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.

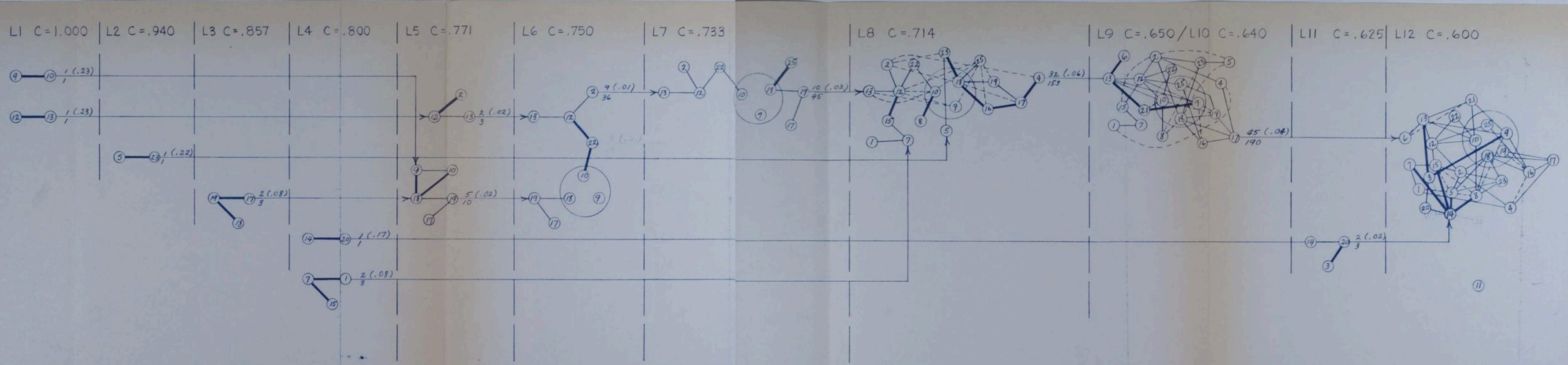
Usage guidelines
We have provided this low-resolution, digitized version for research purposes. To inquire about publishing any images from this item, please contact the Institute.

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The Hunt Institute Archives contains hundreds of thousands of pages of historical content, writing and images, created by thousands of individuals connected to the botanical sciences.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of time and social context in which these materials were created, some of the collections contain material that reflect outdated, biased, offensive and possibly violent views, opinions and actions.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 Archival records are historical documents, and the Hunt Institute keeps such records unaltered to maintain their integrity and to foster accountability for the actions and views of the collections'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.

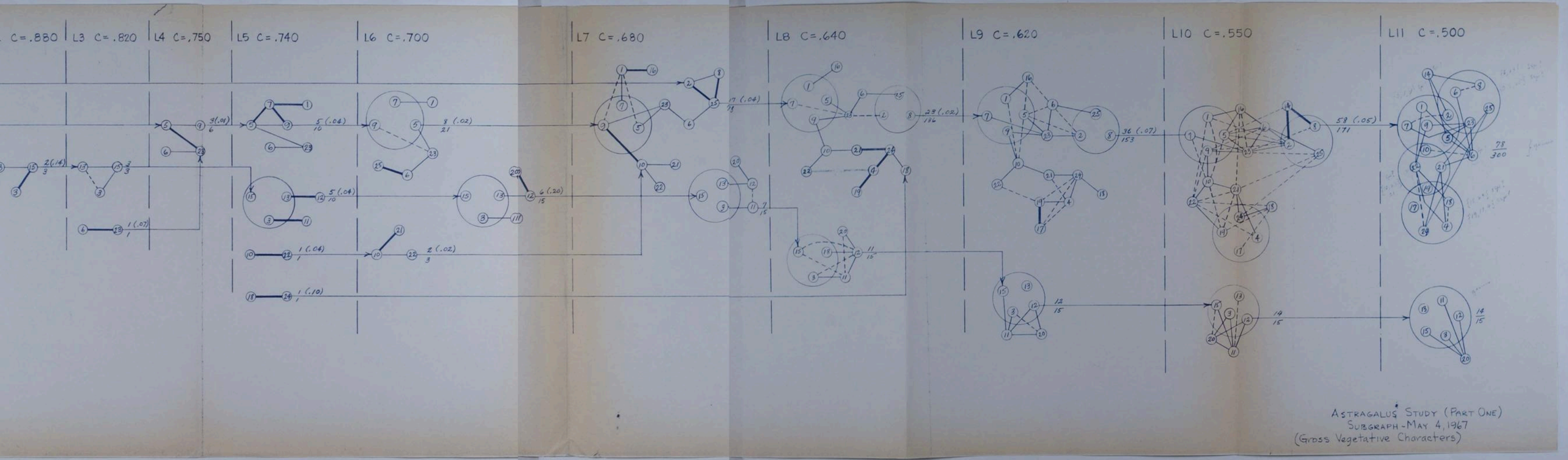
About the Institute
The Hunt Institute for Botanical Documentation, a research division of Carnegie Mellon University, specializes in the history of botany and all aspects of plant science and serves the international scientific community through research and documentation. To this end, the Institute 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.



