The Development of and Rationale for Algorithmic Classification in the Biological Sciences

W. A. Green

Abstract

Biological scientists generally base their classifications on what they call 'natural' groups, or entities that are supposed to exist in the physical world independently of whether or by whom they are described or classified. Geologists are more prone to classify the things they work with on the basis of explicitly arbitrary divisions along important gradients. Why are biologists so concerned with 'objective' classification? Is their search for 'natural' groups meaningful and important (as it has been assumed to be for the past half century), or is it based on a confusion about what classification is?

Introduction: what is classification?

The word *classification* has been used in two fundamentally different but seldom explicitly distinguished ways. On one hand the concept of classification as analogy seems to be a necessary prerequisite for learning and abstract communication in humans and other animals; on the other hand the notion of set-theoretic classification is a formal abstraction that does not seem to date to much before the 19th Century. The similarity between the duality of classification and the duality of probability discussed by Hacking (1975) may be more than a trivial coincidence.

This duality of classification has been recognized in biology ('Philosophers as well as taxonomists have realized almost from the beginning that classifications serve a dual purpose, a practical and a general (that is, scientific or metaphysical) one.' Mayr 1982:148) yet the two meanings continue to be mixed up in debates over classificatory procedure.

The first concept of classification, which can be called *hermeneutic* because it generally involves an interpretive (as opposed to analytic or explicitly algorithmic) procedure, clearly underlies abstract thought, language, and in fact, learned behavior in general. For instance, a dog can only be taught to sit on command if he is capable of grouping together the sounds that the command 'sit' produces when articulated by different people at different times. Empirically we seem to find that the variation in sound between different articulations of 'sit' is small enough that it is generally resolvable from the similar cluster of sounds produced by articulations of the word 'come'. So an average dog can correctly cluster variations on /sɪt/ with respect to /kʌm/ but would probably not distinguish /sɪt/ from /ʃɪt/, which the average English-speaking human generally does.

The use of a linguistic example for illustration is not arbitrary: linguistics is a field in which the duality of classification has been discussed in some detail. The terms *phonetic* and *phonemic* refer respectively to the sounds that it is possible to make verbally (and distinguish aurally) and to the sounds that convey differences of meaning in a particular language. For instance, the difference between the International Phonetic Alphabet sounds /y/ and /u/ is non-phonemic in English: that is, we would not distinguish between the word 'food' pronounced as /fud/ (the standard English pronunciation) and /fyd/ (with the vowel brought from the back of the mouth to the front but still rounded and closed). In French, on the other hand, this distinction produces the distinct words *vu*=seen and *vous*=you. By analogy with this distinction, Pike (1967) coined the terms etic and emic for attributes that are respectively extrinsic (objective) and intrinsic (subjective) to a society.

An etic classification is objective in the sense that it comes from outside the context in which it is used while an emic classification is context-dependent. It is reasonable to equate hermeneutic classification at least roughly with Pike's version of emic classification.

In comparison to this hermeneutic concept, there is the idea of set-theoretic classification which developed along with axiomatic logic and set theory from the time of Peano (1922), who introduced the symbol $\in$ to indicate set membership in his axiomatization of set theory. Under this etic concept of classification, a classification is a formal process of assigning elements to sets (or classes) under the restriction that:

$$\forall A, B : A \in B \vee A \notin B,$$

i.e. for all As and Bs, the entity A is either a member of a class B or it is not a member of the class B. This is logically equivalent to the statement that the entity A has the property B or does not have the property B, and under classical logic is assumed to be analytically true. (Despite the developments in axiomatic set theory between Frege (1893) and Cohen (REF), some concept of set membership is basic to all forms of set theory. As Zermelo stated it in 1906 (translated in Moore 1982:156), 'The property that an object a is an "element" of a set M is treated as a primary fundamental relation.')

We see, however, from the history of usage in the Oxford English Dictionary, that the word classification appeared at the very end of the 18th Century along with high enlightenment thought, and considerably before set theory:

> 1. The action of classifying or arranging in classes, according to common characteristics or affinities; assignment to the proper class. 1790 BURKE Fr. Rev. Wks. V. 332 Montesquieu observed very justly, that in their classification of the citizens the great legislators of antiquity made the greatest display of their powers. 1804 ABERNETHY Surg. Observ. 18 In attempting a classification of tumours. 1847 CARPENTER Zool. 2 The object of all Classification..[is] to bring together those beings which most resemble each other and to separate those that differ. 1874 BLACKIE Self Cult. 19 Nothing helps the memory so much as order and classification. 2. The result of classifying; a systematic distribution, allocation, or arrangement, in a class or classes; esp. of things which form the subject-matter of a science or of a methodic inquiry. 1794 SULLIVAN View Nat. II. 196 De Saussure gives us this brief classification of volcanic substances. 1834 J. M. GOOD Study of Med. (4th ed.) I p.x, A syllabus of its classification for the purpose of lecturing from. 1856 SIR B. BRODIE Psychol. Inq. I. vi. 230 The classification of faculties which these writers have made is altogether artificial. 1860 MAURY Phys. Geog. Sea. xi. 505 Red fogs..do not properly come under our classification of sea fogs. Mod. Several classifications have been made.

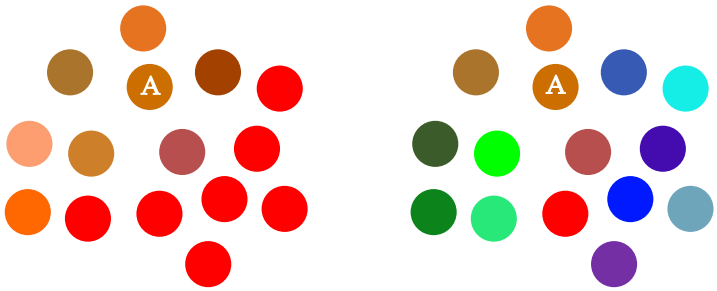
It seems somewhat strange that this precise set-theoretic notion of classification has come to be the dominant one in biology but this is clearly the case: an organism either is or is not a member of a species; there is generally considered to be no middle ground. The assumptions of naive set theory, of course lead to a number of contradictions (like Russell's Paradox), so biologists have been concerned with avoiding these by separating individuals (elements) from sets (classes):

'When rigid criteria for the limits of taxa are established, it is an easy thing to lose sight of the fact that organisms, as evolving entities, should not be regarded as fixed individuals.' (Delevoryas 1964:35)

'A species, indeed, is always a particular (individual), never a class.' (Mayr 1995)

As can be seen from these quotations, however, there has been little consensus in biology about what is really being classified, not to mention how the classification should be carried out.

