1-1-2014

Identification of unknowns within a probabilistic system: The diagnostic value of attributes

D W. Goodall
Edith Cowan University

10.1080/11263504.2014.913731
This is an Accepted Manuscript of an article published by Taylor & Francis in Plant Biosystems on 12 Jun 2014:
Goodall D.W. (2014). Identification of unknowns within a probabilistic system: The diagnostic value of attributes. Plant Biosystems, 148(6), 1346-1354. Available here

Identification of unknowns, within a probabilistic system: the diagnostic value of attributes

David W. Goodall*

School of Natural Sciences, Edith Cowan University,

Joondalup 6027, Western Australia

* E-mail: d.goodall@ecu.edu.au

Abstract

Using a data base underpinned by probability considerations, in which a variety of attributes, some of which may be quantitative, are recorded for a number of "operational taxonomic units" (OTU's), a key system is described by which an unnamed specimen may quickly be identified.

The concept of "diagnostic power" is introduced, by which each attribute is evaluated in terms of its potential contribution to identifying the unnamed specimen.

Besides coverage of different types of attribute and the introduction of "diagnostic power", the system has the advantages of incorporating multiple values of an attribute for each OTU, and offering short-cuts to identification

Key Words: identification; diagnosis; probability; classification; taxonomy; *Drosera*

1. Introduction

Diagnostic keys have been part of the biological world for centuries. They have made it possible to allot a specimen – let us call it the propositus – to one of a number of recognized categories, usually taxonomic groupings. Traditionally, the diagnostic key has been a printed key in a book – often dichotomous. The enquirer is offered alternatives for a particular attribute. If he can decide which alternative for that attribute applies to the propositus, the key then directs him to another attribute, about which he again has to make a decision. Finally there are no choices – the set of attribute values observed for the propositus are found in only one of the taxa included in the key, and the propositus is identified as belonging to that taxon.

A great difficulty in the use of such keys has been that an attribute which needs to be defined may not be observable in the material in hand -- perhaps the propositus is incomplete, perhaps it is at the wrong stage of development, perhaps the enquirer lacks necessary tools or skills. It would be an enormous advantage if the enquirer could choose the attributes about which he is to make a decision, rather than having them chosen for him.

The new possibilities opened for diagnostic keys by the development of the computer were early recognized (Goodall, 1968, Pankhurst 1970, 1975), but remained little more than possibilities until computers of smaller size became available. Initially, as might be expected, computer applications simply made possible the rapid development of keys of the traditional type. However, it soon became evident that the possibilities were much wider. The advent of the personal computer led to the development of electronic keys. A number of computer programs, many listed by Norton (2002) and Dallwitz (2009) , have been developed for this purpose. General questions about the principles underlying identification keys were early discussed by Payne and Preece (1980), and have since been considered by Pankhurst (1991) and by Hagedorn et al. (2010). Much attention has recently been devoted to simple keys for use on portable devices in the field, often to cover rather limited ranges of taxa, and for use by non-specialists (e.g. KeyToNature (2011); see also Nimis et al (2012))

In computer-oriented keys, rather than the enquirer being guided through a series of decisions in an order prescribed by the designer of the key (who in the past was usually following the systematics of plants), he could have a virtually unlimited choice of search through the information available in the data base. But this was almost an *embarras des richessses*. One might raise the question: is unguided

choice of attributes really in the best interests of fast and accurate identification? Clearly attributes are not of equal value for this purpose. Sometimes a single attribute suffices without further ado to identify an unknown specimen as belonging to a single category and that category only. For instance, among Australian species of *Drosera* (Goodall and Marchant 1996), one species only (*D. burmannii)* has styles described as "penicillate" (brushlike). A penicillate style thus serves instantly to identify an unknown as *D.burmannii.* Another example: most *Drosera* species have three or five styles, but there are just two species in Australia with only a single style (*D. fimbriata* and *D.hamiltonii*). These style characters clearly have high diagnostic power within *Drosera*. Should an identification program for Australian species of *Drosera,* then, direct an enquirer to style chararcters, as likely to lead most directly to identification?