Copies made daily Ozalid Process
Boulder Reproductions
1225 Spruce
10F 3-8

ASTRAGALUS STUDY (PART TWO)
SUBGRAPH - MAY 4, 1967
(Floral Vegetative Characters)



A TAXIMETRICAL APPROACH
TO THE
GENUS ASTRAGALUS

by

Tom Whitfield

B 170 Taximetrics

Colorado State University

June, 1967

METHOD

In its broadest sense, taximetrics is a method, but more specifically, a tool to assist the scientist, be he biologist, geologist or chemist, in the reduction and analysis of large quantities of data. In name, taximetrics is related to "taxonomy", the science of classification and naming of organisms. In methodology, it is not only related to taxonomy, but to the entire spectrum of scientific endeavor in the "synthesis" of information about the biological and physical (i.e., living and non-living) world. The underlying principles of taximetrics are those of "classification", thus the obvious relationship to taxonomy. However, classification is essentially only a process of dividing and subdividing a group of objects according to observed similarities and differences, and in taxonomy placing the subgroups in some hierarchical order. This process, though not always recognized or called classification, is basically the same method used by scientists in many fields to analyze (i.e., synthesize) data on many different forms of matter, thus the broader relationship to science.

Taximetrics then, though a new word to most, is simply the embodiment of a few tried and true methods of classification, heretofore undefined. Taximetrics subsumes these in its methodology and adds two powerful tools to reinforce them--mathematics and the computer. The computer was chosen because of its vast computational and data handling capability, and mathematics provides the means to express numerically (i.e., quantify) intuitively known, learned and verified methods (e.g., classification) in the form of a mathematical model. Construction of the model required a careful and rigorous definition of the actual methods of traditional taxonomy, and further a reduction of those necessary and sufficient ones common to all classification. The methods and principles thus derived may then be expressed in mathematical

symbols so that the various operations may be performed numerically. These expressions are then transformed into an algorithm (i.e., a series of instructions to a computer) which when fed into a computer along with the "data" collected on a set of objects, makes it possible for the computer to perform all the operations required to effect a "classification" of the objects, in a minute fraction of the time it would take the scientist to perform the same required operations by hand. Not only is the computer incomparably faster, but it is completely objective and consistent in its performance.

At this point it is necessary to "classify" classifications somewhat, according to purpose. It may be desirable to classify objects (e.g., organisms) for taxonomic purposes, or simply to discover relationships between objects or variable properties of a group of objects, among others. In either case the basic methods of classification are the same, but in the former many other considerations must be made (e.g., the definition of "taxa", and the selection of "appropriate" data) while in the latter the main concern may only be the most accurate collection of specific data.

Since there is such a diversity of disciplines and purposes for classification, taximetrics takes on a unique position and significance for science. It becomes a most flexible tool because it provides an "objective" means to peform the hard core of classificatory work without placing undue restriction on the "subjective" considerations that must be made and that vary widely from discipline to discipline, and purpose to purpose. These subjective considerations are easily imposed in the "professional judgment" required in selecting data and in interpreting the results of the computer output.

Now you may rightfully ask: just what are these "principles of classification", and what is this "data" you refer to? The basic principles of classification and thus of taximetrics are (1) a classification is a hierarchical

(i.e., smaller classes being wholly contained in larger ones) sequence of "partitions" of a given collection of objects, each partition dividing the collection into "classes" or groups; (2) within any given partition of a classification, two "similar" objects should not be placed into different classes, and (3) two "different" objects should not be placed in the same class. This is to say that the classes in any partition of the collection should be "exclusive" of one another (or distinct) and also be "exhaustive", (i.e., exhaust the collection).

So, you say fine, well and good--but how do you determine what objects are "similar" and which are different? When anyone looks at two objects to determine if they are similar or different and to what degree, he doesn't just stare at their entireties in an unfocused way and hope the answer will just come to him. Aware of it or not, he compares the two to determine what "properties" they have in common (such as shape, hardness, hairyness, appendages and etc.). These properties may be "intrinsic" (as above) having to do with the object itself or "extrinsic", having to do with the surroundings (or environment) of the object. Whichever, he compares the objects for each property they have in common to determine to what degree each possesses that property and some measure of the difference between the two degrees (e.g., if they possess the same degree of the property they are identical for that property). Then somewhere in the dark reaches of the gray matter all these degrees of all the observed properties are summed up and the conclusion is reached that these two objects are similar or they are different or--well, they aren't similar and they aren't different! When these compaisons are made for every possible pair of objects in a collection, then those pairs that seemed most different are kept apart, with the rest falling somewhere in between. The classification is then complete. If the man happens to be doing a taxonomic

classification, then he consults his authorities, uses his best professionally trained intuition, and deems certain groups of objects "taxa" and assigns to them a name based on their position in the hierarchy.