At any empirical level, also, it remains far from universally evident what is meant by a set, still less that set membership is necessarily bivalent (taking only two values). Most current variant axiomatizations of set theory (von Neumann-Bernays-Gdel, Zermelo-Frankel, New Foundations, etc.) while avoiding formal paradoxes, nevertheless restrict themselves an empirically (psychologically) vague notion of what a set is and to bivalent notions of set membership. Assuming that $\forall A, B : A \in B \vee A \notin B$ is analytically true is very convenient as a method of constructing numbers and arithmetic from primitive ('self-evident') logical assumptions, but it unfortunately seems less effective when used as tool for describing empirical data from the natural world, much less when used to describe such data filtered through the observational biases of human observers. For instance, for the set of all red things R and the set of all maroon things M, it is not at all clear that M is a subset of R or that is is not a subset of R. This sort of difficulty is studied as the phenomenon of 'vagueness' in philosophy (Raffman 1994) and is not resolved by description of 'red' light as a band of wave-lengths on the electromagnetic spectrum.



Is the circle labled A one of the red ones?

It is possible that our mental predisposition in favor of dichotomies is to blame, as some biologists have argued: 'Our cultural bias to categoized things discretely, to fit even continua into pigeonholes, is at fault. This bias extends even into mathematics and is why ordinary set theory is partly inappropriate for biological systematics' (Van Valen 1988:55). But it seems more likely that this insistence is a result of the inheritance of the set-theoretic notion of classification from logic and mathematics. Clearly our mental propensity is to classify things hemeneutically and be quite comfortable with the idea that maroon is red compared to green but maroon compared with cherry-red, apple-red, and candy-red.

One solution is the adoption of fuzzy set theory, generally attributed to Zadeh (1965) though a biologist has also laid claim to the idea: 'In 1964 I proposed what were soon aptly renamed fuzzy sets' (Van Valen 1988:56). Fuzzy sets, where set membership is described by a probability has already been shown to have some application in clustering applications (REFS), and their realm

of applicability may continue to grow. Neverthless, the basic incompatibility may remain between set-theoretic classification as a precisely defined axiomatic property in a formal system, and any practical notion of classification of emiprical data.

The problem seems to be that almost all explicit classifications we use are perceived to be set-theoretic in nature and therefore are only strictly relevant to abstract sets. It may therefore be unreasonable of us to be surprised when we find they in fact bear only an analogical relationship to the empirical data from which they are constructed. Actually, there is an analytic method for producing hermeneutic classifications: Pavlovian conditioning. By applying the appropriate stimuli to subjects like dogs or graduate students, it is clearly possible to produce a belief in the relationships between otherwise unrelated phenomena.

Classification in the biological and geological sciences

Until the advent of algorithmic classification, this was in fact the dominant procedure. For a student learning to identify or classify any sort of objects from the natural worldand many man-made artifacts, for that matterthere is a common progression. In the beginning, all the objects look the same. Then one slowly learns to recognize a few of the most common kinds and may, as each new type is recognized, for a time identify most things as that newly-recognized type. As the number of recognizable types increases, the student begins to have an opinion about the parameters of the various types, how clear-cut the lines are that divide them from each other, and with what degree of reliability each can be recognized. This learning process, some variant of which has surely been experienced by anyone who has dealt with large numbers of similar objects whether they are sports cars, fossil brachiopods, or impressionist paintings, is usually mediated by a teacher or teachers; learning to identify objects is more a skill to be learned than a concept to be understood, and as such is is better taught through personal instruction than through books or articles. Thus, in a scientific field like systematic biology where identification of objects has been central in both practice and principle, there is at some level a reliance on expert opinion. Until relatively recently there was little criticism of this procedure of passing down expert knowledge from teacher to student and in any generation allowing seniority to set the standards for identification and classification of organisms. But around the middle of the twentieth century, serious questions began to arise about the logical validity of such a procedure: if an expert and a neophyte look at the same material and produce different classifications, is the expert right and the neophyte wrong, or has the expert merely disgorged established dogma while the neophyte takes a fresh, unbiased look at the material? Furthermore, no two experts will entirely agree on a classification and consensus carries the same disadvantages as reliance on received wisdom.

Between 1950 and about 1970, two schools of thought in systematic biology, phenetics and cladistics, arose to address this issue in classification of biological organisms. Both these schools aimed to design objective and biologically meaningful ways of classifying organisms or, alternatively, to invent a procedure for approximating or estimating the true natural relationships among organisms. (These two alternative formulations are functionally equivalent but the former assumes that classifications are inherently artificial, human creations while the latter presupposes that there exists a true natural system of relationships to discover. Both phrasings are met in the systematic literature.) In addition to sharing an ultimate goal, both schools of thought relied on numerical or computational methods (algorithms) to provide apparent objectivity and took full advantage of the electronic computer as it became available.

