Botanical Keys Generated by Computer

R. J. PANKHURST

Computer Laboratory, University of Cambridge

INTRODUCTION

This paper describes the application of a computer program to produce botanical keys. A brief summary of this work has already appeared (Pankhurst, 1971) and the methods used are discussed here in detail, the techniques of computation having been considered in earlier papers (Pankhurst 1970a, b). In all cases the taxonomy from which the data is derived is taken as already established. The key generating program should be thought of as an aid to the identification of specimens, not as a method for classification.

ADVANTAGES OF THE METHOD

The generation of keys by computer offers the following advantages:

(1) The manual labour of actually constructing the key is avoided.

(2) The key is as accurate as the data on which it is based. In manual key construction, it is very difficult to be certain that the textual descriptions of taxa correspond exactly with a key to the same material.

(3) An attempt can be readily made to shorten the key. The strategy by which the key is constructed allows for an effort to make the key as efficient as possible in the following sense. The most efficient key is taken to be that which has the shortest average path (i.e. number of leads used) to a correct identification of all taxa. It has been shown (Osborne 1963) that the key with the shortest average route to an identification is dichotomous with groups of taxa of equal size falling into each lead of a pair. The construction of such a key can also take into account the different usefulness or ease of observation of the characters used by the key and the commonness or rarity of the taxa. Improvement of keys in the above sense is not feasible with manually constructed keys above a certain size without considerable effort.

(4) It is easy to edit the key. If one wishes to add or remove taxa, or to add, remove, correct or re-arrange the characters used, then a new key is produced just by running the computer program again. In particular, this makes it easy to produce keys based on the same taxonomy but tailored for different needs, as for example, keys for herbarium specimens, or keys to perennials in winter. This flexibility is achieved by altering the relative importance (weighting) of the characters.

PERFORMANCE OF KEYS

A key is expected first of all to be accurate, and then in addition to be efficient. An accurate key gives a correct identification with the great majority of specimens. The most efficient key is one that requires the least effort on the part of the user, and the theorem of Osborne indicates how this may be done.

A key may give a wrong identification because of a number of different types of error, which are discussed below.

(1) The specimen may belong to a taxon which is not included in the key in any case. This should become apparent from either definite disagreement or an unlikely result. An unlikely result is recognised by previous experience, or failing that, by disagreement with other details not in the key, such as habitat or distribution. Hence, as many details as possible of each taxon should be used in the key, in order that taxa which do not belong are seen to disagree. However, since a key can often be constructed without using all the available characters, some of the remaining ones are often included at the point where a taxon keys out. This helps the key-user to check his identification.

(2) Leads in a key can be misunderstood or characters of a specimen can be misread. If there is only one character per lead in the key, then a wrong branch can more easily be taken. This is much more serious at the beginning of a key than towards the end, because the error causes the key-user to consider a larger number of taxa, none of which is the right one. As a precaution against errors of this kind, keys which have several characters in each lead are preferred, since there is less doubt if several characters agree. Alternatively, several keys using different sets of characters should be tried. If we give a numerical value to the probability of answering each lead correctly, then these probabilities have to be multiplied together to give the probability of arriving correctly at the end of a sequence of leads, and finally to keying out. Suppose as an example that this probability is 0·9, and consider a key for 50 taxa. Suppose that the key is dichotomous, in which case about 6 leads will be used as a specimen of each taxon is identified. This is because, ideally, two taxa could be separated by one lead, four taxa by two leads, and so on. In general, a key for N taxa involves the use of at least n leads per taxon, where 2^n is approximately equal to N. In the above example, the expected probability of getting a correct result for an average taxon is 0·9 to the power of 6, i.e. about one half. In such a case, the key is as likely to be right as it is likely to be wrong. Therefore, if errors occur in the leads of a key, the key should be as short as possible to increase the likelihood of a right result. This is, incidentally, a reason why keys to more than a few hundred taxa are rare; even with a high probability of answering each lead correctly, the chance of a correct final result can be small, so that large keys are rather impractical.

It is often stated that 'good' characters should be used in keys. A good character is one which is both easy or cheap to determine and which has a high probability of being correctly read. The ease of determination might be a matter of time spent, or the need for equipment, training, or previous experience, or the cost of chemicals or a technician's pay. A character will more likely be read correctly if it is, among other things, constant, available on average specimens, and not too technical. These aspects of the quality of a character are partly subjective, so measurement of the cost and probability is not generally practical. For reasons of accuracy, those characters with the highest probability will be used first in a key, since it is at the beginning of a key that an error would have the worst effect. For reasons of efficiency, the easiest or cheapest characters should be used first and the others, if necessary, used later. In general, therefore, the characters in a key encountered in successive leads are often in the order of decreasing quality and increasing cost.