Taximetrics replaces all the foregoing process of comparison and summation and grouping and division, with a mathematical model, which when given the "data" in the proper format, quantifies the process and turns it over to a computer (as an algorithm) to operate upon the data and output the desired classification. The "data" referred to is simply the same "properties" selected before, now called "characters", and their respective "degrees of possession" now referred to as "states" of those characters, for all objects in the collection (or study). These characters must be selected for the purpose of the classification and according to your best professional judgment as to what constitutes a good character for the study. States for each character must likewise be chosen and number no less than 2 and preferably no more than 6 or 8. As such, each character is a partition or classification of the study for it should divide the entire study placing each object into one and only one of its states.

Once the data is collected on a study, it is then punched on IBM cards in a language the computer can interpret. This information is then fed into the computer along with the program (or algorithm) containing the mathematical model, or the quantified methodology of classification. The computer takes over from there, computes a "similarity" (defined in the model) value for each possible, unique pair of objects in the study. This value ranges from 1.000 (identical objects) to 0.000(maximally dissimilar). Then based on these computed similarity values it begins grouping objects into "clusters", beginning with those objects with the highest similarity values. These clusters are the "classes" of the partitions (defined by the model). Each "partition"

of the study is made up of clusters of objects (or single object clusters) that are connected by a chain of pairs of objects having at least the similarity value defined for that partition. A series of partitions is formed as the similarity value is reduced to bring objects into the clusters sequentially until all objects are grouped in one cluster (the study). The computer simultaneously computes values of "connectedness" and "moat" for each of the clusters as they form. The connectedness value is a measure of the interconnectedness of the objects in the cluster (or the degree of interrelationship). The moat value is a measure of the distance (or difference in similarity value) between the cluster and the next object to join it (or the degree of distinctness of the cluster).

Once the computer has completed its task and printed out the foregoing values, the output is then transformed into a graphical form for easier interpretation. This graphical representation (called subgraphing) enables one to see and follow the "flow" of similarity within the study. Careful study will reveal many relationships between the clusters, and when combined with an analysis of characters for the study, enables one to determine just what properties of the objects causes them to be so related.

From this point on, it is up to the scientist to draw his own conclusions. The method only illuminates the relationships of similarities and differences. What these relationships mean is entirely up to the scientist--who knows his discipline, his purpose and has selected the objects and properties of interest to him.

APPROACH

The first step in an approach to any problem should be to define the problem, determine what is of particular interest about the problem, and to check ones premises in regard to the discipline or disciplines involved. The problem herein referred to was a group of plants collected by my professor in Taximetrics, purported to be representatives of the Family heguminosae, Genus Astragalus. So it seemed they were. My objective was to be a classification of a number of these plants by the taximetrical method I was to learn in class.

Basically, the plants interested me because the genus contains species that are poisonous range forage plants quite unhealthy for cattle, and are widely distributed problem for ranchers in Colorado. This is primarily because the genus contains over 300 species with much intergradation, which makes it very difficult to tell the poisonous species from the not so, or non-poisonous ones. My first steps were to consult respected authorities on the genus such as "The Atlas of American Astragalus" by Barneby and "The Manual of Colorado Plants"by Harrington (containing a key and descriptions by Porter). These works gave me some insight into what characters are important (i.e., help to classify) for the genus. There were of course many such characters, but I selected only 20 (those I felt would be easiest to collect) to keep the study within the scope of the class. Also from my readings in Barneby I discovered that historically the genus had been classified in two ways: one based primarily on gross morphology and a second based primarily on floral morphology. It struck me that it might be interesting to contrast these two ways of classifying the genus to determine if there would be a difference in the results. I hypothesized that given the assumptions of "biological order," a common genetic make-up, and reasonable care in selection and measurement of characters, the

results of two classifications, one based on gross vegetative characters, the other on floral vegetative characters, should not be substantially different, and second that the classification based on the combination of the two sets of characters should be the "best" one in terms of weight of evidence. So, again to restrict the scope of the work involved, I chose ten of each type of character and then 25 of the plant specimens on which I felt the required data could be collected. A list of the characters and characters states selected is included in the appendix of this paper. In collection of the data I discovered that much information was missing on the fruit and flowers as these are very difficult to obtain together. The effect of this missing information became apparent in the analysis of the subgraphs.