The primary difference between these schools actually predated their establishment; that is, the two schools arose independently rather than one in reaction to the other. Phenetics drew heavily on a French or Anglo-French tradition typified by the work of the French botanist Adanson (see Jardine and Sibson 1971:137 for a brief bibliographic history of the antecedents to phenetics) while cladistics came out of the German system of which Haeckel is the most prominant representative and essentially began with the work of Hennig (initially published in 1950; translated into English in 1966, at which time its influence began to be felt). Here, we will distinguish the Adansonian perspective, implying merely an absence of *a priori* assumptions about whether organisms form a tree, from phenetics, which espoused the Adansonian perspective, but really describes the program of numerical classification on Adansonian principles that was started by Sokal and Sneath (1963) and explicitly renounces *a priori* weighting of the characteristics used to classify organisms. It has frequently been said that phenetics classifies organisms based on their 'degree of overall similarity. We will also distinguish the Haeckelian view, which merely presupposes a single, strictly dendritic tree of life, from cladistics which is a program of classification in general agreeing with Haeckels *a priori* belief in a dendritic tree of life but primarily based on Hennigs reccomendations for estimating phylogeny based on synapomorphy and subsequent rules for applying names to the estimated phylogeny. A slight element of confusion is produced by the group of cladists called 'pattern (or 'transformed or 'methodological) cladists who have rejected the Haeckelian *a priori* belief in a dendritic structure but are willing to retain the methods that cladists use to reconstruct a tree of life and name its branches.

The main argument in support of an Adansonian approach is that it has fewer untestable assumptions (i.e. it does not assume a dendritic structure in the history of life) while this has also been urged as a criticism: that since most scientists believe in descent from a single origin of life, a method of classification that does not take this as a premise is less likely to be biologically meaningful. The same criteria have been used to evaluate the Haeckelian perspective and the cladistic methods derived from it: on one hand that they have more assumptions than strictly necessary in a classificatory system, and on the other hand, assuming a dendritic structure is reasonable insofar as we also assume that Darwinian evolution takes place.

Through the 1980s there was an active debate between the pheneticists and the cladists (REFS) with the sometimes inflamatory, sometimes palliative influence of a third school, the evolutionary taxonomists (also called Mayrian, Simpsonian, synthetic, syncretistic, gradistic and eclectic taxonomist) who believed that taxonomy had to be a compromise between an evolutionary system and one that considered overall similarity as well as descent.

The triangular controversy between cladists, pheneticists, and evolutionary taxonomists essentially ended with cladistics acquiring a definitive preeminence over the two competing schools of thought; since 1990, the evolutionary taxonomists have largely been branded as old-fashioned or reactionary while the pheneticists are usually dismissed as incorrect when their point of view is considered at all. Phenetics (the algorithmic classification of objects with no *a priori* weighting of the importance of their characteristics) was hardly ever used for the routine classification of biological organisms and now seems only to be studied in statistics departments where it is of interest per se as an application of clustering algorithms (Hartigan 1975).

There are a number of possible explanations for the success of cladistics: first of all its supporters may be correct in arguing that a true, strictly branching tree of life exists, that cladistic methods

of phylogenetic reconstruction offer the best available way of finding progressively better approximations to it, and that biological nomenclature should have a one-to-one correspondence with the best available approximation to the true phylogenetic tree. Second, the historically contingent association between molecular sequence data and cladistics may have contributed to the dominance of cladistics. In other words, the early molecular systematists may have had a cladistic or Haekelian perspective so when they began to construct cladograms from molecular data they continued to think of themselves as cladists even though maximum likelihood algorithms for phylogeny estimation from sequence data seem to occupy a fuzzy ground between phenetic and cladistic methodologies. Finally, the dominance of cladistics in systematic biology may be due to its better concordance with prior views independently of its real success as a tool for classification and taxonomy. A corollary to this possibility is that it is not inherently in concordance with prior dogma, but is merely easier to manipulate and therefore can be easily, perhaps unconsciously, forced to give acceptible views. Whatever the reason for the current dominance of cladistics, its agenda is stated very clearly: it will provide rules for the express purpose of naming the parts of the tree of lifeboth species and cladesby explicit reference to phylogeny. (Cantino and de Queiroz 2000:2)

From before Darwin to the present, the implicit assumption has underlain most practical taxonomy that, 'all organic beings are found to resemble each other in descending degrees, so that they can be classed in groups under groups. This classification is evidently not arbitrary like the groupings of the stars and constellations.' (Darwin 1859:411)

In biology, much of the debate over classification has focussed on the adjectives 'natural' and 'artificial': a 'natural' classification is one that is perceived to be objectively present among organisms, regardless of whether or not and by whom they are classified; an 'artificial' classification is imposed by the subjective, question-driven intent of a particular systematist.

But how does this natural-arbitrary distinction related to the heuristic-set-theoretic duality discussed above? In fact the categories are reversed: a natural group is determined hermeneutically and is therefore emic and in reality subjective while an artificial group is set-theoretically defined, and therefore etic and objective.

'There are two kinds of classification systems, one reflecting properties of the classifier alone (the subject), or artificial systems, the other mirroring relations having real existence in nature (the object), or natural systems. (Note that "objective reality" and "objective classification" do not refer to the opinions of classifiers who are sincere or consistent; "objective" and "subjective" have, again, both logical and psychological meanings.) (Ghiselin 1966a:212)

Again linguistics provide a useful example: a real language (like English) is defined by the phonemes—subjectively recognized characteristics—that compose it, while the International Phonetic Alphabet is an entirely artificial compilation of sounds that appear phonemically in different languages.

In set-theoretic terms, there is no difference in degree of arbitrariness between constellations of stars and species of mammal; it is just that our Copernican view of the stars makes it vividly apparent how alternative perspectives can make the constellations as seen from the earth seem merely a particular point of view, while we are less practiced in imagining alternative ways of classifying mammals. A heuristic classification of mammals sees 'tigers' as an important thing to name and 'striped animals' as unimportant, but in set-theoretic terms the two classes are equally abstract.

This leads us towards the conclusion that biological classification should be governed by the pragmatic concerns of data analysis, rather than by some notion of discovering 'natural groups'. A salutary example is provided by geology, a field that also must classify natural (empirical) data, but which has shown little angst about arbitrary classifications and no debate over classificatory methodology.

This can be illustrated by examination of a typical geological problem, the classification of sandstones, in which arbitrary delimitation is not seen as a defect. There are three alternative classifications of sandstones given in a standard textbook on sedimentology (Boggs, 1995). All three focus on the relative proportions of quartz, feldspar, and rock fragments that make up the grains of the sandstone, and all three restrict the category 'quartz arenite' to a very restricted area (though this area is differently shaped in each system). The lines separating the categories are clearly arbitrary in the sense that 5% is not markedly better than 6%, but no geologist would consider rejecting such a classification because of the way real sandstones were distributed on the triangle.