(3) Characters used in the key can be missing from the specimen. This can be

due, for example, to not having a complete plant, or to deterioration after pressing, or to collection at the wrong season. If the key has only one character in a lead at some point, and that character is missing, it is impossible to continue. As above, this may be avoided by having more than one character in a lead, or by having a variety of keys using different sets of characters. If a key with several characters per lead can be provided, then the user may use the key by selecting whichever characters he does happen to have. This is then really a matter of selecting one of a number of alternative keys which are expressed in the form of one. In this case, therefore, multi-character leads may be used as if they are single-character leads. Key-users may often select only one character from a multi-character lead, even when all the characters are available, in order to save effort.

Once the considerations of accuracy have been considered, one may turn to questions of efficiency. Dichotomous keys (where the leads are in pairs) are often preferred over polychotomous keys (where there can be more than two leads as alternatives). This is simply because a choice between two alternatives is easier than between many. Each lead considered adds to the effort required. It is also easier to choose between single character leads than between multi-character leads in which every character is considered. Reasons have been given for preferring multi-character leads, but, unfortunately, it is generally difficult to find sets of characters in natural groups for use in leads which give an efficient dichotomy.

More explicitly, multi-character leads producing an efficient dichotomy are hard to construct at the top of a key, but easy to construct at the bottom, where they occur as lists of differences between alternative taxa. A device which is often used to counter this difficulty is to qualify auxiliary characters with adverbs such as 'usually' or 'rarely'. This is of limited usefulness because of a clear element of doubt in the extra characters, as for example, in the following pair of two-character leads:

1. Leaves hairy, flowers usually white.
   Leaves glabrous, flowers pink.

As has been mentioned above, it has been shown using formal mathematics (Osborne 1963) that the key with the shortest average route to an identification is dichotomous with groups of taxa of equal size falling into each lead of a pair. Such an arrangement is an ideal which can only be approximated to in practice, because of unequal distribution of characters and odd numbers of taxa. This does however give a very useful rule of thumb for choosing an efficient key. Notice that the 'shortest average route' should take into account the frequency with which specimens are found, so that common taxa should key out after only a few leads, and rarer ones after more. Many keys will be seen to key out unusual taxa near the beginning. If an unusual taxon is taken as one which differs in many of its characters from the others, it may therefore be expected to key out after rather few leads. If it is unusual in only a few characters, then the efficiency of the key may be reduced by making it key out near the beginning.

From the above discussion it is evident that the choice between few or many leads, or single or multi-character leads, is a matter of compromise. The computer program discussed below takes the following general strategy. The order of the characters is used to decide on how to satisfy accuracy before efficiency. If no order is specified, then only considerations of efficiency are used, to try and shorten the key. If an order is specified, and if, after this has been satisfied, there is still some freedom of choice, rules of efficiency are used. The program will try out as many characters per lead as it is asked to consider, but will check that the multi-character leads are reasonably efficient, and so may still prefer single character leads. All the remaining distinctive characters are given when a taxon keys out, since this is useful and it does not interfere with efficiency.

PROGRAM DATA AND OPTIONS

There is no restriction on the rank of taxa which may be put in to the program. One may equally well work with families, genera, species, subspecies, varieties or microspecies. There is nothing in fact in the program's design to restrict its application to botany. The taxa are described as a rectangular matrix of taxa in rows versus characters (features, properties, attributes) in columns, filled with values. Both the character and value are usually phrases in natural language: for example, *Epilobium roseum* has an entire stigma, which is expressed as a character 'stigma' with a value 'entire'. The choice of the characters, and the values assigned to them, are entirely in the hands of the program user. Two or several values can be used with each character, i.e. characters can be two- or many-valued. A value can just as readily be qualitative (e.g. 'long') as quantitative (e.g. '10 to 20 cm'). Values can also be missed out if they are not known, inapplicable or too variable to be useful.

The program prefers to produce dichotomous keys where possible, but can create polychotomous keys if necessary. This depends on what proportion of many-valued characters are given. The user may specify the maximum number of characters that the program is to consider in combination as each branch of the key is formed. The characters can be given equal weights, or any sequence of weighting that the botanist prefers. It is also possible to give weights to the taxa, the highest weight going to the commonest, so that the program will try to arrange for a short path to common taxa, as far as the data allow.

Either of the two common forms of key can be produced, with an option for indentation if required. Provision is made for highly distinctive taxa to be recognised and keyed out early, although this is deliberately restricted because it may interfere with the process of shortening the key. If the taxa are not all distinguishable from one another with the given data, then a partial key can be formed, such that several taxa may key out together. The character matrix has to be sufficiently well scored before it is possible for keys to be constructed. This requires that, initially, all the values of at least one of the characters must be supplied. The names of the taxa form part of the data, and these may all be different, or some of them may be the same. In the latter case, unnecessary key branches involving only different descriptions of the same taxon are avoided. As an aid in checking the data, the program will print a synopsis of all taxa. This could be useful to a Flora writer who wishes to make his text agree exactly with his key.