Once all the data had been taken from specimens and recorded, it was punched onto IBM punched cards for insertion into the computer. The computer run was made on three sets of data; (1) characters 1-10 for all objects, (2) characters 11-20 for all objects and (3) all characters on all objects, hereafter referred to as parts One, Two and Three, respectively. The computer printout was received and is included in the appendix for reference. Now the process of interpreting the results began. The first step was to "subgraph" the results. This is a graphical means of displaying the relationships between objects of the study. On the subgraph, objects are denoted by numbered circles (numbers identify the individual plants). Within each partition only those objects are connected by a line which have a similairty value at least equal to the "C" value associated with the partition (or level). A heavy line indicates that the two objects first joined (or clustered) at that level (i.e., one of the objects was added to the cluster). A dashed line indicates those objects connected after they joined the cluster. A thin solid line just means they were connected at a previous level. The fraction following each cluster is the "connectedness" percentage of the cluster (i.e., number of actual connections, number of possible

connections) and the number in parenthesis is the "moat" value (or the distinctness") of the cluster. The subgraphs for parts One, Two and Three are included in the appendix.

CONCLUSIONS

In conclusion, I found that due to too much missing information on the floral vegetative characters, part Two of the study did not have good results and therefore could not be fairly compared with part One. However, part Three of the study did have a high degree of correlation with part One, so it may be inferred that part Two did not greatly detract from the combined classification. In fact, the clusters obtained from part Three were wholly contained in those found in part One, and seemed to contain fewer objects of higher similarity. Both part One and part Three divided the study into five main clusters. One cluster being so distinct from the others that it was determined (on the basis of key characters) that this cluster was made up of members of the Genus Oxytropis. The remaining four clusters, though articulated (i.e., joined by commonly similar members) are distinct enough to represent four subgenera of the Genus Astragalus. There were also two individual objects which stood out from all clusters to such an extent that they might also represent distinct sub-genera of the genus. It was also my hypothesis that certain subgroups within each major cluster formed at such high similarity values, and remained distinct for such a great distance that they represent members of species within those subgenera. Later identification of certain individuals within the groups verify that this is the case.

APPENDIX

Astragalus Study - Characters

Gross Vegetative

K1 Overall shape (s)

- A - Tall (12" and up)
- B - Medium (6" to 12")
- C - Low (3" to 6")
- D - Mat-like (up to 3")

K2 Overall Vesture (m)

- A - Glabrous
- B - Puberulent to Pubescent
- C - Strigose
- D - Pubescent to Pilose
- E - Sericeous

K3 Stems (s)

- A - Many (11 or more)
- B - Few (2 to 11)
- C - Solitary
- D - Logical (K5-A)

K4 Stipules (o) k = 1, n = 4

- A - Scarious
- B - Chartaceous
- C - Subfoliaceous
- D - Foliaceous
- E - Logical (K5-A)

K5 Leaves (s)

- A - Radical
- B - Basal
- C - Caulescent

K6 Leaflets (o) k = 1, n = 5

- A - Linear
- B - Linear to elliptic
- C - Elliptic
- D - Elliptic to ovate
- E - Ovate

K7 Peducles (s)

- A - Many (5 or more/stem)
- B - Few
- C - Solitary

Floral Vegetative

K11 Flower Color (m)

- A - White
- B - Creamy to yellow
- C - Yellowish to lt. Blue
- D - Lt. blue to dk. purple

K12 Banner Length (s)

- A - Long (21 to 30 mm)
- B - Medium (11 to 20 mm)
- C - Short (0 to 10 mm)

K13 Banner Arch (o) k = 1, n = 4

- A - Strongly
- B - Moderately
- C - Mildly
- D - Straight

K14 Calyx Shape (s)

- A - Campanulate
- B - Campanulate to cylindrical
- C - Cylindrical

K15 Calyx Teeth (s)

- A - Broad
- B - Narrow
- C - Spinulose

K16 Calyx Vesture (s)

- A - Light
- B - Dark
- C - Logical (K2-A)

K17 Keel Dot (s)

- A - Present
- B - Absent
- C - Logical (K11-D)

K18 Fruit Length (s)

- A - Long (21 to 30 mm)
- B - Medium (11 to 20 mm)
- C - Short (0 to 10 mm)

Astragalus Study - Characters

Gross Vegetative

K8 Racemes (s)

A - Capitate (less than $\frac{1}{4}$ of peduncle)
B - Subcapitate ($\frac{1}{4}$ to $\frac{1}{2}$ of peduncle)
C - Elongate (more than $\frac{1}{2}$ of peduncle)

K9 Caudex Branching (s)

A - Profuse (11 or more)
B - Moderate (4 tO 11)
C - Little (2 to 4)
D - None

K10 Root Thickness (s)

A - Thick (very woody)
B - Average (slightly woody)
C - Thin (slightly herbaceous)

Floral Vegetative

K19 Fruit Curvature (s)

A - Pronounced
B - Slight
C - Straight
D - Sigmoid

K20 Fruit Compression (s)