This is equivalent to there being three definitions of the genus *Homo*, one as 'the set of all featherless bipeds', another as 'the set of all organisms more closely genetically related to Joe Dimaggio than to any living chimpanzee', and a third as 'the set of all humans'. All of these refer to more-or-less the same group of empirical phenomena (us) and in many cases none has any advantage over the others. Few biologists, however, would be ambivalent about which of these to use.

Why are biologists so concerned that their classifications reflect 'real' boundaries in nature? This is not a universal preoccupation:

> 'It has often been said, particularly by those biologists who consider the species to represent a more natural level of integration than other taxa...that *phena* delimited by taxometric methods are arbitrary. This criticism becomes less serious if it be accepted that the concept of biological species is a model which is actually realized in nature only rarely.' (Silvestri and Hill 1964:96)

So, when species are looked at closely, they almost never look anything but arbitrary. In the way that statisticians conventionally assume that variables are independent even though they almost never are, biologists assume that species are real even though they are in practice abstract set-theoretic approximations of biological clusters or patterns that happen to be interesting.

Here be a Figure

The minority of biologists who support this practical view admit that:

> 'Classification...was aptly described...as a concise key to a great deal of information of a rather varied nature specifically pertaining to a group of associated organisms (a taxonomic group or taxon), the link between the taxonomic assembly and the stored information being formed by a code word,...the scientific name....If one agrees with this, the *practical* value of a classification is...manifest. Nevertheless, many workers are apparently reluctant to admit that it is exactly this useful aspect of classification which often decides the methods and the criteria to be employed in *arriving* at a classification....Practical reasons tend to overrule alternative considerations (such as a more "scientific" approach to classification, whatever that is).' Meeuse 1964:115f

And yet the dominant view remains focussed on 'natural' and primarily 'natural because phylogenetic' classifications:

> 'If a system is not phylogenetic, it really cannot be judged "better" or "worse" than another that is not. It is really pointless to argue about the relative merits of various artificial schemes.' (Delevoryas 1964:30)

In some ways, the current state of biological classification has been determined more by the economics and institutional structure of the academic world than by any logical or scientific concern. Support for the development of genetic technology (from the agricultural and pharmaceutical industries) has led to the extraordinary flourishing of departments of molecular biology with respect to the organismal and evolutionary branches of the field. Genetic data is new; genetic data is hot; running a laboratory that produces genetic data guarantees a certain level of funding in academia. Genetic data is not necessarily the way everyone should be classifying organisms and the methods developed for handling genetic sequences (as with many highly specific, technical methods) tend to obscure the problems with using them to reconstruct phylogeny, not to mention the problems with classifying things based exclusively on their genetic relatedness.

So a preoccupation with genetic data and the confused logic that measures the quality of a biological classification scheme by whether it reflects 'natural' groups (which are actually emic, subjective groups, as discussed above) has led to a dogmatic approach, which equates biological classification with measurement of genetic distance to the exclusion of real issues of practical classification of different types of empirical data.

Classification as applied data analysis

When discussing classification as a branch of applied data analysis, the question of whether the groups being produced are 'real' seldom appears, though some similar issues have been discussed: 'All clustering algorithms will, when presented with data, produce clusters—regardless of whether the data contain clusters or not....Data which do not contain clusters should not be processed by a clustering algorithm' (Jain et al. 1999:267). This concern has led to the the production of stopping rules for hierarchical algorithms and various coefficients that measure the degree of clustering found

in a data set. In addition, practical application of classificatory algorithms has its own issues, of which three in particular have received particular attention.

Distribution mixtures

In a univariate case, classification is simplified to the question of clustering: in the distribution of data along a single axis of variation uniform, or is the data arranged in clusters? This is the question of discriminating mixtures of different distributions, which has received a fair amount of attention in statistics.

If the forms of the mixed distributions are known with some degree of precision, very precise tools can be designed to discriminate them. A good example of this are the Rayleigh criterion and Sparrow limit, which are used by astronomers to decide whether to recognize as separate closely spaced point light sources at long distances.

Since the diffraction pattern (Fraunhofer pattern) due to a single slit has a well known functional form,

$$y = (i * ((sin(pi * a * (x/sqrt(x^2 + b^2))/l))/(pi * a * (x/sqrt(x^2 + b^2))/l))^2,$$

where a is the width of the slit, b is the distance of screen from slit, i is the maximum light intensity, and l is the wavelength of the light.

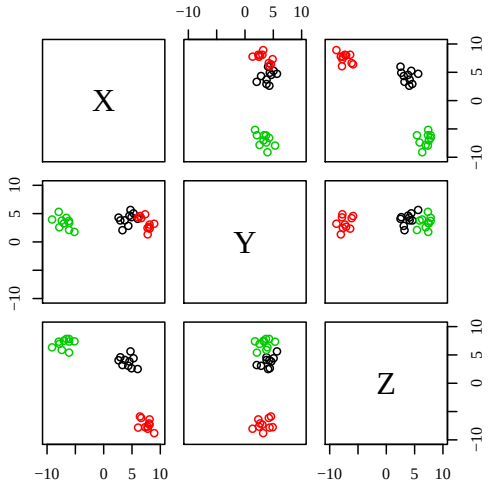
it is possible to propose a precise criterion for acceptance of polymodality; there are two that have been suggested, the first due to Rayleigh (REF), who argued that point sources were resolvable when the maximum of one was lined up with the first minimum of another. This is equivalent to a 26.5% dip in brightness between the sources in the two-dimensional case (point sources); the dip in the case of a one dimensional case (slit experiment) is slightly different. A less conservative version of this was proposed by Sparrow (1916), who proposed that any dip in intensity was enough to discriminate sources of known frequency (modulation transfer function = 0).