METHOD OF CONSTRUCTING KEYS

It should be remarked that since computer programs can evolve very rapidly, the following discussion is only strictly accurate at the time of writing.

At the beginning of the key construction process, all the taxa and all their characters are available for consideration. The characters which have missing values are then temporarily dismissed, since they cannot be used without uncertainty. The fully scored characters are then examined to see which one will form the best division. Suppose for the moment that all the characters have equal weighting. If one character has two different values, then this character could be used to form a pair of leads. If we ignore taxon weighting for the moment, then this character would be a good choice if the two groups of taxa with the two different values are about equal in size. Other characters might have more than two values, but could still form conveniently equally-sized groups with the different values. Another possibility is to choose any pair of characters and construct leads with two characters at a time. This may be expected to give rise to more than two leads and polychotomous branching. Then one may consider three characters at a time, and so on, up to the maximum number of characters available. In practice, a limit is set in the data as to the maximum number of characters in a lead to be considered. Up to this limit, all the possible combinations of characters are examined. The choice between different possible ways of branching the key is made by calculating which is the best from the number of leads and the number of taxa in each group. This is described in detail below.

Once a branch has been made, then the existing leads and a list of the taxa belonging to each lead are remembered. The process is then repeated on each unfinished branch of the key. Notice that this time only those taxa which belong to the previous lead are considered, and only those characters which have not already been used in the preceding sequence of leads. To put this another way, some of the rows and columns of the character matrix are ignored when lower branches of the key are created.

It was assumed above that all the characters were equally weighted. If they have different weights, then each time a new key branch is to be made, an unused character with the highest weight is considered. It is possible for more than one character to have the same weight. If just one character has a higher weight than the others, then this one is used at once to make the next branch, and the other characters are not examined. If several characters share the highest weight, then these are examined as described above to choose the best way of making a new branch, and those of lower weight are ignored, unless the higher weighted characters turn out to have missing values.

A simple method of calculation is used to choose the best way of branching; it is based only on convenience and has no theory to back it. The result of the calculation is arranged so that it is always positive, and would be zero in the ideal case. The key branching which gives the smallest result is the one which is chosen. The result has two parts. The first part depends only on the number of leads at a branch, and is given by $(K-2)^2$, where K is the number of leads. A dichotomy is the case $K = 2$, which makes the first part of the result zero. A polychotomy of four branches ($K = 4$) would give $(4 - 2)^2 = 4$, which is larger and not so favoured.

The second part depends on how equally the taxa are distributed in groups, and is small for groups of equal size and large for unequal groups. A numerical example will be helpful. Suppose we have a dichotomy with 2 taxa in one group and 8 in the other, which is a total of 10 taxa. Since there are two groups, the ideal would be if each group were of size $\frac{10}{2} = 5$. If we take the ratio of the actual sizes to the ideal we have $\frac{2}{5} = \cdot 4$ and $\frac{8}{5} = 1{\cdot}6$. Ideally these ratios would each be unity. The differences from one with the sign removed are $1 - \cdot 4 = \cdot 6$ and $1{\cdot}6 - 1 = \cdot 6$ and these are added to give a result for the second part of $1{\cdot}2$. A dichotomy of 4 and 6 respectively gives the second part equal to $\cdot 4$, and would be taken as a better grouping. This calculation depends only on the relative number of taxa in different groups. For example, if we take 4 and 16 taxa in a group of 20, the second part of the result is still $1{\cdot}2$. The two parts are added together, dichotomies with equal groups therefore giving the lowest result.

Taxon weighting, if required, is allowed for by using the numerical value of the weighting to be an indication of the frequency of the taxon concerned. A common taxon will be given a high weighting. If the original example above consists of two common species with weighting 6 each, and 8 local species each with weighting 1, then the effective number of species in the first group is now 6 times greater i.e. 12. The total effective number of species is now $12 + 8 = 20$. The result for the second part of the calculation will now be $\cdot 4$ instead of 12, which makes it more favourable to branch the key this way. The two common species would then be separated by one more dichotomy, whereas the eight local species would need about three more branches. Hence the desired effect of shortening the path through the key to the identification of common taxa is achieved. This method of taxon weighting was derived from that of Hall (1970).

To express the above as a mathematical formula for the case of no taxon weighting, suppose that the size of each of the K groups is n_i for $i = 1$ up to K, and that the total number of taxa is N, where $N = \sum_{i=1}^{K} n_i$. Then the next key branch is chosen by finding the minimum value of

$$(K-2)^2 + \sum_{i=1}^{K} \left| 1 - \frac{n_i K}{N} \right|$$

where the vertical lines stand for the absolute value, i.e. the expression between the lines with the sign removed. The corresponding formula for taxon weighting is

$$(K-2)^2 + \sum_{i=1}^{K} \left| 1 - \frac{n_i' K}{N'} \right|$$

where n_i' is the effective number of taxa in group i (the sum of the taxon weights) and N' is the effective total number of taxa.

