL THEORY OF CLASSIFICATION AND PREDICTION*

T. T. Tanimoto

International Business Machines Corporation
New York, New York

Introduction.--The analysis of qualitative data is one of the areas which, to a great extent, has defied mathematical treatment. With the use of the modern large-scale electronic computers in mind, we shall propose a simple procedure which, when used as a tool, will assist us in solving problems in many fields where most of the important data are qualitative; e.g., taxonomy, organization theory, medical and psychiatric diagnosis, etc. The method does not preclude any quantitative data, since significant specified quantitative intervals can be considered as attributes. A simple enumeration process may be called a system of classification and may be very useful at times, but it certainly is not a scientific system of classification. It is generally felt that a good scientific classification system should be based on over-all similarity of attributes of objects which are felt to be pertinent to the particular purpose of the classification.[1] Although scientific classification is a system by which we can compress much information about individual objects or ideas into smaller systems, we certainly would not attempt to classify all the objects in the universe. We must localize our classification to some field or subfields of endeavor. The more local a scientific classification is, the more specific is the information yielded by the classification about the individuals. The decision as to how local or how

chosen a classification system should be is determined by the degree of specificity of the information desired in the classification. Obviously objects or ideas may be classified in many different ways depending upon what we wish to accomplish and this lead us to consider different sets of attributes; e.g., airplanes may be classified as flying objects whose class would include such things as birds, flying saucers, bats, etc.; or airplanes may be classified as means of transportation, which would take in automobiles, ships, etc. In short, the attributes of the objects or ideas which we wish to classify must be chosen in light of what we wish to accomplish by the classification. One of the most useful results of a good scientific classification system, besides that of gaining new over-all information, is its value in the prediction of the existence or non-existence of an object or an attribute. Our method will be such that the prediction will be of a probabilistic nature.

It is clear that any procedure or theory based on attributes can not be completely objective in so far as an expert in a particular field must make the decisions as to

(1) which objects are to be considered

(2) what attributes are pertinent

(3) whether a particular object does or does not possess a specific attribute of the set of pertinent attributes.

This degree of subjectivity is not only necessary, but is probably a good thing, in that an expert's personal experience and insight are incorporated into the procedure. We shall propose as our fundamental assumptions:

(i) All the objects with which we are concerned must be, to the best of our knowledge, be distinct kinds of objects.

(ii) All the attributes considered must be distinct. This does not preclude any attributes which we, from experiences, feel are comprehended by other considered attributes. If a comprehension exists, our procedure will bear this out.

We shall develop our theory upon the subjective aspects (1), (2) and (3) and the assumptions (i) and (ii).

Suppose that B is a finite set of n objects, and let a be a particular attribute possessed by some elements of B; then the classical definition of the probability p that an element of B chosen at random, assuming equal likeliness, will possess the attribute a, is given by

$$p = \frac{N(a,B)}{N(B)}$$

where N(a,B) is the number of elements of B which possess the attribute a, and N(B) is the number of elements in B. We now extend this definition to include the concept of similarity of a pair of attributes.

Let $A = \{a_i\}$, $i = 1, 2, \ldots, m$ be a finite set of m attributes (of which some may be the absence of particular attributes) associated with the finite set $B = \{b_j\}$, $j = 1, 2, \ldots, n$ of n objects. Define the $m \times n$ matrix $R = (r_{ij})$ so that $r_{ij} = 1$ if b_j possesses the attribute a_i and $r_{ij} = 0$ if b_j does not possess the attribute a_i. (Note that not considering any particular attribute in the system is completely distinct from considering its absence in the system.) Let us denote by A_i the i^{th} row

vector of B (and B_j the j^{th} column vector of B) and define $A_i \cup A_k$ as the n-dimensional row vector consisting of ones and zeros in such a way that if α_i^u and α_k^u are the u^{th} component of A_i and A_k respectively, the u^{th} component of $A_i \cup A_k$ is given by $\alpha_i^u + \alpha_k^u + \alpha_i^u \cdot \alpha_k^u \pmod 2$; i.e., if the u^{th} component of A_i or A_k (or both) is one, then the u^{th} component of $A_i \cup A_k$ is one, otherwise zero. Also define $A_i \cap A_k$ as an n-dimensional row vector so that its u^{th} component is given by $\alpha_i^u \cdot \alpha_k^u$ if α_i^u and α_k^u are the u^{th} component of A_i and A_k respectively; i.e., the u^{th} component of $A_i \cap A_k$ is one if and only if the u^{th} components of A_i and A_k are both one, otherwise zero. We now define the similarity coefficient σ_{ik} of a pair of attributes a_i and a_k with respect to the given set B of objects by