A - Ovoid (inflated)
B - Terete
C - Lateral
D - Dorsal

K2 Matrix

	A	B	C	D	E
A	1				
B	0	1			
C	0	.50	1		
D	0	0	0	1	
E	0	0	0	.50	1

K11 Matrix

	A	B	C	D
A	1			
B	.50			
C	0	0	1	
D	0	0	.25	1

Paul Holden A

TAXIMETRICS

First, I would like to make the point that numerical taxonomy and taximetrics are not synonymous. Numerical taxonomy refers to classification by a computer, with the biologist's main role that of data collection. Taximetrics on the other hand is a tool that the biologist may use in making a classification. Numerical taxonomy professes to be a "new" method of classification. Taximetrics is not a new method but is simply the one used by biologists for hundreds of years quantified in mathematical terms which can then be expressed numerically and programed for a computor.

How does one go about translating a biologist's thought processes to mathematical terms? Primarily the procedure is one of defining specifically, or as best we can, what we mean by such words and processes as clustering, similarity, character, etc., so a mathematician can define them in his language. First of all we must define what we do in classification. Basically, we select characters that will be used to compare our organisms, do the actual comparison, decide how similar our organisms are, and from this similarity form clusters which are the units of the classification.

Now that we know what the processes are, let us define them so they can be put into mathematical terms. First, what is a character? This we find is a rather nebulous thing that varies from organism to organism, and biologist to biologist. The question of what is a good character for classificatory purposes is another problem, one that is best solved by the biologist. He must know his organisms well enough to be able to say if a character has value although the computor program will often help him in this regard.

Basically a character is a piece of information about your organisms that can be broken into seperate subsets or states. A character is such

that ~~each~~ of the organisms (objects) in the study can be placed in one of the states, and only one. Mathematically, character states form a partition of the study (we may talk of Partitions of Sets only) ~~character~~ for they exhaust or contain all the ~~information in the character,~~ objects in the study and are disjoint, or non-overlapping. A character is a function that assigns one of its states to each object in the study. In a sense each character is a classification for the study. note here that a character is a function - not a set.

Each organism in the study is assigned the state it belongs in for each character. Now the comparison of organisms can be made. Basically what we do here is compare states of the objects in the study. This is where the similarity measure of taximetrics and the computer come into play.

If two objects have the same state for a character, we say they are similar for that character. Let us give a numerical value of 1 for this situation. When the states are different, the objects are not the same, let us express this with a value of 0. Certain situations arise where two states are not exactly similar, but the biologist knows there is some similarity between them. This method allows him to place a value somewhere between 0 and 1 to express this similarity. There are two procedures available, ordering or matrix, a discussion of these is irrelavent here but a knowledge of their presence is essential. Each pair of objects in the study can be given a value in the range 0-1 for each character. Now if these values are summed for each pair of objects over all the characters in the study, and divided by the number of characters used, a value in the range 0-1 can be given to describe the total similarity of the pair. This can be written in formula form as:

$$S(a,b) = \frac{\Sigma S_k(a,b)}{\text{Total \# of characters}}$$

S(a,b) is the total similarity value for two objects, a and b; $S_k(a,b)$ is the similarity value for a and b for each character, k.

The computer goes through and calculates a similarity value for each possible pair of objects in the study.

Once the biologist has figured out which objects are similar, he groups them into clusters. We must instruct the computer to do the same, therefore we define a cluster as a group of objects connected together by some chain of similarity. The machine starts this process by finding the highest S(a,b) in the study, and then picks out all the pairs of objects that similar. It clusters them according to our definition of a cluster. For example, if the first level is .9500 and the pairs that similar are (a,b), (b,c) and (d,e); two clusters will be formed. One contains a, b and c; the other d and e. The similarity value is then lowered until the next pair or pairs will be admitted. These pairs may be between a member of a previous cluster and a new member, between two new members which will form a new cluster, or between objects in two previous clusters, or any combination. This process continues until all the objects in the study are in one big cluster. At each level a value is given for the individual clusters telling how far the similarity value must drop before there is a change in the cluster membership. This value is called moat and is a measure of how far the cluster is seperated from the closest object in the rest of the study. This closest object will be the next one to join the cluster. A large moat value means a group is quite dissimilar from the rest of the study.

Internal connections of a cluster, that is connections between objects already in the cluster, are shown at the levels they appear. The more connectedness a cluster has, the stronger it becomes.

The computer prints out the various levels and what is happening at each. The biologist must interpret the results and decide on the level and clusters he will use in his classification. A method of subgraphing the printout is employed to show, level by level, how the clustering developes. An evaluation of the classificatory value of the individual characters can be made from the printout. This may help the biologist decide which characters are good.

Often, several runs of the study differing in character structure will be made before a satisfactory classification results.

The most important element in this process is the characters, for they determine the results. Characters with many states tend to be splitters, those with few states, lumpers. Ordered or matrix characters, those with partial similarity between states, also tend to be lumpers.