Some identification programs confront this problem of guiding the user's choice of attributes, selecting for him what are often called the 'best' attributes; but the principles of selection are rarely addressed. Gower and Payne (1975) considered the question at some length – but only for binary attributes (albeit taking account of missing values). Sneath (1980) also discussed the problem in the limited case of presence-absence variables. According to Hagedorn et al. (2010) "the fastest algorithms are those that provide a division into equally sized partitions" This is true if all taxa included in the key (or still available as options) are equally likely to be presented for identification, but this will rarely or never be true in practice.

Dallwitz et al. (2008) say that a program, in selecting the 'best' attribute, might take into account (1) what they call the '**cost**' of the different attributes – the effort required to determine them, and the risk of error; (2) the **frequency** with which different OTUs are presented for identification; and (3) the **distinctiveness** of the character values observed. '**Cost**' seems to be highly dependent on the particular circumstances of the identification – including the user's expertise –so that it is probably better to leave it to the user to take account of it himself. The **frequencies** in question are in principle hardly determinable; in programs as constructed, the frequencies (prior probabilities) are usually assumed to be equal. The **distinctiveness** of the character, however, is inherent in the data and the way they are handled by the program, so should very properly be presented to the user in the course of the identification, each time he selects an attribute for definition. INTKEY (Dallwitz 2009, Dallwitz et al. 2008) presents attributes to the user in decreasing order of a function reflecting the evenness of numbers of OTUs with the different values of the attribute in question. This may be a good rule of thumb where all attributes are qualitative and defined for all OTUs, and where a single value of each attribute characterizes each OTU . The general case, where any taxon may have more than one value for an attribute, where not all attributes are qualitative, and

where an attribute may be undefined for many taxa, is not satisfactorily addressed under this principle. Grosser et al. (2010) adopted a different approach to selecting attributes for attention, using the whole data set, taxa and attributes, for the purpose. This question of the diagnostic value of attributes is considered afresh in Section 5 below.

2. A proposal based on probability

Over the years, a probabilistic system has been described for numerical classification, using a particular type of data base (Goodall 1966, 1993, Goodall et Al. 1991, Goodall and Marchant 1996). The system handles all types of attributes, whether purely qualitative, ordered, quantitative (expressed as a real number), spatial (expressed by coordinates e.g. Goodall 1994) or angular (ordered on a circular scale e.g. Goodall 1993), with multiple values possible for each OTU[1] x attribute combination. Combination of probabilities for the different attributes depends on the assumption that the attributes are logically independent of one another. The fact that some OTU x attribute combinations may be unknown ("missing values"), or even logically indeterminate, presents no difficulty; those combinations are ignored. Since the relationship (similarity /dissimilarity) between two OTU's can be expressed in the same terms of probability (within the data base) for attributes of all types, the values can easily be combined.

Though designed for purposes of classification (similarity, clustering), data bases built on these principles can also be used effectively for identification. The present paper describes a computer key for this purpose.

3. The probabilistic data base

In these data bases, each attribute has a number of alternative values, and for each OTU x attribute combination, the proportion of the OTU with each alternative value for the attribute is specified in the data base. For each OTU x attribute combination, these proportions must, of course, sum to unity. In the case of quantitative attributes, the different alternatives are ranges of values, each defined by a median. Similarly, for spatial attributes, each alternative is defined by the coordinates of its centroid. These data bases have much in common with those of the well-known DELTA system, developed from 1970 onwards (Dallwitz 1974, 2009; Pankhurst (1991).

---

[1] OTU: "Operational Taxonomic Unit" – the items classified under this system.

An example is taken from the data-base for Australian species of *Drosera* [Goodall and Marchant 1996, Dallwitz 1974), to be used later (Table 1).

Table 1. Example of probabilistic data base entries. Selected attributes for *Drosera arcturi.* The proportion is n estimate of probability.