Similar criteria have been proposed for mixtures of Gaussian distributions as well as for other functional forms (Hartigan 1985), but their practical application is limited by the accuracy with which distributional forms are known. Even in so simple a case as measurement with a ruler, the total measurement error will in fact be the sums of the measurement errors associated with aligning the tick marks at each end. These are usually assumed to be equal, uncorrelated, and equivalent in sum to about half to one tick on the ruler, but in practice they are liable to be different (as, for instance, is the case when using a tape measure where the near end can be read carefully but the far end is out of reach and cannot be hooked over a convenient corner.

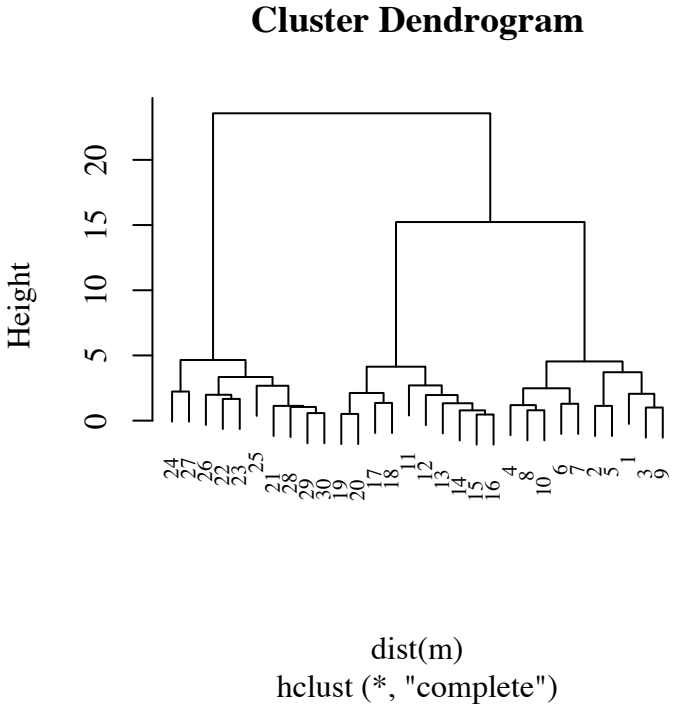
In fact one should also add in other errors like the error of reading ticks incorrectly or adding or dropping a whole number of units. The problem with dealing with this component of the error equasion is that such errors have no convenient functional form, so they are conventionally assumed to be zero....Note the though they are *conventionally* assumed to be zero, it is not true that they are *usually* assumed to be zero; misreading or typing error are usually the first things checked when unexpected results appear.

Orthogonality

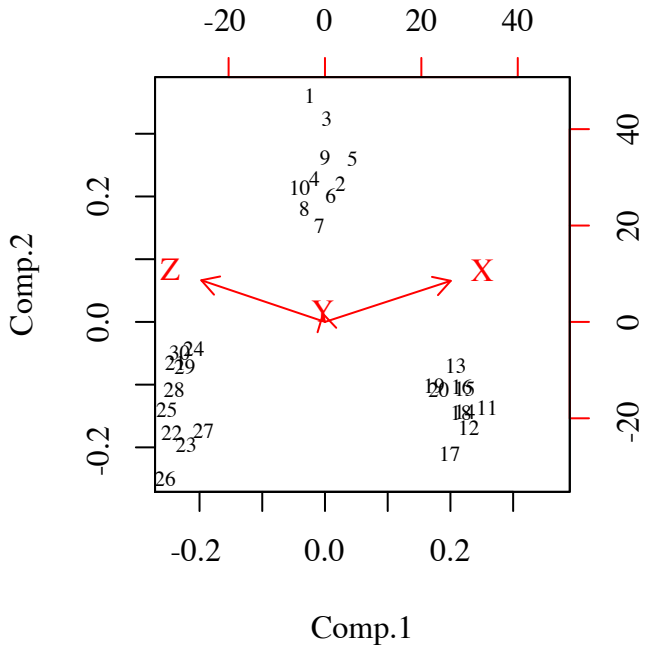
When multiple variables are considered, as in most actual classifications, there is the added issue of orthogonality: for each variable there it the question of whether the data is clustered with respect to it, and then there is the question of whether enough of the (relevant) variables show clusters, and if so whether the clusters that they show are the same.



For instance, here is a (complete linkage, agglomerative) hierarchical cluster plot and a pairs plot of a set of thirty points in three well-separated tri-variate-normal spherical clusters. The presence of these three clusters is obvious, but a careless examination of these plots would also seem to suggest that two of the clusters are more similar to each other than either is to the third cluster.

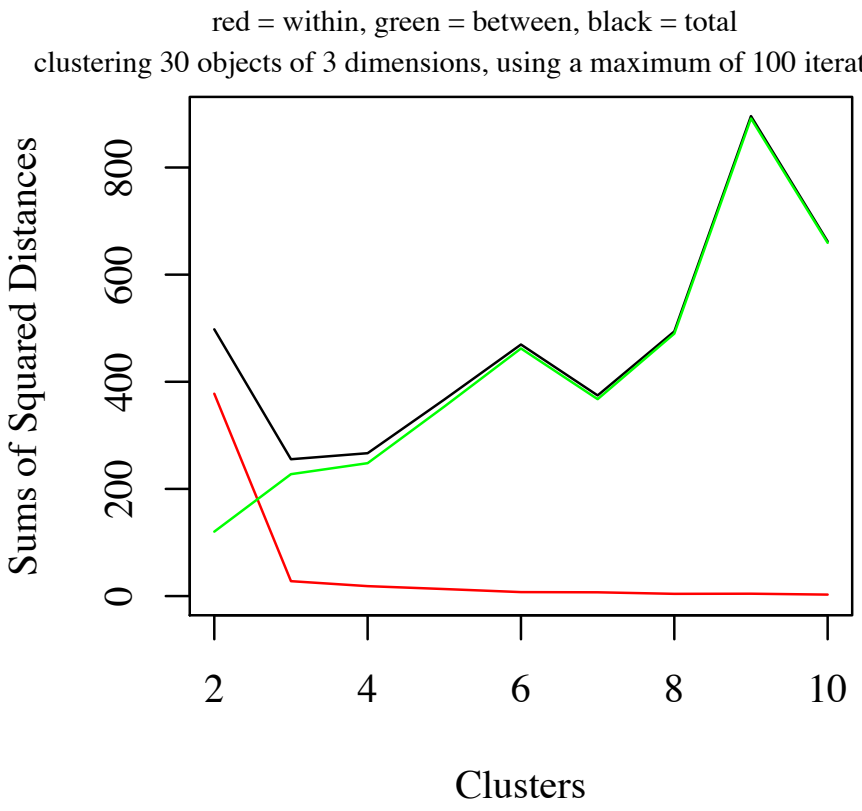


Even in this trivial case, it requires a rotation of the axes via eigenvector extraction to show that the clusters are in fact equally spaced in three dimensions:



The great variety of eigenvector methods remain essentially graphical devices (like pairs plots) for plotting more than two variables on a flat surface; there is no reason to expect that the linear combination of variables with the largest variance (the first principle component) is more likely than any other combination of variables to show clustering.

The closest thing to an algorithmic procedure that will actually evaluate the presence of clusters in many dimensions is a partitioning algorithm like k-means clustering or a hierarchical clustering algorithm with a stopping rule. Both can be used to illustrate clusters that are known to be present and they can provide some guidance about how many clusters are present in data known to be clustered (as illustrated by the figure below in which the presence of exactly three clusters is recovered from the synthetic data by minimizing the average total sum-of-squares distances between points), but neither indicates whether some variables are better than others for classificatory purposes.



The practice of variable shopping is common in all branches of applied statistics, and applies to clustering in much the same way that it applies in multiple regression. Because a variable can always be found to induce a particular clustering, the temptation is great to choose data that provide unambiguous answers instead of data that are conceptually related to the question of interest.

Scale

Finally, in addition to the issue of separating mixtures of similar distribution (discrimination) and reconciling orthogonal measurements, there is the practical issue of dealing with large amounts of data.

Classification of tens or even hundreds of points with two or three continuous variables is so trivial that there would never have arisen a need for algorithmic clustering methods. There are almost two million described biological species, however, and if the clustering of organisms into species is also considered a matter of interest, then the number of points to be classified by systematic biologists becomes in practical terms uncountable.

'In *Solanum* there are at least 1000 species and no one knows how to break them up

> into major groups. This is a case where the use of the computer would be of extreme value. I am quite certain that there must be many such groups in which we cannot produce any good system by our present methods simply because of the limitations of the mind in holding more than a certain number of characters at the same time.' Stearn in Heywood and McNeill, eds. 1964:162

These are the situations in which a discussion of the methodology of classification becomes particularly interesting. How robust are different algorithms in the face of different kinds of noise? Since complete sampling is generally impossible and there are seldom even vague *a priori* assumptions about the hyperdimensional distribution of the data, how can classifications be produced based on irregular and scarce sampling?

In biology, attempts to deal with these questions have generally taken second place to extremely technical and elaborate discussions of optimal methods of dealing with relatively small sets of particular types of data (like genetic sequence data or morphological character codes) with no examination of when or whether such methods are applicable to orders of magnitude more data. The efficacy of graphical tools has been grossly underrated (Chernoff 1973, Tufte 2001, Basford and Tukey 1999), and the quest for resolution of dendritic structure as far as possible has helped minimize the importance of error analyses. It is often more important to determine when classification is meaningless than to show how a procedure can be developed to classify every point exactly with respect to every other point however lacking in content such a procedure may be.

Conclusions

In *The Merchant of Venice* Bassanio and the Princes of Morocco and Arragon have the task of classifying three caskets of different materials (gold, silver, and lead) into two groups (the group that leads to marriage with Portia and the group that leads to rejection). To all three suitors are available the same apparent empirical data (the materials of and inscriptions on the caskets). The actual empirical contents of the caskets (two mocking doggerel verses and one portrait of Portia) are hidden. Had the contents of the caskets been known, naturally all three suitors would have made the same classification because their ultimate aims, marriage with Portia, are identical. Because of differences in the classifiers, however, the conclusions that are drawn from the same apparent empirical data are different. The situation of empirical scientists applying classification procedures to real objects and organisms is similar to the situation in which Portia's suitors are put: they agree on overall theoretical goals, but not on the way in which the available empirical data is likely to relate to the overall goal. Therefore their choices of classification are likely to reveal more about their own biases and preoccupations than they are about patterns in the data. Thus the areas in which classification has been a methodological issue are those fields in which there has been most doubt about the sort of empirical data that should be considered relevant for any given question. In the classification of sandstones, the variables of interest have remained clear and constant over the past century; in the classification of organisms, the purposes for classification have changed radically over the same time period.

Therefore the advantage provided by numerical or algorithmic methods over raw human prejudice is sometimes equivocal:

> 'The "numerical" approach may produce a useful classification where other methods have failed; but I honestly believe that on the whole its merits are highly overrated, its objectivity is based on feelings of false security, and its results are not "better" nor more reliable and certainly not preferable to those of conventional methods of classification based on a well-considered judgment.' Meeuse 1964:120

So some biologists have rejected a 'numerical' (algorithmic) approach because it seems to make false (or at least problematical) claims to objectivity rather than because it lacks useful practical functionality. Almost certainly in the case of dealing with very large data sets, probably in the case of synthesizing orthogonal variables, and maybe in the case of distinguishing mixtures, algorithmic methods have obvious practical advantages. The focus on metaphysical validity over practical utility has preoccupied biologists, narrowed the range of questions for which useful classifications can and should be designed to essentially a single axis of variation (genetic dissimilarity) and deflected attention from the practical differences between comparable algorithms.