The great value in this method is that it lets you compare many objects using many characters, which would be too tedious to do "by hand". Are there other values?

It can be seen that this technique is only as good as the biologist, for he directs all the action.

It would be difficult, I think, to derive from your discussions a precise understanding of the model - however they serve well to describe the proper attitude which one should take.

A

A LETTER TO JOHN DAVID HOPKIRK
OR
WHAT TAXIMETRICS MEANS TO ME

There is a distinct difference between computerized taxonomy and a system of taxonomy utilizing a computer as one of the tools in the process of building a classification. The former implies that all the biological decisions leading up to a classification are made by a computer; whereas, the latter operation permits the biologist to construct the classification based on his knowledge of his organisms. Taxametrics is such a process.

Taxametrics is a system of classification whereby similarities between organisms are determined by the total number of selected characters between any two organisms. The characters are selected by the biologist and thus are judged more important than other characters for the classification. Weighting of characters can be done another way, by reiterating the character, that is, by splitting a morphological or any other feature into different aspects of that feature and categorizing each aspect as a character.

A character, as defined by Taxametrics, is a "real valued function describing the body" or a way to make an association. It is intrinsic to the animal and is an expression of a gene or a combination of genetic interaction. It will split and group (partition) a number of organisms on the basis of similarity.

A character is divided into several categories called states. States reflect the biologists interpretation of the variation within a character. It is obvious that some states will be well defined (e.g.presence or absence of an adipose fin) and that others are less so(e.g. as in shades of color or in overlapping ranges of lateral line counts between species). Those characters which have distinct states are called simple characters. Those characters which have states which are harder to distinguish, can be coped with through two modes of operation. If the states can be separated into logical sequence, it is placed into an ordered character.

An ordered character cushions the educated guess by providing for a margin of overlap into the neighboring states. If the character shows a reticulate pattern, it is placed into a matrix formula. A matrix character is given values between 0 and 1 for its states, depending upon the percentage of similarity between the relationship of each of the states to each of the other. A matrix formula is similar to the formula one is taught when learning Mendelian genetics. Both matrices? Otherwise not the same

The study will have a number of organisms compared by a number of similar states for a discrete number of characters. The comparisons will be tabulated by a binary computer. Each state of a character will act as a positive value as it is a real valued function. NOT THIS NOT.

A SIMPLE CHARACTER will correlate thusly for two organisms; for any character A with two states, a and b, the value 1 will be given to a perfect match (a,a) and (b,b), and the value 0 will be given to a mixed match (a,b) and (b,a). One will soon observe that a simple character will tend to split if it has a large number of states.

An ORDERED CHARACTER will, depending on the number of overlaps give the value of less than 1 to a series of mixed matches; for example, a character A with the states a,b,c,d and an overlap of 1 will be given the value of less than 1 greater than 0 to (a,b) (b,a) (b,c) (c,b) (c,d) as well as the perfect matches. The value 0 will be given to matches wich are logically separated two sequences away, for example (a,c (b,d). One can thus realize that an ordered character tends to lump depending on the degree of overlap and that an ordered character tends to lump depeding on the degree of overlap and that a high degree of overlap for a low number of states is a poorly differentiated character.

A MATRIX CHARACTER has the value 1 for its perfect matches and values between 0 and 1 for the relationships beween the states; so that for the character A with the relatioships and values of (a,b)=.5,(a,c)=0, (b,c)=.25,

one can see the reticulate pattern. The concepts of the matrix and ordered characters organizes the thinking processes of the biologist about certain characters which I feel helps him analyze them, whereas before he relied on intuition.

The output from the computer organizes the data into a hierarchical classification. A hierarchical classification can be explained in terms of set theory. A set is composed of any number of objects. These objects may be recognized as members of a collection. A deck of poker is a set composed of 52 members. A set can be composed of subsets that are exclusive and exhaustive. A deck of poker is composed of four suits and all the fifty two cards are exhausted into those four suits no one card belonging to any two suits. The 52 cards are partioned into the four suits. In the same manner an organism is partioned into species, genus, family , order, class, phylum.

The computer is like an electronic sorter, sifting the organisms(as represented by computer punch cards) into partions. It will first seek to find organisms which have identical states for each character in the study. If failing to find such a comparison, it will seek out the organisms with the highest number of identical states over the entire study. It will cluster the organisms accoording to their relations. This level is called a c level and it is somewhat analogous to a percentage because the level of similarity is determined by the number of states ofscharacters without missing information that are similar for a number of organisms OVER the total number of states of characters in the study without missing pieces of information. The computer continuously lowers the level of similarity until an organism clusters to a new object. The amount in the level of similarity which is lowered for that organism is called a "moat" and is de-

fined as the amount of similarity by which space that the closest organism to the cluster fails to qualify for membership. The computer drops the level in similarity until it has included every object (organism) into a large set(exhausted the study) composed of a number of subsets or partitions (made exclusive) which the biologist organizes into taxa.