Attribute	Type	Alternatives				
Leaves in rosette?	Qualitative	Yes	No			
	Proportion	1.0	0.0			
*Rosette shape	Ordered	Hemispherical	Convex	Slightly convex	Flat or concave	
	Proportion	0.0	0.8	0.2	0.0	
*Radical leaves, max.dimension	Quantitative	Range 1: 3 – 5 cm	Range 2: 5 – 7 cm	Range 3: 7 – 20 cm		
	Median (cm)	4.0	6.0	13.0		
	Proportion	0.4	0.4	0.2		
Flowering date	Circular	Jan	Feb.-Sept.	Oct.	Nov.	Dec.
	Proportion	0.2	0.0	0.2	0.3	0.3
Geographical distribution	Spatial	Name	N.Z.	Tasmania	Victoria	N.S.W	Elsewhere
		Centroid N-S	42.0	42.5	37.0	32.0	--
		E-W	172.0	146.0	146.0	152.0	--
		Proportion	.25	.25	.25	.25	0.0

- * These attributes are indeterminate if leaves are not in a rosette

The example in Table1 is the form of data base used by the diagnostic keys described below; the proportion of OTU i with value or range k for attribute j is symbolized as $P_{ijk;}$; if the attribute is quantitative, the median of that range is A_{ijk}.

These data bases do not need to be complete -- some attributes for some OTU's may remain undefined without causing difficulties; and their structure has the advantage of permitting multiple states for each OTU x attribute combination. This is indeed highly appropriate in biological taxonomy, where multiple states for OTU x attribute combinations are commonplace. For instance, Marchant and George (1982), in their description of *Drosera pulchella,* include

> "Calyx 1.5-2.5 mm long, divided into obovate lobes, glandular-pubescent, the apices of the lobes entire or slightly fimbriate"

In the data base mentioned above, the translation of this description includes two attributes with multiple states (length, and lobe apex).

4. Diagnostic key using a probabilistic data base

Some features of the probabilistic data base described above make it particularly suitable for an identification key. That certain attributes may be indeterminate or unknown for some OTUs is a commonplace of taxonomic knowledge, as also is the possibility of alternative values of an attribute within the same OTU. These are handled without difficulty by the probabilistic approach.

The key user with a specimen for identification (the propositus) is offered a free choice among the attributes recorded. He chooses one, and describes this attribute for the propositus. Often the description will be multiple – the atttribute in question may have more than one value for the propositus − in which case the user is asked to estimate their proportions. For quantitative attributes, he may describe the propositus in one of three ways:

1. He may specify two or more ranges of values, each with a median value and a proportion (as for the OTU's included in the data set);
2. A maximum and minimum may be stated; or
3. A single value may be specified.

If option 1 or 3 has been chosen, limits are estimated as described in the Appendix. The program then compares the propositus description with the records for that attribute in each of the OTUs for which it has been recorded, and finds, for each, the proportion of overlap. "Overlap" is here defined as the proportion of the OTU in question, as described in the data set, which is compatible with the description of the propositus. An OTU should not be discarded as a possible identification unless there is no overlap -- the values reported for the propositus lie completely outside the range reported for that OTU. All OTUs with positive overlaps remain possible diagnoses. They are listed, each with its overlap proportion. The enquirer then has an opportunity to describe another attribute for the propositus.

Whenever a new attribute is described for the propositus, the overlaps for the new attribute are calculated, and multiplied by those previously calculated for other attributes to give the overall overlap of each OTU with the propositus as described to date. A number of OTUs previously possible as identifications now have zero overlaps and are excluded. The field is narrowed. The process of attribute selection, propositus description, and calculation of overlap is repeated until only a single OTU remains with a positive overlap, and hence as a possible identification.

The propositus is within the range of variation of this OTU, and outside the ranges of all others.

The calculation of "overlap" is clearly critical to the process of identification. For any non-quantitative attribute, where P_{ijk} is the proportion of OTU i with the k'th value of attribute j, and subscript *x* is used in place of *i* to denote the propositus, the "overlap" of OTU *i* with the propositus is defined simply as:

ΣP_{ijk} for all *k* where $P_{xjk} > 0$

In quantitative attributes, "overlap " is based on the extreme limits for the propositus and the limits of each range for each OTU in the data base with which it is to be compared. The limits required are estimated from the medians in the data base as described in the Appendix. If one had a continuous distribution curve of values for each OTU, the solution would be straightforward – the overlap would simply be the area of the distribution curve within the extremes for the propositus. However, the data base provides only the proportions within ranges, the medians,and the limits as derived from the medians. The limits for the *k*'th range of attribute *j* in OTU *i* will be symbolized as L_{ijk} and $L_{ij(k+1)}$, the extremes for the propositus (*x*), with n ranges, being L_{xj1} and $L_{xj(n+1)}$. The overlap between the *k* 'th range of OTU *i* and all *n* ranges of the propositus, as a proportion of the entire range of OTU *i*, is then estimated as

$$P_{IJk}\{\max(0,(\min(L_{ij(k+1)},L_{xj(n+1)})- \max(L_{ijk} ,L_{xj1})))/(L_{ij(k+1)} -L_{ijk})\}$$

and these quantities are summed over all values of *k* to give the total overlap of OTU *i* with the propositus for attribute j.

5. Diagnostic Power of Attributes

At any point in the process of identification, the choice of a new attribute to specify for the propositus may be rather critical. An optimal choice may lead one directly to the correct answer, whereas other choices, while not actually leading one astray, may be far less fruitful, and take one to the correct answer only after many time-consuming steps. The program can help and guide this choice, without determining it. To meet this need, the user is given the opportunity, at any stage, to enquire after the diagnostic power of attributes not yet specified for the propositus, among all the OTUs still remaining as possible solutions.

The diagnostic power of an attribute is defined as the probability that specification of a value for that attribute for the propositus will distinguish between a random pair of OTUs within the system; this is close to the "separation coefficient", as defined by Pankhurst (1991). It answers the question: what proportion of values of the attribute are different in a random pair of individuals?. All possible pairs of OTUs must be considered, and within any OTU pair all possible pairs of values. Only for pairs which differ can the attribute in question have any diagnostic value. It may be noted that, where the attribute is indeterminate or has not been recorded for one or both of the OTUs, the contribution to diagnostic power is explicitly zero

For OTUs *a* and *b,* and non-quantitative attribute *j,* the non-overlap of *a* with respect to *b* is a contribution to diagnostic power , and may be expressed as

$$D_{abj} = \Sigma_k P_{ajk} \text{ where } P_{bjk} = 0$$

The non-overlap of *b* with respect to *a* is

$$D_{baj} = \Sigma_k P_{bjk} \text{ where } P_{ajk} = 0$$

It will be noted that the two are different (except where both OTU's have records for the same subset of values), and so must be calculated separately.

In the case of quantitative attributes, since the OTU description includes only the proportions in specified ranges, disagreement between individuals is indicated by the extent to which the ranges in which they respectively fall do not overlap.

In determining their non-overlap, one has recourse to the limits, estimated as described in the Appendix. The non-overlap of *a* with respect to *b* includes all ranges of *a* for which the lower limit exceeds the uppermost limit of *b*, or the upper limit is less than the lowermost limit of *b,* together with a proportion of any shared ranges , thus:

$$D_{abj}=\Sigma_k P_{ajk}[\{\max(0,(\min(L_{aj(k+1),}L_{bj1})-L_{aj1}))+$$

$$\max(0,(L_{aj(k+1)}-\max(L_{ajk,}L_{bj(u+1)}))\}/(L_{aj(k+1)}-L_{ajk})]$$

where *u* is the number of ranges of attribute j for OTU *a*

Similar calculations, *mutatis mutandis*, give the non-overlap of *b* with respect to *a:*

$$D_{baj}=\Sigma_k P_{bjk}[\{\max(0,(\min(L_{bj(k+1)},L_{aj1})-L_{bj1}))+$$

$$\max(0,(L_{bj(k+1)}-\max(L_{bjk},L_{aj(v+1)}))\}/(L_{bj(k+1)}-L_{bjk})]$$

where *v* is the number of ranges of attribute *j* for OTU b.

Though the range of values shared {($L_{aj(p+1)} - L_{bjq}$) or ($L_{bj(q+1)} - L_{ajp}$), whichever is positive} is common to the two OTU's, the contribution to power (based on the non-overlap) may be quite different. For instance, the extreme range for one may lie entirely within that for the other. The two contributions must each be calculated and added to the overall power assessment for the attribute in question.

