TAXON 24(5/6): 603-608. NOVEMBER 1975

A POSTERIORI WEIGHTING OF DESCRIPTORS

*Pierre Legendre**

Summary

The Adansonian and the more traditional taxonomic principles may be combined in an iterative information-seeking strategy of classification. Uniqueness is not a property of biological classifications, and even the wishfully fundamental principle of parsimony has not brought to taxonomy the strong basis on which uniqueness could rest. On the other hand, it can be shown that descriptors have different values along the scale of taxonomic categories, and that different descriptors have different values for the same partition. It is then suggested that an ideal taxonomic treatment should begin with a large body of equally weighted descriptors, and lead to deletion, re-structuring and *a posteriori* weighting of the descriptors, through recognition of the main clusters and a measure of the contribution of each descriptor to each partition of the objects. The analysis could then continue, for each partition independently, by an iteration between clustering, discrimination and weighting, until stability is reached.

Résumé

Il est possible de combiner les principes d'Adanson avec ceux de la taxonomie classique, en une stratégie itérative qui améliore à chaque cycle la classification produite. Les classifications biologiques n'ont pas la propriété d'unicité; même le principe de parcimonie, que l'on a voulu fondamental, n'a pas apporté à la taxonomie cette base sur laquelle on aurait pu rechercher des classifications uniques. D'autre part, on peut démontrer que les descripteurs n'ont pas la même valeur tout au long de l'échelle des catégories taxonomiques, et que différents descripteurs ont aussi des valeurs différentes pour une même partition. Un traitement taxonomique idéal devrait donc débuter avec un grand nombre de descripteurs à poids égaux, mais permettre par la suite de laisser tomber, de re-structurer ou d'attribuer *a posteriori* des poids aux descripteurs, après que l'on ait reconnu les principaux groupes d'objets et que l'on ait mesuré la contribution de chaque descripteur à chacune des partitions des objets. L'analyse pourrait alors se poursuivre de façon indépendante pour chaque partition, par un cycle: groupement des objets, discrimination, et modification du poids attribué à chaque descripteur, jusqu'à ce que l'on en arrive à une solution de stabilité.

Introduction

As a contribution to the discussion about the validity of the Adansonian principles for numerical taxonomic studies, we would like to present hereafter a reconciling point of view concerning an iterative process that achieves stable classifications. Indeed, in the past decade or so, there has been an increasingly large body of conflicting literature relative to this quarrel, which could be resolved by considering the positive contribution of each of the opposing strategies to stabilization of a classification.

* Centre de recherche en sciences de l'environnement, Université du Québec à Montréal, B.P. 8888, Montréal, Québec, Canada.

The opponents

The tenants of a pure Adansonian strategy argue that an objective classification can only be reached if the information is uniformly manipulated. Since the information is found in the descriptors used, a usually large body of descriptors is assembled, and the states of the meristic and metric descriptors are assigned by dividing the range of variation into a number of equal segments. After perhaps a scaling operation (normalization, standardization, etc.), all those descriptors – the most numerous, the best – are used in a non-weighted manner to obtain a classification. It is argued that this constitutes a *robust* strategy, since it leads consistently to about the same results with the same objects, whatever clustering procedure or descriptor-set might be used.

Members of the more traditional school of taxonomists counter with the argument that these results are rather a clear demonstration of the robustness of the evolutionary-systematic structure investigated, which manages in some way to survive in spite of all the bad treatments the information has to suffer. Those who share this way of thinking prefer to select "good" characters before proceeding with the similarity computation and subsequent clustering. One way to establish the "goodness" of descriptors is through a preconceived idea of the classificatory value of descriptors, which was the rule in the early ages of taxonomy. Since this preconceived idea originates, at least partly, through intuition or by a decision as to what the final classification ought to be, this approach can be, with reason, accused of circular reasoning. But the goodness of descriptors may also be established through a descriptor analysis, which is the analysis of the structure shared by groups of descriptors through parametric or non-parametric correllation analysis, information-theoretic redundancy measure, principal components analysis, and the like, after which the choice of the descriptors can be determined and their states-structure modified. It is probably through such procedures, carried on by an intuitive process by outstanding pattern-analysts, that good classifications have been produced in pre-computer times, classifications that have survived generations of revisers as well as modern computer testing.

These instances of "good" taxa may lead the taxonomist to believe that classifications are unique, which needs to be discussed. But first we have to look at the steps involved in classifying.