$$\sigma_{ik} = \frac{N(A_i \cap A_k)}{N(A_i \cup A_k)},$$

where the numerator and the denominator of σ_{ik} are the number of ones in the vectors $A_i \cap A_k$ and $A_i \cup A_k$ respectively. Note that $N(A_i \cup A_k) \neq 0$, since a_i and a_k are considered to be attributes which are pertinent to B and hence must be possessed by at least one element of B. If $B^* \subseteq B$ is the subset consisting of elements possessing either the attribute a_i or a_k, and since $N(A_i \cap A_k) \leq N(A_i \cup A_k)$ so that $0 \leq \sigma_{ij} \leq 1$, σ_{ij} is exactly the probability of choosing at random, assuming equal likeliness, an element of the set B^* which has both the attributes a_i and a_k simultaneously. In a similar way we define the dual similarity coefficient s_{jh} of a pair of objects b_j and b_h with respect to the set of attributes A by

$$s_{jh} = \frac{N(B_j \cap B_h)}{N(B_j \cup B_h)}$$

Note that $s_{jj} = 1$, i.e., any object is perfectly similar to itself. The $n \times n$ matrix $S = (s_{jh})$ will be called the matrix of the similarity coefficients of the objects B with respect to the set of attributes A and dually $\Sigma = (\sigma_{ik})$ the $m \times m$ matrix of the similarity coefficients of the attributes A with respect to the set of objects B. Both S and Σ are symmetric matrices with ones along the principal diagonal. Note that mathematically, the problem of classifying objects with respect to attributes is exactly the same as the problem of classifying the attributes with respect to the set of objects. One problem will be called the dual of the other.

A simple geometrical interpretation of the significance of the matrix S is easily established if we consider the objects b_j, $j = 1, 2, \ldots, n$ as points in a semi-metric space R with the distance $d_{ij} \geq 0$ between the points b_i and b_j defined by

$$d_{ij} = -\log_2 s_{ij}$$

Thus if two objects b_i and b_j are very similar, i.e., s_{ij} is nearly one, then b_i and b_j, considered as points in R are very close to each other in the sense that the distance between them is small so that our usual notion of closeness of two objects, attribute-wise, is carried over in a geometrical sense in R. If $d_{ij} < \infty$, we shall say that the point b_i is connected to the point b_j and if $d_{ij} = \infty$, then the point b_i is not connected to the point b_j. Thus the matrix (d_{ij}) defines a graph G in R. If (g_{ij}) is the point-to-point incidence matrix determined by G, i.e., $g_{ij} = 1$ if $s_{ij} \neq 0$ and $g_{ij} = 0$ if $s_{ij} = 0$, then the

*The problem was originally proposed to the author by Dr. David J. Rogers of the New York Botanical Garden.

[1] H. A. Sneath, J. Gen. Microbiology, 17, 201, 1957.

[2] D. Konig, Theorie der Endlichen und Unendlichen Graphen, Chelsea, New York, 1950.

[3] C. E. Shannon, A Mathematical Theory of Communication, Bell System Tech. Journ., 27, 379, 623, 1948.

Let μ_i be the number of zeros in the i^{th} row of the matrix S then the ramification order of the point b_i $O(b_i)$ (i.e. the number of rays emanating from the ~~point~~ $\{b_i\}$ in the graph G) is exactly $n-\mu_i$ and we write $O(b_i) = n-\mu_i$. If $O(b_i) = n$, we say that G is saturated and that ~~this~~ b_i's are completely pairwise ~~interdependent~~. If $\max_i O(b_i) = k < n$ we say that G_n is of rank k and if i_0 is not unique $\max(i_{0_j}, j_0) = i_0$.

Quasi Unimodular Matrices.

Since all the elements s_{ij} are such that $0 \leq s_{ij} \leq 1$ ~~$0 \leq |\det S| \leq 1$~~ $|\det S| \leq 1$ We have $|\det S| \leq 1$ by Hadamard's theorem on the bound of a matrix. ~~and~~ we define $\Delta = 1 - |\det S|$ as the degree of redundancy in of the b_i's with respect to the set $\{a_i\}$.

Theorem A quasi unimodular matrix is singular