Once the key is complete, the program works through it again to put in the lead numbers, both before and after each lead. The key is remembered in the order suitable for the indented type of key. If it is indeed this form of key that is wanted, then the leads are sorted in the order of groups of taxa ascending size, since this gives a better printed layout. If it is the bracketted type of key that is required, then the details of the key are re-ordered after generating the key and before printing it. Indentation is achieved by remembering with each lead how many previous leads preceded the current lead, and this gives the number of print positions to allow for the indentation.

In addition to the procedure outlined above, the program makes an attempt to key out unusual taxa earlier than the rest. This is omitted if any of the taxa to be compared have missing values of characters. The similarity of the taxa to each other is estimated and those which are most unlike the rest may be keyed out separately, without the usual branch-choosing procedure.

EXAMPLES

The keys printed here as examples (Figs. 1–3) cover the species of *Epilobium* L. found in Great Britain. Fig. 1 is produced by giving all the characters equal weighting, so that the program alone has decided which to use and in what order. Fig. 2 emphasises floral features while Fig. 3 emphasises vegetative features. The second and third keys use character weighting to achieve the different emphasis. All three keys were produced from one computer run, and are all indented and in one of the two possible styles. This is a small example, as it involves only 13 species. The flexibility of the computer approach shows well in this example. The ability of the program to produce keys based on fewer characters and with fewer leads as compared with hand-written versions only begins to show with the order of 50 or more taxa. The keys in all three figures have been reproduced as far as possible in the form that the computer prints them. There are restrictions on the letters that can be used in FORTRAN programming, which is why the keys cannot appear in the conventional style of printing without being re-typed.

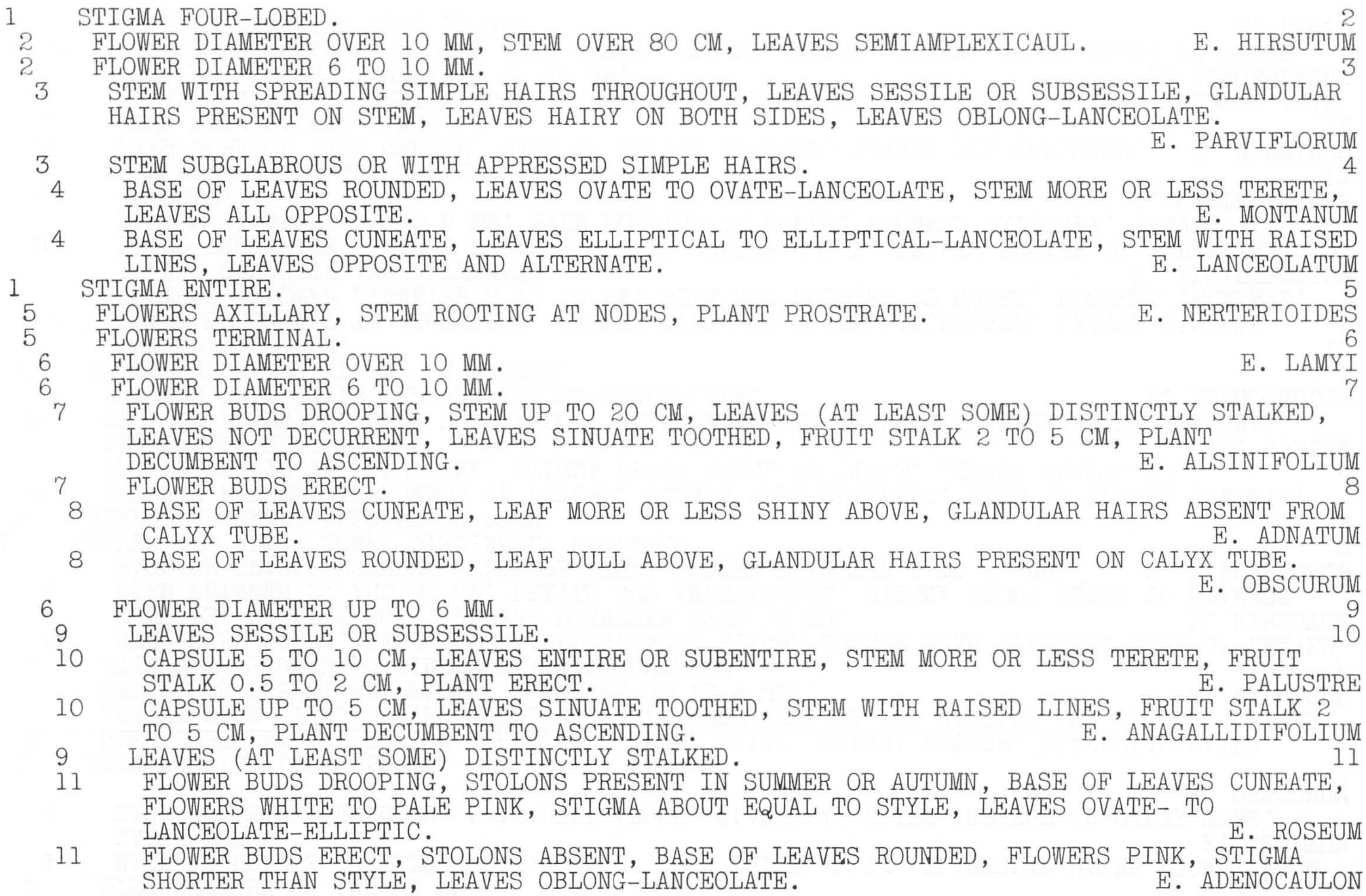
The data used for the *Epilobium* keys were derived from the *Flora of the British Isles* (Clapham, Tutin & Warburg 1962) and the *Illustrated Flora of Great Britain* (Butcher 1961). Certain rules of thumb were used in transcribing the written descriptions of species to the data matrix. It is sometimes necessary to make deductions about character values from descriptions of the form 'species B is like A but . . .'. Care is needed with adverbs such as 'usually', 'rarely' and 'sometimes' used in describing characters. One may choose to leave out the value of such a variable character altogether, or repeat the taxon with different character values, or just score the 'usual' value. The last choice means that a variable character is incompletely scored, and the finished key can then fail on certain 'unusual' specimens. It may be convenient to make simplifications, for example in *Epilobium* 'pale purple-lilac' was scored as 'pink'.