The classificatory algorithm

We can define a classification for a collection of objects (OTU's) as a hierarchical, two-dimensional partitioning of the objects, which characteristically groups in the same sub-collection, for any given partition, the objects obeying the similarity rule which has been chosen and defined. Furthermore, sub-collections of a given partition are isolated from one another.

This definition leaves three variables open: the choice of the similarity rule, of the clustering strategy, and of the information which will be used to represent the objects. *Represent,* because we cannot use the objects themselves in the classificatory process, and so we have to choose a certain finite number of descriptors (characters, variables, attributes) to represent them, and eventually we have to structure the states of those descriptors in a given way which will remain the same all through the classificatory procedure. The states may be measurements or meristic data, grouped or not, qualitative descriptions of a character, or yes-no data (cardinal, ordinal, interval, nominal or binary).

Interestingly, it is the choice and structure of the information to be used where most of the discussion occurs, while on the contrary, one may use whatever similarity-rule or clustering-strategy he wishes without generating too much protest against his classification. This may be because the criteria for making a decision as to what similarity-coefficient and clustering-model to choose, are not based upon general principles on which taxonomists could rely. In the case of the

structure of descriptors on the other hand, the neo-Adansonian principles, as summarized in Sneath and Sokal (1973: 5) form such a basis on which the discussion can be built. Of course, the clustering strategy as single linkage, complete linkage, or the intermediate solutions proposed by Sokal and Michener (1958), Sneath (1966), Lance and Williams (1967), Clifford and Goodall (1967 – corrected and programmed in L. Legendre, 1971), and Shepherd and Willmott (1968), is chosen in relation to the problem at hand, and opposite strategies may give complementary information about the cluster structure. In the same way, the choice of the similarity rule depends on the information structure, since some coefficients are applicable only to binary data, while others do better with continuous or with discontinuous data; it depends also on the nature of the problem, since, for instance, ecology often requires different coefficients than those used in taxonomy. However, as Reyment (1970: 68) puts it, "if one examines the logic behind the concept of the similarity coefficient of the numerical taxonomists . . ., it soon becomes apparent, that this is an area in which personal opinion is permitted considerable rein".

Uniqueness

Let us come back to the question of the uniqueness of classifications, and in particular morphological classifications, since they are the ones concerned with the classificatory algorithm described above. We are not speaking here about the biosystematic strategy of classification which has its own similarity rule based upon reproductive isolation.

An entity is unique in a given set, relative to given conditions, if, and only if, it is the only member of the set satisfying the conditions. The conditions are then said to determine the element. So, for all the various ways of arranging a group of biological objects, which would also correspond to the definition we have given of a classification, does there *exist* a set of biologically defendable rules which would determine that one and only one of these arrangements is valid? The answer is No. Indeed, there is no *biological* law which could help us to choose between two similarity rules, for instance between the mean character difference rule and Goodall's probabilistic similarity index. The choice of a similarity rule is based on other considerations, the structure of the information available, for one thing.

In the same way, there is no better biological reason to use a large body of descriptors divided arbitrarily into states of equal width as there is to rearrange descriptor states to get a better redundancy of the information content, for example. And the reason for this is that taxa are not being formed after evolution of all characters at an equal rate, but rather by a cytogenetic process which has little direct relation with the morphological modifications which usually accompany or follow the appearance of a new taxon. This may be the reason why the principle of parsimony (which was first presented as a representation of *the way biologists thought* about evolutionary history, and not as a biological principle), otherwise quite appealing to the evolutionary biologist, has never been supported by independent evidence (Cavalli – Sforza and Edwards, 1967; Sneath and Sokal, 1973: 322). It is even disputed by evolutionists and numerical phylogenists, who argue that in real cases, evolution may at best have proceeded in the neighborhood of the most parsimonious path. On the other hand, the wishfully fundamental principle of parsimony, when made operational, can often lead to a multiplicity of computationally equivalent optimal solutions (for example in Estabrook, 1968), showing its incompleteness as a basic rule of evolution. So, as long as we are talking about morphological classifications, we cannot talk of uniqueness, since there does not exist a single set of biologically defendable rules to establish a unique morphological classification. Uniqueness may be sought – at least potentially – only when we consider biosystematic properties like reproductive isolation.