Whether the attribute j be quantitative or not, the values of D_{abj} are then averaged over all pairings of the *n* OTU's which remain relevant (including those where values of this attribute are unknown or indeterminate for one or both OTU's, in which case $D_{abj} = D_{baj} = 0$) to give the overall diagnostic value of attribute j:

$$V_j = \Sigma_a \Sigma_{b\neq a} D_{abj} /\{n(n-1)\},$$

n being the total number of OTUs still possible as diagnoses.

It may be noted that, since the diagnostic power depends on the set of OTU's remaining, it needs to be recalculated whenever this set is diminished as a result of exclusions following the description of an attribute for the propositus.

Though the procedure for calculation differs between quantitative and non-quantitative attributes, the results are fully comparable, expressing in each case the proportion of non-overlap.

The discussion of diagnostic power so far has the underlying assumption that all OTUs in the system are equally likely as identifications of the propositus. Often this assumption may be patently false (cf. Dallwitz et Al. 2008) and the key user may have specific information to the contrary. Most obviously, if the propositus is collected from a specified geographical area, there may be prior knowledge of the abundance and breadth of distribution of the various taxa. If this knowledge is incorporated into a table of prior probabilities, it can be used by the program in calculating attribute power. Then, in the equation for V_j above, the contribution for each pair of OTU's is weighted by the product of the prior probabilities of the two OTU's in question

In the practical application of the programs, if the user has information on prior probabilities, he is invited to specify them for the more probable OTUs (summing, of course, to less than unity), and the balance of probabilities is then divided equally among all remaining OTU's.

6. Exceptional Features as an Aid in Identification

As noted by Hagedorn et al. (2010), certain exceptional features within a group of organisms can be very helpful for identification, and may even make it possible to bypass the key. Only a few of the OTUs within the group possess them. The unsophisticated enquirer will not know the value of looking for these unusual characters, but the specialist's eye leaps to them immediately. "Penicillate styles" were mentioned above as such a character for *Drosera* spp. It seemed worth while, within the key program, to list such exceptional features as soon as work on a propositus begins.

The program gives the user the opportunity to define "exceptional" for this purpose as he chooses. In the present exposition, "exceptional" is defined as "shared by no more than 3% of the OTUs for which the attribute was recorded". For qualitative or spatial attributes, this test is applied to each value; for ordered or circular attributes, it is applied only to the values limiting a sequence in either direction; for quantitative attributes, it is applied to the medians of the extreme ranges.

7. An example of use

To illustrate the use of the program, I take a set of data on Australian species of *Drosera* used by Goodall & Marchant (1996), and since extended to cover a total of 89 species and subspecies, and 144 attributes, including fifteen calyx characteristics. The user (let us call her "Estelle") is attempting to identify a specimen purely on the calyx characters. She is first asked whether she has information on the prior probability of the various taxa; she has not. She then indicates that she would like to be told of exceptional features that might facilitate diagnosis, and defines "exceptional" as "shown by no more than 3% of the taxa included". She is then told that there are 51 "exceptional" characters, of which nine are calyx attributes, namely:

Attribute	Value
Sepal shape – position of maximum width	Near apex
Sepal shape – length/width ratio	> 8: "linear"
Sepal apex shape	Acuminate
Sepal apex fringe	Shallowly fringed
Sepal length	< 1.5 mm or > 8.0 mm
Sepal iridescence	Iridescent
Sepal concavity	Deeply concave
Gland distribution on sepals	At base
Calyx enlarging in fruit?	Enlarging

Estelle does not recognize any of these descriptions as applying to her specimen, and chooses to report that the sepal length is 2.0 mm. The program responds that 33 taxa remain as possible identifications, the proportion compatible with the propositus ranging from 1.0 (for 7 taxa) down to 0.002. Estelle then indicates that the calyx lobes in her specimen are free to the base. This reduces the number of taxa still consistent with the description to 28, with overlapping proportions ranging from 0.002 to 1.00. She then asks about the diagnostic value of the remaining attributes. Those for calyx attributes range from 0.071 for all 28 of the OTU's remaining to 0.731among 16 of them; the most promising seems to be the apex fringe of the sepals, with a diagnostic value of 0.643, for which all 28 remaining OTU's have information. Estelle reports that the apex is "shallowly fringed", and is informed that the possible identifications are now reduced to two: *D. macrantha* ssp*.macrantha,* and *D. pulchella.* She again enquires about the value of different calyx attributes in distinguishing between these two taxa, and is told that each of the following three attributes has a diagnostic power of 1.0 at this point:

Position of maximum width,
Pilosity, and
Glandularity

She chooses the last attribute, and reports that glands are present on the sepals; this defines the propositus as *D. pulchella,* and completes the identfication. She is offered a complete description of the taxon from the data base.

It may be noted that Estelle had been told at the outset that a shallow fringe at the sepal apex was an "exceptional" character; if she had recognized immediately that this applied to her specimen, she could have shortened the identification process, and gone directly to the final two taxa as possibilities.

Estelle has now another specimen to identify, and again wishes to use calyx characters. Not recognising that any of the "rare" attributes applies to her new specimen, she chooses to report that the sepal length is 7.0 mm. The program responds that 29 taxa in the data set include such values, the proportion compatible with the propositus ranging from 1.0 (for 8 taxa) down to 0.003. Estelle indicates further that the calyx lobes in her specimen are free to the base. This reduces the number of taxa still consistent with the description to 21, with overlapping proportions ranging from 0.033 to 1.000. She then asks about the diagnostic value of the remaining attributes. Those for calyx attributes range from 0.095 to 0.800, the most promising seeming to be (a) the sepal margins, with a diagnostic value of 0.668 , and on which the system has information for 10 OTUs still among the

possibilities, and (b) the distribution of glands, with a diagnostic value of 0.800, among 5 OTUs. Estelle chooses the latter attribute, but is reminded that this is dependent on another attribute, not yet defined for the propositus -- whether the calyx is free of glands or not. She confirms that, in her material, all the calyces are glandular, and then states that, of the four possibilities offered (glands throughout, sparsely glandular, glands at base or glands near edges), the second describes her specimen best. She is then informed that a single taxon in the data base matches this description, namely *D. menziesii* ssp. *thysanosepala.* So the identification is complete. Again, she is offered a full description of this taxon from the data base.

It did not apply in Estelle's case, but a user may make a mistake, leading to a description of the propositus which is incompatible with any OTU in the data base. If that happens, he or she is so informed, and invited to begin again.

8. Concluding Remarks

In conclusion, one may indicate certain advantages distinguishing this system from more traditional keys. One of these is the possibility of including multiple values, both for the OTUs in the data set and for the propositus. If the user has several differing specimens, or is merely uncertain as to which description best fits his specimen, he can incorporate the variation or uncertainty in his description of the propositus. A second advantage is the combination of quite different types of attribute (qualitative, quantitative, etc.) within the same system. Thirdly, attention is drawn to possible "short-cuts" – unusual and distinctive characters. The fourth special feature is the concept of "diagnostic power" of an attribute in distinguishing among a particular set of OTUs, and the opportunity to calculate it at any stage.

It hardly needs pointing out that the system is not limited to biological identification, but that it could be used in any situation where a propositus is to be allotted to one of a number of OTUs defined by multiple variables – such as arises constantly, for instance, in medical diagnosis.

The computer programs used in this system (INPUT for preparing a data base, and IDENTIFY for identification) are available, without charge, on request to the author.

Acknowledgements

I would like to thank my colleagues Dr.A.Koenders and Dr.K.Lemson for valuable criticism of this paper in draft, and Dr. P.L. Nimis and Dr. S. Martellos for reading and

commenting on the last version of the paper. Finally, I wish to express my appreciation to my friend Enrico Feoli, of Trieste, who undertook to see this paper through the publication process for me.

Appendix

Estimation of Limits for Ranges of Values in Quantitative Attributes

In this system, values of quantitative attributes are described in terms of the proportions (P_{ijk}) within ranges defined by medians (A_{ijk}). The diagnostic process, however, depends on the overlap between values for the propositus and OTUs in the data set, which are not derivable directly from the range medians. Accordingly, a procedure has been developed to estimate the limits of the ranges, so that overlaps can be determined.