It was stated that the key-generating program prefers to work with characters which are fully scored. This gives the program a better opportunity to create a short key. This criterion of completeness is not usually required when descriptions of taxa are prepared for Floras, so that, in the *Epilobium* example, the data matrix prepared from two Floras was more complete than that which could have been prepared from either alone. The experience with key-generating to date shows that the collection of data on which to base a key requires most of the effort.

The earliest example of a computer-generated key was based on the microspecies of *Alchemilla vulgaris*. A key to the 134 genera of the Umbelliferae in *Flora Europaea* (Tutin *et alia* 1968) was computed from data supplied by Professor Tutin. These data came from an intermediate stage in the preparation of the *Flora*, and are somewhat incomplete. The computed key proved to have 20% fewer leads, and was based on fewer characters, than the handwritten version. Keys for up to 116 potato cultivars have been computed for Mr T. Webster of the National Institute of Agricultural Botany. One such key has been published (Webster 1969). These keys have been especially welcomed since, on account of the effort required, no handwritten key has ever been prepared.

```
1    LEAVES SESSILE OR SUBSESSILE.                                                           2
 2    STEM WITH RAISED LINES.                                                                3
  3    STEM UP TO 20 CM, LEAVES NOT DECURRENT, LEAVES SINUATE TOOTHED, FLOWER DIAMETER UP TO
       6 MM, FRUIT STALK 2 TO 5 CM, PLANT DECUMBENT TO ASCENDING.         E. ANAGALLIDIFOLIUM
  3    STEM BETWEEN 20 AND 80 CM.                                                            4
   4    BASE OF LEAVES CUNEATE, LEAF MORE OR LESS SHINY ABOVE, GLANDULAR HAIRS ABSENT FROM
        CALYX TUBE.                                                                 E. ADNATUM
   4    BASE OF LEAVES ROUNDED, LEAF DULL ABOVE, GLANDULAR HAIRS PRESENT ON CALYX TUBE.
                                                                                   E. OBSCURUM
 2    STEM MORE OR LESS TERETE.                                                              5
  5    STEM SUBGLABROUS OR WITH APPRESSED SIMPLE HAIRS, STIGMA ENTIRE, GLANDULAR HAIRS
       ABSENT FROM STEM, LEAVES SUBGLABROUS (EXCEPT PERHAPS FOR MARGINS AND VEINS), LEAVES
       ENTIRE OR SUBENTIRE, FLOWER DIAMETER UP TO 6 MM.                            E. PALUSTRE
  5    STEM WITH SPREADING SIMPLE HAIRS THROUGHOUT.                                          6
   6    STEM OVER 80 CM, LEAVES SEMIAMPLEXICAUL, STIGMA LONGER THAN STAMENS, BASE OF LEAVES
        CUNEATE, FLOWERS ROSE, FLOWER DIAMETER OVER 10 MM.                         E. HIRSUTUM
   6    STEM BETWEEN 20 AND 80 CM, LEAVES NOT AMPLEXICAUL, STIGMA ABOUT EQUAL TO STAMENS,
        BASE OF LEAVES ROUNDED, FLOWERS PINK, FLOWER DIAMETER 6 TO 10 MM.      E. PARVIFLORUM
1    LEAVES (AT LEAST SOME) DISTINCTLY STALKED.                                              7
 7    BASE OF LEAVES CUNEATE.                                                                8
  8    LEAVES DECURRENT, FLOWER BUDS ERECT, FLOWER DIAMETER OVER 10 MM.               E. LAMYI
  8    LEAVES NOT DECURRENT.                                                                 9
   9    STIGMA FOUR-LOBED, GLANDULAR HAIRS ABSENT FROM STEM, STIGMA SHORTER THAN STYLE,
        LEAVES ELLIPTICAL TO ELLIPTICAL-LANCEOLATE, FLOWER DIAMETER 6 TO 10 MM.
                                                                               E. LANCEOLATUM
   9    STIGMA ENTIRE, GLANDULAR HAIRS PRESENT ON STEM, STIGMA ABOUT EQUAL TO STYLE, LEAVES
        OVATE- TO LANCEOLATE-ELLIPTIC, FLOWER DIAMETER UP TO 6 MM.                   E. ROSEUM
 7    BASE OF LEAVES ROUNDED.                                                               10
 10    FLOWER DIAMETER UP TO 6 MM.                                                          11
  11    GLANDULAR HAIRS PRESENT ON STEM, LEAVES OBLONG-LANCEOLATE, LEAVES DENTICULATE, STEM
        NOT ROOTING AT NODES, FLOWERS TERMINAL, PLANT ERECT, LEAVES OPPOSITE AND ALTERNATE.
                                                                               E. ADENOCAULON
  11    GLANDULAR HAIRS ABSENT FROM STEM, LEAVES BROAD OVATE TO SUBORBICULAR, LEAVES ENTIRE
        OR SUBENTIRE, STEM ROOTING AT NODES, FLOWERS AXILLIARY, PLANT PROSTRATE, LEAVES ALL
        OPPOSITE.                                                              E. NERTERIOIDES
 10    FLOWER DIAMETER 6 TO 10 MM.
  12    STEM BETWEEN 20 AND 80 CM, STIGMA FOUR-LOBED, LEAVES DENTICULATE, STEM MORE OR LESS
        TERETE, FRUIT STALK 0.5 TO 2 CM, PLANT ERECT, LEAVES ALL OPPOSITE.         E. MONTANUM
  12    STEM UP TO 20 CM, STIGMA ENTIRE, LEAVES SINUATE TOOTHED, STEM WITH RAISED LINES,
        FRUIT STALK 2 TO 5 CM. PLANT DECUMBENT TO ASCENDING, LEAVES OPPOSITE AND ALTERNATE.
                                                                               E. ALSINIFOLIUM
```