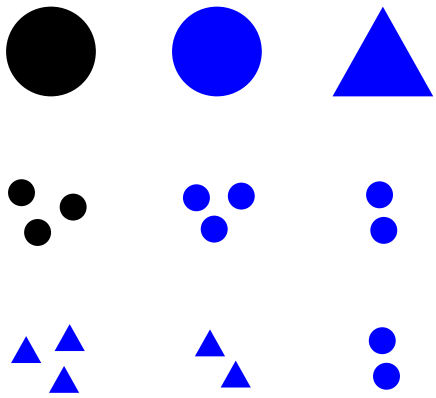
Decisions about classification techniques should begin with a choice about the purpose or purposes to which the classification will be put, and end with and evaluation of how well it serves those ends.

> 'When objective facts have been examined and we are still disagreeing, there has been a tendency to say...."the decision is then arbitrary" or "it is purely subjective"....In fact what has also been emerging...is that when we are in doubt about our classifications we want to ask what the purpose of our classification is.' Crawshay-Williams in Heywood and McNeill, eds. 1964:163

This can also be illustrated with the three-object problem: for three items A, B, and C, there are three possible (non-trivial) classifications:

A(BC), (AB)C, and (AC)B, where (ABC) and (A)(B)(C) are trivial

If for example A = 3 apples, B = 2 oranges and C = 3 oranges, or A is a blue circle, B is a blue triangle, and C is a white triangle, it is clear, as in the three-object problem discussed above, that there is no algorithmic way to organize the objects. Classification is clearly possible, but it explicitly requires a choice about whether to privilege colour, shape, number, or some other variable.



In the case of sandstones, geologists seem to have no difficulty identifying interesting axes of variation and then designing classifications—sometimes several different classifications—for given purposes. When faced with evolving biological organisms, biologists seem to have lost the pragmatics of classification in a morass of hazy epistemology and miss-applied set theory.

Works Cited

Basford, K. E. and J. W. Tukey (1999) *Graphical Analysis of Multiresponse Data* Boca Raton: Chapman and Hall/CRC.

Binder, D. A. (1978) Bayesian cluster analysis *Biometrika* 65(1):31–38.

Binder, D. A. (1981) Approximations to Bayesian clustering rules *Biometrika* 68(1):275–285.

Cantino, P. D. and K. de Queiroz (2000) PhyloCode: A Phylogenetic Code of Biological Nomenclature, draft version published electronically April 8, 2000.

Chernoff, H. (1973) The use of faces to represent points in K-dimensional space graphically *Journal of the American Statistical Association* 68(342):361–368.

Delevoryas, T. (1964) The role of palaeobotany in vascular plant classification pp. 29–36 in Heywood and McNeill, eds. 1964.

Everitt, B. (1974). Cluster Analysis. New York: Wiley.

Faith, D. P, Minchin, P. R. and Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance *Vegetatio* 69, 57-68.

Fielding, A. H. (1999) Machine Learning Methods for Ecological Applications Boston:Kluwer

Fisher, L. and J. W. Van Ness (1971) Admissible clustering procedures *Biometrika* 58(1):91–104.

Ghiselin, M. T. (1966a) On psychologism in the logic of taxonomic controversies *Systematic Zoology* 15(3):207–215.

Ghiselin, M. T. (1966b) An application of the theory of definitions to systematic principles *Systematic Zoology* 15(2):127–130.

Gordon, A. D. (1999). *Classification, Second edition* London:Chapman and Hall / CRC

Hartigan, J. A. (1975) Clustering Algorithms. New York: Wiley.

Hartigan, J. A. (1985) The dip test of unimodality *Annals of Statistics* 13(1):70–84.

Hartigan, J.A. and Wong, M.A. (1979). A K-means clustering algorithm. Applied Statistics 28, 100-108.

Heywood, V. H. and J. McNeill, eds. (1964) *Phenetic and Phylogenetic Classification* Systematics Association Pub. No. 6. London: The Systematics Association.

Jardine, N. and R. Sibson (1971) *Mathematical Taxonomy* London: John Wiley and Sons.

Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Legendre, L. and P. Legendre (1998[1983]) Numerical Ecology 2nd [1st] English edition

Mayr, E. (1982) *The Growth of Biological Thought* Cambridge: Harvard University Press.

Mayr, E. (1995) *Systems of Ordering Data* Biology and Philosophy 10:419434.

Meeuse, A. D. J. (1964) A critique of numerical taxonomy pp. 115–121 in Heywood and McNeill, eds. 1964.

Moore, G. H. (1982) *Zermelo's Axiom of Choice: Its Origins, Developments, and Influence*

Raffman, D. (1994) Vagueness without paradox *The Philosophical Review* 103(1):41–74.

Silvestri, L. G. and L. R. Hill (1964) Some problems of the taxonmetric approach pp. 87–103 in Heywood and McNeill, eds. 1964.

Sokal, R. R. and P. H. A. Sneath (1963) Principles of Numerical Taxonomy San Francisco: W. H. Freeman and Co.

Sneath, P. H. A. and R. R. Sokal (1973). Numerical Taxonomy. San Francisco: Freeman.

Struyf, A., M. Hubert and P. J. Rousseeuw (1996): Clustering in an Object-Oriented Environment. Journal of Statistical Software ¡URL: http://www.stat.ucla.edu/journals/jss/¿

Tufte, E. R. (2001) *The Visual Display of Quantitative Information* 2nd Ed. Cheshire, Connecticut: Graphics Press.

Van Ness, J. W. (1973) Admissible clustering procedures *Biometrika* 60(2):422–424.

Wolda, H. (1981). Similarity indices, sample size and diversity *Oecologia* 50:296–302.

Other sources:

Tukey T.A. Bancroft, ed., Statistical Papers in Honor of George W. Snedecor, (Ames, Iowa 1972), pp. 293-316, copyright 1972 by The Iowa State University Press EPIDEMIOLOGY & PUBLIC HEALTH LIBRARY Call Number: QA276.16 S83

van Heijenoort (ed.): From Frege to Gdel: A Source Book in Mathematical Logic, 1879-1931, Cambridge, Massachusetts, Harvard University Press 1967

Anderberg, M. R. (1973). Cluster Analysis for Applications. Academic Press: New York.