In defining limits, a distinction is made between attributes in which random variation tends to be uniform as between different ranges of values (called "arithmetic" here), and those in which it tends to proportionality (here called "logarithmic"). In attributes for which zero values have been recorded, it is assumed that variation is "arithmetic". If there are no zero values, a decision between “arithmetic” and “logarithmic” is based on consideration of the extreme range of values for the attribute in the entire data set. If the maximum is more than twice the minimum, it is assumed that variation in this attribute should be treated as "logarithmic", otherwise "arithmetic"; i.e. it depends on the value of

$$E_i = \max_{ik} A_{ijk} / \min_{ik} A_{ijk}$$

$E_J > 2$ causes j to be treated as “logarithmic”, $E_j \leq 2$ as “arithmetic”.

The estimation of limits for the ranges around the medians A_{ijk} then proceeds as follows:

If two or more ranges of attribute j are defined for an OTU i, then the limits for range k (out of n) $(1 < k < n)$ are defined as

$$L_{ijk} = (A_{ij(k-1)} + A_{ijk})/2 \qquad \text{("arithmetic")}$$
$$L_{ij(k+1)} = (A_{ijk} + A_{ij(k+1)})/2$$

or

$$L_{ijk} = \sqrt{(A_{ij(k-1)} * A_{ijk})} \qquad \text{("logarithmic")}$$
$$L_{ij(k+1)} = \sqrt{(A_{ijk} * A_{ij(k+1)})}$$

For the lower limit of the first range,

$$L_{ij1} = \max[0, (3A_{ij1} - A_{ij2})] \qquad \text{("arithmetic"), or}$$
$$L_{ij1} = \sqrt{(A_{ij1}^3/A_{ij2})} \qquad \text{("logarithmic")}$$

and for the upper limit of the uppermost range

$L_{ij(n+1)} = (3A_{ijn} - A_{ij(n-1)})/2$ ("arithmetic"), or

$L_{ij(n+1)} = \sqrt{(A_{ijn}^3/A_{ij(n-1)})}$ ("logarithmic")

If only a single range has been defined for an OTU, its limits are defined in relation to the values of this attribute in other OTUs. If one or more of the other OTUs have been recorded with multiple values for this attribute, the range of the different medians (arithmetic or logarithmic) is averaged:

$\Sigma_i (A_{ijn} - A_{ij1})/2m$ ("arithmetic"), or

$\Sigma_i \{A_{ijn}/A_{ij1}\}^{1/2(n-1)}/m$ ("logarithmic")

where summation over *i* only applies to the m OTUs for which multiple ranges of values have been recorded.

Where no ranges of values are recorded for any OTU in the data, the putative range of values around any single stated value for one of the OTUs is related arbitrarily to the extreme range for the attribute in the data set. Thus, one calculates

$B = \max_i A_{ijn}$, , $C = \min_i A_{ij1}$

B and C being the limits of values of attribute *j* for the m OTUs for which this attribute has been recorded. One then calculates G as the assumed half-range around the stated median, where no more apposite information is available:

$G = (B - C)/2(m-1)$ ("arithmetic"), or

$G = (B/C)^{1/2(m-1)}$ ("logarithmic")

These values of G are then used to set the ranges around the single values recorded for each OTU:

$L_{ij1} = \max[0, (A_{ij1} - G)]$, $L_{ij2} = A_{ij1} + G$ ("arithmetic") or

$L_{ij1} = A_{ij1} / G$, $L_{ij2} = A_{ij1} * G$ ("logarithmic")

For the propositus (x), the data provided by the key user may also not include limits for ranges, so indirect estimation of limits may also be needed here. As indicated above, the value of this attribute in the propositus may be described in three distinct ways:

(1) Two or more ranges of values may be specified, each with a median value and a proportion (as for the OTU's included in the data set);

(2) A maximum and minimum may be specified; or

(3) A single value may be specified.

These three cases are treated differently in converting the data to the form required, and then setting limits for ranges.