Figure 1. Key to *Epilobium* species



```
1    STIGMA FOUR-LOBED.                                                           2
 2    FLOWER DIAMETER OVER 10 MM, STEM OVER 80 CM, LEAVES SEMIAMPLEXICAUL.      E. HIRSUTUM
 2    FLOWER DIAMETER 6 TO 10 MM.                                                  3
  3    STEM WITH SPREADING SIMPLE HAIRS THROUGHOUT, LEAVES SESSILE OR SUBSESSILE, GLANDULAR
       HAIRS PRESENT ON STEM, LEAVES HAIRY ON BOTH SIDES, LEAVES OBLONG-LANCEOLATE.
                                                                          E. PARVIFLORUM
  3    STEM SUBGLABROUS OR WITH APPRESSED SIMPLE HAIRS.                            4
   4    BASE OF LEAVES ROUNDED, LEAVES OVATE TO OVATE-LANCEOLATE, STEM MORE OR LESS TERETE,
        LEAVES ALL OPPOSITE.                                                 E. MONTANUM
   4    BASE OF LEAVES CUNEATE, LEAVES ELLIPTICAL TO ELLIPTICAL-LANCEOLATE, STEM WITH RAISED
        LINES, LEAVES OPPOSITE AND ALTERNATE.                              E. LANCEOLATUM
1    STIGMA ENTIRE.                                                               5
 5    FLOWERS AXILLARY, STEM ROOTING AT NODES, PLANT PROSTRATE.          E. NERTERIOIDES
 5    FLOWERS TERMINAL.                                                            6
  6    FLOWER DIAMETER OVER 10 MM.                                             E. LAMYI
  6    FLOWER DIAMETER 6 TO 10 MM.                                                 7
   7    FLOWER BUDS DROOPING, STEM UP TO 20 CM, LEAVES (AT LEAST SOME) DISTINCTLY STALKED,
        LEAVES NOT DECURRENT, LEAVES SINUATE TOOTHED, FRUIT STALK 2 TO 5 CM, PLANT
        DECUMBENT TO ASCENDING.                                        E. ALSINIFOLIUM
   7    FLOWER BUDS ERECT.                                                         8
    8    BASE OF LEAVES CUNEATE, LEAF MORE OR LESS SHINY ABOVE, GLANDULAR HAIRS ABSENT FROM
         CALYX TUBE.                                                         E. ADNATUM
    8    BASE OF LEAVES ROUNDED, LEAF DULL ABOVE, GLANDULAR HAIRS PRESENT ON CALYX TUBE.
                                                                            E. OBSCURUM
  6    FLOWER DIAMETER UP TO 6 MM.                                                 9
   9    LEAVES SESSILE OR SUBSESSILE.                                             10
   10    CAPSULE 5 TO 10 CM, LEAVES ENTIRE OR SUBENTIRE, STEM MORE OR LESS TERETE, FRUIT
         STALK 0.5 TO 2 CM, PLANT ERECT.                                    E. PALUSTRE
   10    CAPSULE UP TO 5 CM, LEAVES SINUATE TOOTHED, STEM WITH RAISED LINES, FRUIT STALK 2
         TO 5 CM, PLANT DECUMBENT TO ASCENDING.                      E. ANAGALLIDIFOLIUM
   9    LEAVES (AT LEAST SOME) DISTINCTLY STALKED.                                11
   11    FLOWER BUDS DROOPING, STOLONS PRESENT IN SUMMER OR AUTUMN, BASE OF LEAVES CUNEATE,
         FLOWERS WHITE TO PALE PINK, STIGMA ABOUT EQUAL TO STYLE, LEAVES OVATE- TO
         LANCEOLATE-ELLIPTIC.                                                 E. ROSEUM
   11    FLOWER BUDS ERECT, STOLONS ABSENT, BASE OF LEAVES ROUNDED, FLOWERS PINK, STIGMA
         SHORTER THAN STYLE, LEAVES OBLONG-LANCEOLATE.                    E. ADENOCAULON
```

