COMPUTER IDENTIFICATION OF HARDWOOD SPECIES

by

R.G. Pearson* and E.A. Wheeler*

Summary

The data for the 4700 entries in the "Oxford card" multiple entry key for wood identification have been computerized. The format of the entries in the database and the procedure for doing wood identification using this database are described.

Introduction

During the preparation of Metcalfe and Chalk's (1950) classic treatise, "Anatomy of the Dicotyledons", Chalk and co-workers at Oxford University recorded wood anatomical data on perforated cards. The anatomical data assembled by Chalk are considered the property of the Department of Forestry of Oxford University.** Subsequently, L.H. Daugherty examined additional taxa for those features chosen as useful by Chalk and recorded his observations on the same type of card; these data have been distributed by F.W. Hankins. A detailed history of the Chalk key and "Oxford cards" is presented by Barefoot and Hankins (in press). The "Oxford cards" with approximately 4700 entries are the most comprehensive of the existing multiple entry keys for wood identification.

The major advantage of multiple entry keys over dichotomous keys for wood identification is the greater flexibility they allow in choosing the characters to be used. This flexibility is particularly valuable when studying fossil wood or recent decayed wood as many dichotomous keys use at one or more of the dichotomies features such as specific gravity and color that cannot be determined in such specimens. Also, dichotomous keys often use the presence or absence of a feature at a key step. In a petrified wood, the preservation may be such that the determination of the actual presence or absence of minute features, such as septate fibers or vestured pitting, is not possible. If specific gravity, color, or the presence or absence of a minute feature are used at the beginning of the dichotomous key, then that key cannot be used for fossil woods or decayed woods.

The disadvantages of multiple entry keys are mainly mechanical ones associated with the sorting of the perforated cards and have been discussed by Miller (1980). These disadvantages can be easily overcome if a computer of sufficient capacity is available. This paper describes a computer program for searching a master dataset which contains the anatomical information recorded on the "Oxford cards". The program devised for this database can also be used to search other databases with records which can be arranged like those in the "Oxford card" database.

Description of computer database

A database containing coded information on anatomical and a few other features (geographic source, density, distinctive color, and odor) has been formed from information on the "Oxford cards". There are 4724 entries, most of which are for different individual species. However, a single entry sometimes represents more than one species of a genus when those species are identical in the anatomical features used in the key. On the other hand, there are multiple entries for some individual species which exhibit variation in one or more key characters.

The records in the database are sorted in alphabetic order by family, genus, and species. If desired, the information for any particular species, genus, or family may be easily retrieved. Each entry contains the following items of information in this order: A number equal to the

* Department of Wood and Paper Science, North Carolina State University, Raleigh, North Carolina 27650, U.S.A. This is Paper No. 6523 of the Journal Series of the Agricultural Research Service, North Carolina.

** Information about the reproduction of the Chalk key may be obtained by writing to Dr. J. Burley, University of Oxford, Department of Forestry, South Parks Road, Oxford, OX1 3RB, United Kingdom.

number of letters in the generic name, a number equal to the number of letters in the specific name, three letters representing the family, a blank, the generic name, a blank, the specific name, a blank, a two digit number equal to the number of features listed as present in the species, and finally, the two digit code numbers in serial order for the features listed on the Oxford card for the species.

All numbers occupy two spaces, numbers less than ten being preceded by a blank. No other blanks, except those mentioned above, occur. The following is a typical entry.

15 7ACK TABERNAEMONTANA ARBOREA 11 2 3 61011121619212325

The perforated cards provide for a maximum of 82 features, some of which are mutually exclusive. The number of features listed as present for each entry ranges from 4 to 26, the mean being 14.

The database requires 23 tracks of storage on an IBM 3330 disk pack, the entries being of variable record length and arranged in blocks of 4240 bytes.

Description of the procedure for identifying an unknown species

The initial procedure is the same as that for using any multiple entry key and involves judgment in choosing the combination of features to be used. The absence of a feature may be regarded as significant as the presence of another. After examination of the sample, the code numbers of the significant features, as given on the Oxford key card, are listed and arranged in serial order. Each code number is followed by the numeral 1 if that feature is present or by the numeral zero if absent. Two numbers are added at the beginning of this list, the first being the minimum number of feature matches desired for identification, while the second is the total number of code numbers. A match is either the presence or absence of the same feature in the unknown species and in a known species from the dataset. Too low an estimate of the number of matches will lead to a large number of species names being printed out. It is recommended that for the first trial, the minimum number of matches should be set equal to, or one less than, the number of features listed. If none of the species in the database yield that number of matches, then the number may be reduced for a second trial.

This list of numbers, each of which must be separated from its neighbor by at least one blank or by a comma, forms the data on which the search program operates. An example is shown below:

12 13 2 0 6 1 11 1 21 1 23 1 28 1 30 1 33 1 36 1 39 1 43 1 44 1 61 1

Here 13 features are listed, all being visible in the wood sample except feature number 2. The leading number indicates that species from the database are to be listed if their features match any 12 of the 13 listed features.

If more than one species is to be identified during the same computer run, then the data for each species are to be entered according to the above format immediately after the end of the data for the preceding species. If not entered on a new line or card, a comma or blank should separate the two sets of data.