Classificatory value of descriptors

Another way to look at the problem of the legitimacy of reworking the information structure of the descriptors is to wonder whether all descriptors can really be thought of as having the same classificatory value. Actually, it is common experience to realize that while certain descriptors have a splitting effect at the subspecific level, the splitting is generated by another group of descriptors at the generic level, for instance. This phenomenon is so well known that it has been codified in the old rules of thumb as to which kind of descriptors one must use in the lower, intermediate, and higher categories.

The phenomenon is easy to test and quantify: given a tree-like classification of a group of objects, let us consider each partition, or category level, just as a descriptor, that is, a single basis of comparison for all the objects under study. This partition, as a descriptor, can be compared with all the descriptors used in arriving at the classification. Several methods are available for so doing: we could use, for instance, coefficients like Kendall's *Tau* or Spearman's *r*, or we could measure the information they share through conditional entropy. Whatever the descriptor analysis tool we use, it becomes possible after such a study to rank all the descriptors for their contribution to the given partition. It becomes clear then that all the descriptors do not have the same classificatory effect for a given category level, and that the same descriptors vary as to their classificatory power on the category scale. Accordingly, the so-called non-weighting strategy, which consists actually in attributing a weight of exactly 1.000 (Moss, 1972) to every descriptor, has no better theoretical grounds than weighting in any other way.

Non-weighting or equal weighting, *as the final strategy* of a classificatory process, is strictly applicable only when in a small taxonomic group, all components have diverged at exactly the same rate with respect to aspects of their morphology.

The way around these difficulties used by the statisticians has been to consider a much larger group of descriptors than used by the classical biologist, with the hypothesis that, if one uses a large enough body of descriptors, the "splitting" and "lumping" inherent in and due to inappropriate choice or structuring of certain descriptors, will cancel each other and what will emerge will be the "true" taxonomic structure, more descriptors leading to stability.

But if we consider for a moment well-known biological phenomena such as cryptic taxa, or, on the other hand, polymorphic taxa, such a treatment would not lead to the "right" structure, at least from the phyleticist's point of view. In this case, proper weighting seems essential to get the right relationships.

Another advantage is that a larger body of descriptors is less likely to convey the idiosyncracies of a worker than a small number of descriptors carefully chosen and structured. Others advocate however, that a "competent biologist" can do a better job at discovering and meaningfully structuring the information characterizing the objects, since he can use his experience and that of his colleagues for doing so.

Iterative procedures

We believe that those seemingly diverging points of view represent in fact two phases of a taxonomic treatment, and that an ideal procedure would iterate between clustering of objects and re-examining the structure of the descriptors, beginning with a non-weighting cycle (large body of equally weighted descriptors), and going on to reworking the taxonomic information by deletion, restructuring descriptor states, and eventually weighting *a posteriori*.

Even though we are not always dealing with cryptic, polymorphic or otherwise "head-shrinking" problems, we have to consider that the ideal situation for non-weighting, mentioned above, is not encountered any more often, so that a certain amount of weighting might always be appropriate, althoug not always essential because of the robustness of the underlying biological phenomenon. To

achieve this weighting, it may be necessary in extreme cases to bring in biosystematic evidence, but we would like to submit that in most cases we can do without it, and that an iterative classificatory procedure might enhance the already existing automated processes enough to produce phenetically based classifications which are also phyletically meaningful. At this point, of course, we have lost our distinguished phenetically-minded colleagues who, like Jardine and Sibson (1971), believe that a phenetic classification cannot also convey phyletic information.

"Many numerical taxonomists express almost as a dogma the requirement that all attributes be equally weighted. If their meaning is examined more closely, however, it is found that they are really arguing against arbitrary weightings depending on preconceptions as to the taxonomic value of different attributes, rather than for strictly equal weighting" (Clifford and Goodall, 1967: 503). Indeed, numerical taxonomy is meant to be an objective process, and the computer must not be used as an excuse to camouflage an already preconceived classification. Computer processing does not make *per se* a method more objective: only the algorithm can be thought of as more or less objective.

There is no question that no-weighting is quite acceptable at the beginning of the classificatory process, to clear up the picture. It is important to recognize at this point, however, that even though the characters may be unweighted, or equally weighted, the information they carry is weighted in varying degrees, depending on it being common to several descriptors (Legendre and Rogers, 1972: 585). An example of an information-generating iterative procedure would be as follows:

I- A numerical taxonomic study could begin with a study of the structure of the information shared by groups of descriptors by parametric or non-parametric correlation analysis, information theoretic redundancy measure, or the like. This is followed by a first structuring of the descriptors into states, trying to make all the pieces of information as equally weighted as possible. This could also mean, for instance, eliminating those descriptors that are largely or completely redundant to others. This is followed by a first clustering, based on largely unweighted information through an appropriate similarity measure.