Figure 2. Key to *Epilobium* species (floral characters weighted).

```
1    LEAVES SESSILE OR SUBSESSILE.                                                  2
 2    STEM WITH RAISED LINES.                                                       3
  3    STEM UP TO 20 CM, LEAVES NOT DECURRENT, LEAVES SINUATE TOOTHED, FLOWER DIAMETER UP TO
       6 MM, FRUIT STALK 2 TO 5 CM, PLANT DECUMBENT TO ASCENDING.        E. ANAGALLIDIFOLIUM
  3    STEM BETWEEN 20 AND 80 CM.                                                    4
   4    BASE OF LEAVES CUNEATE, LEAF MORE OR LESS SHINY ABOVE, GLANDULAR HAIRS ABSENT FROM
        CALYX TUBE.                                                           E. ADNATUM
   4    BASE OF LEAVES ROUNDED, LEAF DULL ABOVE, GLANDULAR HAIRS PRESENT ON CALYX TUBE.
                                                                              E. OBSCURUM
 2    STEM MORE OR LESS TERETE.                                                     5
  5    STEM SUBGLABROUS OR WITH APPRESSED SIMPLE HAIRS, STIGMA ENTIRE, GLANDULAR HAIRS
       ABSENT FROM STEM, LEAVES SUBGLABROUS (EXCEPT PERHAPS FOR MARGINS AND VEINS), LEAVES
       ENTIRE OR SUBENTIRE, FLOWER DIAMETER UP TO 6 MM.                      E. PALUSTRE
  5    STEM WITH SPREADING SIMPLE HAIRS THROUGHOUT.                                  6
   6    STEM OVER 80 CM, LEAVES SEMIAMPLEXICAUL, STIGMA LONGER THAN STAMENS, BASE OF LEAVES
        CUNEATE, FLOWERS ROSE, FLOWER DIAMETER OVER 10 MM.                    E. HIRSUTUM
   6    STEM BETWEEN 20 AND 80 CM, LEAVES NOT AMPLEXICAUL, STIGMA ABOUT EQUAL TO STAMENS,
        BASE OF LEAVES ROUNDED, FLOWERS PINK, FLOWER DIAMETER 6 TO 10 MM.   E. PARVIFLORUM
1    LEAVES (AT LEAST SOME) DISTINCTLY STALKED.                                     7
 7    GLANDULAR HAIRS PRESENT ON STEM.                                              8
  8    STOLONS PRESENT IN SUMMER OR AUTUMN, FLOWER BUDS DROOPING, BASE OF LEAVES CUNEATE,
       FLOWERS WHITE TO PALE PINK, STIGMA ABOUT EQUAL TO STYLE, LEAVES OVATE- TO
       LANCEOLATE-ELLIPTIC.                                                    E. ROSEUM
  8    STOLONS ABSENT, FLOWER BUDS ERECT, BASE OF LEAVES ROUNDED, FLOWERS PINK, STIGMA
       SHORTER THAN STYLE, LEAVES OBLONG-LANCEOLATE.                      E. ADENOCAULON
 7    GLANDULAR HAIRS ABSENT FROM STEM.                                             9
  9    STEM UP TO 20 CM.                                                            10
  10    PLANT DECUMBENT TO ASCENDING, LEAVES OVATE TO OVATE-LANCEOLATE, LEAVES SINUATE
        TOOTHED, FLOWER DIAMETER 6 TO 10 MM, STEM NOT ROOTING AT NODES, FLOWERS TERMINAL,
        LEAVES OPPOSITE AND ALTERNATE.                                   E. ALSINIFOLIUM
  10    PLANT PROSTRATE, LEAVES BROAD OVATE TO SUBORBICULAR, LEAVES ENTIRE OR SUBENTIRE,
        FLOWER DIAMETER UP TO 6 MM, STEM ROOTING AT NODES, FLOWERS AXILLARY, LEAVES ALL
        OPPOSITE.                                                        E. NERTERIOIDES
  9    STEM BETWEEN 20 AND 80 CM.                                                   11
  11    STEM MORE OR LESS TERETE, BASE OF LEAVES ROUNDED, LEAVES ALL OPPOSITE.  E. MONTANUM
  11    STEM WITH RAISED LINES.                                                     12
   12    LEAVES NOT DECURRENT, FLOWER BUDS DROOPING, STIGMA FOUR-LOBED, STIGMA SHORTER THAN
         STYLE, FLOWER DIAMETER 6 TO 10 MM.                              E. LANCEOLATUM
   12    LEAVES DECURRENT, FLOWER BUDS ERECT, STIGMA ENTIRE, STIGMA ABOUT EQUAL TO STYLE,
         FLOWER DIAMETER OVER 10 MM.                                           E. LAMYI
```