Several techniques have been devised to help the biologist organize his taxa from the computer output. The subgraphing technique is a graphical representation of the coutput. Lines are connected between similar objects. Each connection represents a relation so that the more lines connected between objects in a cluster, the more related they have become. This is also a graphical representation of partitions. From the graph the biologist induces taxa, based on similarity levels where strong clusters formed, the moat surrounding the clusters and the connectedness of the clusters.

The classification may not satisfy the biologist; if so, he can analyze which characters were most important in forming connections and clusters, and can rearrange the states within the character to best present his views on the organisms. This is not cheating, for often, certain characters are badly represented by the biologist or have confusing states or, as ~~in~~ with ratios, which refect allometry, spurious information. In fact the biologist can strenthen important characters by this form of analysis by better definition.

The methodology of taxametrics is an important, powerful tool of the biologist and thus the biologist serves more than just an information, data gathering slave for the computer. Good discussion from a Biological point of view. You should be more careful in choosing words which mean what you want to SAY. Think about the precise definitions of words so that you don't SAY meaningless or poorly defined things.

A

An Explanation of

Taximetrics-Tool for the Biologist

May 23, 1967

Judy Li

Taximetrics is a methodology employed as a useful tool for classification of biological organisms. The groupings established are based on an heirarchy; that is, those objects which are similar at lower levels continue to be grouped together in higher taxa. The method depends on certain rules, often expressed in mathematical terms. Furthermore, the biologist's interpretations of the input data and print out from the computer, are essential elements of the method.

In order to group or separate objects in a classification, comparison between those objects is most conveniently made on the basis of degrees of similarity. For discrete criterion for similarity measures biologists use characters. In this method we consider a character as a function (see fig. I). The domain is the set of objects to be studied, the range the list of values determined for each character; the character assigns a particular value for each object. Figure one illustrates how

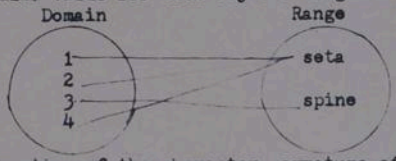


Fig. I Function of the character, armature of the 4th swimming leg in Cyclops vernalis.

the character, armature of the fourth leg of copepods in my study, assigns a value in the range to the domain of the study. Here, animals one and two and four possess setae, while animal three has a spine.

There are three important properties of a character. The first is that it defines a symmetrical relationship, that is, if object one is in relation to two, then two is in relation to one. Therefore if one in the sample study is in relation to two defined by the character of leg armature (both have seta), then two is in relation to one. (We see that relative to this character, two and three are not in relation). A second property is

that a reflexive relationship is established; an object is in relation to itself. For example, one is in relation to itself. Thirdly the relationship defined by a character is transitive; that is, if one is in relation to two, and two in relation to four, then one is in relation to four. In our illustration, if four possesses a seta, it is in relation to one; since one is also in relation to two, then two is in relation to four (all have setae).

When these three properties exist (symmetrical, reflexive and transitive relationships) we call it an equivalence relation. This is important because it follows in mathematics that a character incorporates the entire study and divides it into non-overlapping sets. In the illustration, one, two, three and four are all the members of the hypothetical study. The character has applied to all members of the study: it is exhaustive. Because non-overlapping sets are established, it is a disjoint property. A general term for these qualities, disjoint and exhaustive, is a partition. Each character we employ is a partition. To the biologist

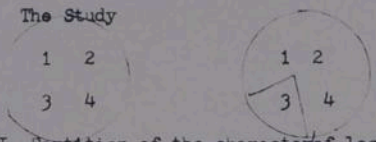


Fig. II Partition of the character of leg armature.

the discrete assignment of a character value (or state) to each object is pleasing, though sometimes difficult, because it necessitates clear thinking, avoiding vague terminology.

After an object is described by a particular character it can be compared to other objects. We do this using a similarity function. All pairs of the study are compared and assigned a value of similiarity between zero and one. If object j is identical to object i for a given

character, then the similarity function for i and j is one(S (i,j,)=1). If they are completely unlike, S(i,j,)=0. To compare objects for all characters used in the study, the average similarity is found. The formula $S(i,j,)= \frac{S_c(i,j,)}{\text{total \# characters}}$ expresses this average. When S(i,j)=0.9, then objects i and j are similar in 90% of the characters. On the "print out" from the computer, there are two tables which tell the function S(i,j). Actually Provides a List of all pairs together with how closely they are related (S (a,b), One provides a list of pairs which are most closely related. For example, if object a is more closely related to c than to any other object in the study, this pair, S(a,c), will appear on the table along with the similarity value at which tney were brought together (a value between zero and one). The second tabulation of S(i,j) indicates the ten closest objects for each object in the study, along with the similarity value for each pair. These are given as nodal distance arrays.