Krebs, C. J. (1999). Ecological Methodology. Addison Wesley Longman.

Mountford, M. D. (1962). An index of similarity and its application to classification problems. In: P.W.Murphy (ed.), Progress in Soil Zoology, 43-50. Butterworths.

Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, edited by H. -H. Bock and E. Diday

Mathematical Classification and Clustering, by Boris Mirkin, 1996 [new link]

Clustering and Classification, Edited by Phipps Arabie, Lawrence J. Hubert, and Geert De Soete, 1996

Murtagh, F. (1985). "Multidimensional Clustering Algorithms", in COMPSTAT Lectures 4. Wuerzburg: Physica-Verlag (for algorithmic details of algorithms used).

Appendix

Numerical Tools

Summary of methods hermeneutic unsupervised algorithmic classification (cluster analysis sensu strictu) hierarchical cluster analysis divisive agglomerative using single linkage distance metric (nearest neighbor) using average linkage distance metric using complete linkage distance metric (farthest neighbor) partitioning hierarchical plus stopping rule k-means liable to be trapped on local min. of squared distances topological minimum spanning tree (MST) closely related to hierarchical methods Delaunay graph relative neighborhood graph (RNG) fuzzy clustering fuzzy c-means (FCM) monothetic clustering supervised algorithmic classification (computer learning) descriminant analysis neural nets (parallel distributed processing, connectionist techniques) require quant. data, number of ouput nodes limited (fixed number of output clusters) self organizing map (SOM) sensitive to initial weights, balls only learning vector quantization (LVQ) Kohonen 1984 similar to k-means adaptive resonance theory (ART) (Carpenter and Grossberg 1990) order dependent, balls only feed forward feed backward rule based (expert systems) searches evolutionary algorithms sensitive to control parameters genetic algorithms (GA) evolutionary strategies (ES) evolutionary programming (EP) simulated annealing (SA) branch and bound tabu search glover 1986 al-sultan 1995

Hierarchical methods generally more versatile but more memory/computation-intensive than partitioning. Jain 1999:277

Distance metrics single euclidean root sum of squared distances manhattan sum of absolute distances minkowsky generalized pth root of sum of distances to the power p canberra sum(—x_i - y_i— / —x_i + y_i—) Terms with zero numerator and denominator are omitted from the sum and treated as if the values were missing. canberra d[jk] = (1/NZ) sum ((x[ij]-x[ik])/(x[ij]+x[ik])) where NZ is the number of non-zero entries. symmetric binary Mahalonobis $d(xi, xj) = (xi - xj)S^{-}1(xi - xj)^{T}$, where S is the covariance matrix of the x's assumes unimodality; multidimensional normality Hausdorff asymmetric binary The vectors are regarded as binary bits, so non-zero elements are 'on' and zero elements are 'off'. The distance is the proportion of bits in which only one is on amongst those in which at least one is on.

gower d[jk] = sum (abs(x[ij]-x[ik])/(max(x[i])-min(x[i])) bray d[jk] = (sum abs(x[ij]-x[ik])/(sum (x[ij]+x[ik])) kulczynski d[jk] 1 - 0.5*((sum min(x[ij],x[ik])/(sum x[ij]) + (sum min(x[ij],x[ik])/(sum x[ik])) morisita d[jk] = 2*sum(x[ij]*x[ik])/((lambda[j]+lambda[k]) * sum(x[ij])*sum(x[ik])) where lambda[j] = sum(x[ij]*(x[ij]-1))/sum(x[ij])*sum(x[ij] -1) Morisita index can be used with genuine count data only. horn Like 'morisita', but $lambda[j] = sum(x[ij]^2)/(sum(x[ij])^2)$ Horn-Morisita variant is able to handle any abundance data. Jaccard index is computed as 2B/(1+B), where B is Bray-Curtis dissimilarity. Mountford index is defined as M = 1/alpha where alpha is the parameter of Fisher's logseries assuming that the compared communities are samples from the same community (cf. 'fisherfit', 'fisher.alpha'). The index M is found as the positive root of equation exp(a*M) + exp(b*M) = 1 + exp((a+b-j)*M), where j is the number of species occurring in both communities, and a and b are the number of species in each separate community (so the index uses presence-absence information). Mountford index is usually misrepresented in the literature: indeed Mountford (1962) suggested an approximation to be used as starting value in iterations, but the proper index is defined as the root of the equation above. The function 'vegdist' solves M with the Newton method. Please note that if either a or b are equal to j, one of the communities could be a subset of other, and the dissimilarity is 0 meaning that non-identical objects may be regarded as similar and the index is non-metric. The Mountford index is in the range 0 ... log(2), but the dissimilarities are divided by log(2) so that the results will be in the conventional range 0 ... 1.

mutual neighbor distance (MND) d(xi, xj) = NN(xi, xj) + NN(xj, xi) where NN (a, b) is the rank of b in the ordered list of neighbors of a Gawda and Krishna (1977) in Jain et all 1999:273 NON METRIC (does not satisfy triangle inequality) conceptual clustering d(xi, xj) = f(xi, xj, C, K) Michalski and Stepp 1983 in Jain 1999 273 NON METRIC (does not satisfy triangle inequality)

group complete linkage (farthest neighbor) Maximum distance between two components of x and y (supremum norm) compact clusters average linkage single linkage (nearest neighbor) suffers from chaining; can extract concentric clusters minimum variance =? ward's method, Ward(1963) J. Am. Stat. Assoc. 58:236 ward's method mcquitty median centroid minimal spanning tree

Similarity measures Bray-Curtis

Synthetic data binary factorial ordered unordered hierarchical continuous

Graphical Tools graphical techniques direct matrix operations Braun-Blanquet profile analysis

Basford and Tukey (1999) Hartigan(1967) scatter plot lattice matrices glyph analysis Chernoff faces trees and castles heatmaps eigenvector methods principal components analysis (PCA) Factor analysis Gradient analysis (direct, indirect) Multidimentsional scaling (MDS) (Detrended) Correspondence analyisis (DCA) Canonical Correspondence analysis (CCA)