Figure 3. Key to *Epilobium* species (vegetative characters weighted).

Other keys have been created for common species of *Cortinarius*, microspecies of *Rubus fruticosus* in Cambridgeshire, European species of *Veronica*, and species of *Jurinea* (Compositae). The *Veronica* key used fewer leads and less than half the characters than the corresponding handwritten key.

Keys relevant to other disciplines have been computed, in particular in zoology, geology, medicine and mechanical and electrical engineering. Further keys, generated at other institutions which have been given copies of the program by the author, are not discussed here.

OTHER WORK

Several efforts have been made to compute the structure of keys, although nearly all of these are numerical calculations without any arrangement to write out the key at the finish (discussed in Pankhurst 1970a). Research by a variety of authors in that branch of computer science known as artificial or machine intelligence has produced a number of papers on decision trees, which are similar to keys (discussed in Pankhurst 1970a). None of these authors appeared to be aware of the use of keys in biology.

The only comparable program known to me is that due to Morse (Morse 1968, Shetler *et alia* 1971). This is similar to mine in its aims and achievements, but differs considerably in detail. For instance, leads are not split into pairs of characters and values, but each character value combination is treated as an indivisible 'couplet'. Character weighting takes effect by being included in the computation for choosing key branching along with the number and size of groups of taxa.

CONCLUSIONS

Diagnostic keys, until recently always composed by hand, can now be generated by computer. A program to do this, written in standard FORTRAN computer language, is available from the author on request.

ACKNOWLEDGMENTS

The author is indebted to Dr S. M. Walters of the Botany School, Cambridge, for encouragement and valuable criticism throughout this research project.

REFERENCES

BUTCHER, R. W. (1961). *A New Illustrated British Flora*, vol. 1. London.

CLAPHAM, A. R., TUTIN, T. G. & WARBURG, E. F. (1962). *Flora of the British Isles*, 2nd ed. Cambridge.

HALL, A. V. (1970). A computer-based system for forming identification keys. *Taxon*, **19**: 12–18.

MORSE, L. E. (1968). Construction of identification keys by computer. (Abstract). *Am. J. Bot.*, **55**: 737.

OSBORNE, D. V. (1963). Some aspects of the theory of dichotomous keys. *New Phytol.*, **62**: 144–160.

PANKHURST, R. J. (1970a). A computer program for generating diagnostic keys. *Comput. J.*, **13**: 145–151.

PANKHURST, R. J. (1970b). Key generation by computer. *Nature, Lond.*, **227**: 1269–1270.

PANKHURST, R. J. (1971). Computer-generated keys, in Exhibition Meeting, 1969. *Watsonia*, **8**: 336–337.

SHETLER, S. G., *et alia* (1971). Pilot data processing systems for floristic information, in CUTBILL, J. L., ed. *Advances in Data Processing for Biology and Geology*. London.

TUTIN, T. G., *et alia*, ed. (1968). *Flora Europaea*, vol. 2. Cambridge.

WEBSTER, T. (1969). Developments in the description of potato varieties, 1. Foliage. *J. natn. Inst. agric. Bot.*, **11**: 455–475.