When the biologist assigns a character state to an object he has many ways in which to do it. Some characters, such as the example given earlier with setae or spines, are very straight forward. This kind of character is called a simple one. Generally most characters are this type. There are other situations in which two objects being compared may be neither absolutely identical nor unlike. Linear measurements are an example of this condition; shorter objects may be more related to medium length objects than they are to longer ones. In this case we might make the character an ordered one. Generally we "order" states adjacent to each other; then short and medium lengths are in relation but not short and long. The similarity value for this condition would be one for object pairs assigned the same character state (if they are both short, for instance), and zero for those not in adjacent states (a short and a long object). An intermediate value can be assigned to the adjacent states. Another char-

acter assignment for non-simple characters is a matrix, in which values of relative similarity are given an appropriate value by the biologist.

The establishment of non-simple characters sounds like a good way to give a pair that are somewhat alike a similarity value proportional to their relationship. However these must be employed with caution. While ordered and matrix characters do tend to group pairs, at the same time they do not allow the given character a good opportunity to make a clear division between objects. Therefore if I am certain about a very gradual change in linear measurements or hesitant about possible accuracy in measurement the ordered character expresses the graduations rather than distinct states. if distinct linear states are observed, such as short, medium and long, they are best left as simple rather than ordered characters. Predominant use of simple characters produces classifications that are easier to analyze.

An important attribute of this method is that all characters are of equal weight. A biologist observes that two organisms which are very similar have many attributes in common. Taximetrics works in the same way. Objects that are similar in many respects are clustered together more closely than those with which they share only a few common points. Thus, if two organisms are closely related, the fact that they are of different lengths (and coded into different states of a simple character) will be overidden by other, more numerous, similarities.

we have to be careful about this as weight is not always a well defined concept.

Sometimes information about one character is unavailable for an object. This is coded as "no information" and the object is compared to others on the basis of only those characters in which information is provided. Using such objects can produce confusing clusters. An object with missing information tends to bring together very early objects which are like it in all characters with information present. It might be

that the objects brought together are actually very different with respect to the characters in which information was missing for one; the similarity value could be much lower than the one indicated. Another circumstance arises when a particular character description is not logical. For example, if the swimming leg armature were a spine on #3 of the illustration given earlier, then length of the seta for this leg would not be logical. In such a case, the character is coded as "not logical".

The results from the computer are called the "print out". Levels of similarity appearing on the print out depend upon total similarity of all characters for which information is present for pairs of objects. Those objects which are most similar appear first. Each level of similarity is given a value, called the "c value", which represents the degree of similarity between zero and one. The first level, c_1 can be 1.0000 or less. Each time a cluster of objects acquires a new member a c level is indicated. Since the earliest levels express the most similarity, once a cluster is formed, all members remain members of that cluster. The similarity decreases in later levels because new additions are less similar to the earlier members. Therefore members forming at c=0.9 would be much more like one another than all the members of the larger cluster at c=0.6. Because the degree of simlarity determines the members of clusters, objects quite unlike remain distinctly separated for quite awhile, and those which are very similar are grouped together earlier.

Degrees of separation between a cluster at any level and the next object to become a member of the cluster is given as the moat. This is actually a measure of how much a pair of ~~objects~~ (clusters) are unlike. On the other hand, it is also desirable to show interconnectedness within a group. This is given as the number of internal connections within a cluster, out of the total number of possible connections.

Just as the initial information about characters is determined by the biologist, in the same manner the interpretation of the output is also up to him. Since the print out groups distinct clusters, which are then contained within larger clusters at less similar levels, this method produces an heirarchial classification. But merely grouping objects of smaller, more related, sets into larger clusters does not define what taxonomic levels we are working with. Graphing the study by connecting objects as they come together at different levels, produces an overall picture of how the classification develops. A skyline is another kind of graph which shows the moats, that is, degrees of separation, very well. While looking at the total study, the biologist also observes carefully the relationships of individual objects. In the latter kind of analysis the S(i,j) table and the nodal distance arrays are very helpful.

Usually the kinds of information used are closely regarded and revisions made. These changes often help to make more distinct clusters, or clusters which better show what the biologist considers to be more valid relationships based on what he knows about the organism and its biology. Many new relationships between objects, especially those which might connect two clusters, can be discovered. Furthermore, the biologist can increase the amount of information in later studies (for example, genetics or ecology) in order to produce a general classification (not based on only one type of character such as morphology). Finally, based upon the distinctness of the clusters formed as well as the biology of the organism in general, the biologist determines the level of taxa upon which he is working.

These discussions indicate an excellent understanding of the graph theory model and how it works. Perhaps the class discussions of the past few days will suggest to you how to think about why it works.