II- In a second step, all the descriptors are taken back into the study, and only the objects which cluster nicely at the levels considered to correspond to the main partitions (systematic levels), are used as the nuclei of a discriminant or canonical analysis. This gives a weight to each descriptor, for every partition considered. Measures of correlation or of information theoretic redundancy could be used in place of discriminant analysis to determine how much each descriptor contributes to the given partition.

III- This is followed by a clustering for each one of the partitions, weighting the descriptors according to the way they contribute to the given partition. Descriptors which do not contribute significantly to a partition can be eliminated or re-structured before this clustering phase. All the objects are used here, and those which had been eliminated from step II should be expected to cluster more easily now, if they do pertain to the clusters so defined.

IV- Steps II and III can be repeated until stability is reached. A test of stability could be based on 1) the elements present in each cluster at iterations (i) and (i + 1), and 2) on a measure of the information shared by the same partition at iterations (i) and (i + 1).

Conclusion

An iterative procedure such as this one has several advantages over the more traditional batch treatment of the information.

1. It has its own intrinsic logic, which uses and reconciles seemingly contradictory points of view about weighting or non-weighting of information, in a progressive iterative procedure.
2. It goes beyond the myth that complete information can be obtained about a

structure after running it once through a single algorithm. Modern taxonomy is usually based on so many factors that there is no unique analytical solution, and consequently we have to search instead for an optimized solution, which can best be attained through iterations.

3. Since different descriptors have varying discriminating powers in different partitions (systematic levels) of the same objects, it is obvious that a better classification will be obtained if each partition is considered independently from the others, and the best structure of the descriptors is established.

4. In taxonomy, it may not always be essential to use such a procedure, because of the hardiness of the underlying biological phenomenon. Indeed, in taxonomy we are trying to recognize existing units which are the results of evolution, and those units may well manage to emerge even though the information is not optimally handled. With other types of classifications, in ecology for instance where the units to be recognized are at best probabilistic, such an iterative processing may be essential in the recognition of meaningful clusters.

References

CAVALLI-SFORZA, L. L. and A. W. F. EDWARS 1967 – Phylogenetic analysis: models and estimation procedures. Evolution 21: 550-570.

CLIFFORD, H. T. and D. W. GOODALL 1967 – A numerical contribution to the classification of the Poaceae Aust. J. Bot. 15: 449-519.

ESTABROOK, G. F. 1968 – A general solution in partial orders for the Camin – Sokal model in phylogeny. J. Theoret. Biol. 21: 421-438.

JARDINE, N. and R. SIBSON 1971 – Mathematical taxonomy. John Wiley & Sons Inc., New York. xviii + 286 pages.

LANCE, G. N. and W. T. WILLIAMS 1967 – A general theory of classificatory sorting strategies. I. Hierarchical systems. Computer J. 9: 373-380.

LEGENDRE, L. 1971 – Phytoplankton structures in Baie des Chaleurs. Ph. D. dissertation, Institute of Oceanography, Dalhousie University, Halifax. 137 pages.

LEGENDRE, P. and D. J. ROGERS 1972 – Characters and clustering in taxonomy: a synthesis of two taximetric procedures. Taxon 21 (5/6): 567-606.

MOSS, W. W. 1972 – Some levels of phenetics. Syst. Zool. 21 (2): 236-239.

REYMENT, R. A. 1970 – Eigen-theory in numerical taxonomy. Bull. geol. Instn. Univ. Uppsala N.S. 2: 8, 67-72.

SHEPHERD, M. J. and A. J. WILLMOTT 1968 – Cluster analysis on the Atlas computer. Computer J. 11: 57-62.

SNEATH, P. H. A. 1966 – A comparison of different clustering methods as applied to randomly-spaced points. Classification Soc. Bull. 1 (2): 2-18.

SNEATH, P. H. A. and R. R. SOKAL 1973 – Numerical taxonomy. The principles and practice of numerical classification. W. H. Freeman and Co., San Francisco. xv + 573 pages.