The program for searching the database is in PL/I. In load module form, it occupies four tracks (13030 bytes per track) of storage on an IBM 3330 disk pack. The search uses 70K bytes of core storage on an IBM 370/165 computer in addition to the disk storage for the load module and database.*

An example of the output is given in the Appendix. The number N is the serial number of the species in the database. The time for a search depends on the number of features to be matched and the number of different woods to be identified in the same job. The time for the search for the identity of the two woods in the appendix was 26 seconds using an IBM 370/165 computer. The total time for another job in which nine different woods were to be identified was 1 minute and 26 seconds. As in any identification procedure, it is still necessary to compare the unknown wood with verified wood samples and/or descriptions to determine if it and one of the species listed in the database are the same.

The search is not done interactively but as a batch job. Typical job control language and output is given in the Appendix. If desired, the output may be displayed on the screen of the searcher's terminal instead of, or in addition to, being printed.

Conclusion

The information contained on the "Oxford cards", a multiple entry key for wood identification devised by Chalk, is much easier to use in computerized form than in the original perforated card format. The manual sorting of

* The authors will be pleased to supply, at cost, a tape containing a copy of the database, the PL/I program, and the load module for the program. Their postal address is Post Office Box 5488, Raleigh, North Carolina 27650, U. S. A.

4700 cards is tedious and subject to error. Using the computer to do the sorting eliminates these problems, and has proved useful in working with both extant and fossil woods.

Acknowledgments

We wish to thank Mrs. E. Wilson for transcribing the information on the perforated cards to IBM cards.

References

Barefoot, A.C. & F.W. Hankins. Identification of modern and Tertiary woods. Clarendon Press, Oxford (in press).

Metcalfe, C.R. & L. Chalk. 1950. Anatomy of the Dicotyledons. Clarendon Press, Oxford.

Miller, R.B. 1980. Wood identification via computer. IAWA Bull. n.s. 1: 154–160.

Appendix:

Typical Input and Output – The job control language may vary with the particular computer installation, but will usually be similar to that given below. For purposes of illustration, the load module has been given the name SPECIES and contains the program named SEARCH. The data are for two pieces of wood, one with 13 features and the other with 30 features listed, most of which are absent. All 13 features are to be matched for the first piece, while at least 28 features are to be matched for the second piece of wood.

```
//jobname JOB account information, etc.
//JOBLIB DD DSN=SPECIES, DISP=SHR
//EXEC PGM=SEARCH
//MASTER DD DSN=database name, DISP=SHR
//SYSIN DD *
data for unknown species, e.g.
13 13 2 0 6 1 11 1 21 1 23 1 28 1 30 1 33 1 36 1 39 1 43 1 44 1 61 1
28 30 1 0 2 1 3 0 4 0 5 0 6 1 7 1 8 0 9 1 10 1 11 0 17 0 19 0 21 1 23 0 28 1 29 0 31 0 33 0 34 0 38
0 39 0 40 0 41 0 42 0 43 0 49 0 55 0 56 0 84 1
//
```

The output for the above job is reproduced below:

```
                *** RESULTS OF SEARCH FOR IDENTITY OF UNKNOWN SPECIES ***

NOTE:
  DATA CARD FOR UNKNOWN SPECIES MUST HAVE THE FOLLOWING NUMBERS IN THE
  ORDER GIVEN BELOW:–
      N1 = MINIMUM NO. OF MATCHES REQUIRED,
      N2 = NO. OF FEATURES LISTED ON DATA CARD,
      #(1) THROUGH #(N2) = CODE NUMBERS OF THE N2 FEATURES,
      THE FEATURE NUMBERS MUST BE IN SERIAL ORDER.
      EACH FEATURE IS TO BE FOLLOWED BY THE NUMERAL 1 IF PRESENT,
      OR BY THE NUMERAL 0 IF ABSENT.
      PUT A COMMA AND/OR AT LEAST ONE BLANK BETWEEN EACH NUMBER.

UNKNOWN SPECIES No. 1 : DATA CARD GIVES FOLLOWING INFORMATION –
13 13  2 0  6 1 11 1 21 1 23 1 28 1 30 1 33 1 36 1 39 1 43 1 44 1 61 1

LIST OF SPECIES POSSIBLY MATCHING THE UNKNOWN SPECIES:–

  N =  1399  BDG  SLOANEA AUSTRALIS:
       FEATURES  6 11 21 23 28 30 33 35 36 39 43 44 61 77

       COMPLETION OF SEARCH FOR UNKNOWN SPECIES NO. 1
```

UNKNOWN SPECIES NO. 2 : DATA CARD GIVES FOLLOWING INFORMATION –
28 30 1 0 2 1 3 0 4 0 5 0 6 1 7 1 8 0 9 1 10 0 11 0 17 0 19 0 21 1 23 0 28 1 29 0 31 0 33 0 34 0 38 0 39 0 40 0 41 0 42 0 43 0 49 0 55 0 56 0 84 1

LIST OF SPECIES POSSIBLY MATCHING THE UNKNOWN SPECIES:–

N = 481 ADN CARPINUS BETULUS:
FEATURES 2 6 9 21 25 32 45 46 63 68 73 74 84

N = 489 ADN OSTRYA SPP:
FEATURES 2 3 6 7 9 14 21 44 45 46 61 63 68 74 80 84

N = 3640 DFG PRUNUS AVIUM:
FEATURES 2 6 9 16 20 21 25 30 46 50 59 64 68 73 74 84 86

COMPLETION OF SEARCH FOR UNKNOWN SPECIES NO. 2

*** ALL SEARCHES COMPLETED ***