In case (1), limits for the propositus are defined exactly as described above for OTUs in the data set; these limits will be symbolized by

L_{xjk} , k=1,(n+1)

In case (2), where maximum and minimum are defined, these minimum and maximum values, B_{min} and B_{max}, become the limits for the single range defined for

this attribute in the propositus: $L_{xj1} = B_{min}$, $L_{xj2} = B_{max}$,with the median of the range being

$A_{xj1} = (L_{xj1} + L_{xj2})/2$ ("arithmetic"), or
$A_{xj1} = \sqrt{(L_{xj1} . L_{xj2})}$ ("logarithmic").

Since there is only one range defined, $P_{xj1} = 1$.

In case (3), where a single value is named for the propositus, this value becomes A_{xj1} and again $P_{xj1} = 1$. The limits L_{xj1} and L_{xj2} are defined as described above for OTUs with a single range:

$L_{xj1} = \max[0, (A_{xj1} - G)]$, $L_{xj2} = A_{xj1} + G$ ("arithmetic"); or
$L_{xj1} = A_{xj1} / G$, $L_{xj2} = A_{xj1} * G$ ("logarithmic")

References

Dallwitz MJ. 1974. A flexible computer program for generating diagnostic keys. Syst.Zool.23:50-57.

Dallwitz MJ. 2009. Programs for interactive identification and information retrieval http://delta-intkey.com.

Dallwitz MJ, Paine TA, Zurcher EJ 2008. User's guide to the DELTA system: a general system for processing taxonomic descriptions. 4th Edition. (2008) http://delta-intkey.com

Goodall D.W. 1966 A new similarity index based on probability. Biometrics 22: 882-907.

Goodall DW. *1968* Identification by computer. BioScience 18 pp. 485-488.

Goodall DW. 1993. Probabilistic indices for classification - some extensions. Abs. Bot. 17:125-132

Goodall DW. 1994. The treatment of spatial data in probabilistic classification. Abs. Bot. 18: 45-47.

Goodall DW, Ganis P, Feoli E. 1991. Probabilistic methods in classification; a manual for seven computer programs, in Computer Assisted Vegetation Analysis, E. Feoli and L.Orloci, eds., Kluwer Academic, pp. 453 – 467.

Goodall DW, Marchant NG. 1996. Consistency in taxonomic rank: an example from *Drosera.* Abs. Bot. 20:1--15.

Gower JC, Payne RW.1975. A comparison of different criteria for selecting binary tests in diagnostic keys. Biometrika 62: 665-672.

Grosser D, Conruyt N, Ralambondrainy H 2010 Identification with iterative nearest neighbors using domain knowledge In: Tools for identifying biodiversity: progress and problems, P.L.Nimis & R.V.Lebbe *pp. 71-76.*

Hagedorn, G., Fanbold, G. & Martellos,S. 2010 Types of identification keys. In: Tools for identifying biodiversity: progress and problems Nimis, P.L. & Cable, R.V. (Ed.) pp. 59-64.

Marchant N G, George AS . 1982. *DROSERA*, in Flora of Australia, *Vol.8*, Aust.Govt. Publ.Serv., Canberra, pp. 9-64.
Nimis PL, Lebbe RV. 2010. Tools for identifying biodiversity: progress and problems.EUT, Trieste Italy.
Nimis PL, Riccamboni R, Martellos S. 2012. Identification keys on mobile devices: The Dryades experience. Plant Biosystems 146:783-788.
Norton G. 2002 Multi-media/internet keys and the taxonomic crisis. Antenna 26: 245- 248
Pankhurst RJ. 1970. A computer program for generating diagnostic keys Computer J. 13:145-151.
Pankhurst *RJ. 1975*. Biological Identification with computers*. (Syst.Assoc. Spec. Vol 7)*. Academic Press, London, New York & San Francisco. 1975.
Pankhurst RJ. 1991. Practical taxonomic computing. Cambridge University Press, Cambridge.
Payne RW, Preece DA. 1980 Identification keys and diagnostic tables: a review. J. Roy.Statist.Soc. Ser.A (General) 143:253-292.
Sneath PHA. 1980. BASIC program for the most diagnostic properties of groups from an identification matrix of percent positive characters. Computers & Geosciences 6 (1980), pp.21-26.