Merlewood Research and Development Paper
Number 35

PLOTTING OF MULTIDIMENSIONAL DATA

J. N. R. Jeffers, FIS, AMBIM

R. & D. 72/35

Merlewood Research and Development Papers are produced for the dissemination of information within the Institute of Terrestrial Ecology. They should not be quoted without preliminary reference to the author. All opinions expressed in Merlewood Research and Development Papers are those of the author, and must not be taken as the official opinion of the Institute of Terrestrial Ecology.

Introduction

The plotting of highly multidimensional data has long been a practical problem in the interpretation of multivariate analysis. Methods such as principal component analysis and canonical variate analysis have the important properties of reducing the dimensions of the total variability to the smallest possible number consistent with little or no loss of essential information. In some cases, the interpretation of the results is possible from consideration of the distribution of the original observations on a plane, or, with the aid of stereoscopic equipment, isometric diagrams, or solid models, from consideration of distributions in three dimensions. Many practical problems, however, require interpretation of the distribution of the observations in more than three dimensions, even after principal component analysis or canonical variate analysis. Various suggestions have been made for the use of different symbols on a two-dimensional plot to give some idea of the effects of a third or fourth dimension, but such methods lack precision and cannot be readily extended to large numbers of dimensions.

A recent paper (Andrews, 1972) has suggested the embedding of highly-dimensional data in a higher dimensional, but easily visualized, space of functions and then plotting the functions. The suggestion is so obviously sensible and so simple in concept that it seems astonishing that it has not been developed before. This paper presents the idea in a simple form, gives a BASIC program for the plotting of multidimensional data on an ordinary teleprinter, and applies the technique to a problem of the morphometrics of species of birch.

Method

The method proposed by Andrews (1972) is very simple. For each point the function:

$$f_x(t) = x_1/\sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \ldots.$$

is defined and the function plotted over the range $-\pi < t < \pi$. This function transforms a set of points into a set of lines drawn across the plot in such a way that the mean of the functions of n observations corresponds to the mean of the observations themselves. The function representation also preserves distances, so that the distance between two functions accords with the distance as judged by the human eye, and this distance is proportional to the Euclidean distance between the corresponding points. Even more important, the presentation preserves variances, so that, if the components of the data are uncorrelated, with a common variance σ^2, then the function value at t, $f_x(t)$, has a variance which is given by:

$$\mathrm{var}\,\left[\, f_x(t) = \sigma^2 \left(\tfrac{1}{2} + \sin^2 t + \cos^2 t + \sin^2 2t + \sin^2 2t + \cos^2 2t + \ldots.\right) \,\right]$$

The variability of the plotted function is almost constant across the graph, a fact which considerably simplifies the interpretation of the plotted functions.

Table 1 gives a BASIC program for the calculation of the values of the function, and the plotting of these values on the kind of teleprinter usually provided for a remote-access, time-sharing system. The program assumes the existence of a disk file containing the calculated values of the principal components or canonical variates, the first variable being associated with the first component or canonical variate, the second variable with the second component, etc. The program can, however, be easily modified to operate as subroutine working on a data matrix held in core. The function is plotted for up to 30 sets with values of t from -3.0 to +3.0 in steps of 0.5. It is, therefore, reasonably fast and suitable for operation from an interactive terminal, the symbols, which may be freely selected from those available on the keyboard, being joined by straight lines.

Only a limited amount of information can be absorbed conveniently from one plot, and experience suggests that about ten functions is the maximum number that can be considered in detail at any one time, unless colours are used to distinguish a priori groupings. Preliminary grouping of the sets by cluster analysis may, therefore, be helpful in reducing the number of sets to be considered, with separate plots to examine within-group variation. The technique is particularly valuable in detecting outliers to the main set of observations.

An Example

Gardiner and Jeffers (1962) analysed measurements of 13 leaf characters for species of Betula, and calculated five components accounting for 99 per cent of the total variation described by the 13 variables. The calculated values of these components are given in Table 2.

```
LIST
   5 REM PROGRAM TO PLOT MULTIDIMENSIONAL DATA
  10 RECORD X(29)
  15 OPEN 8,"DATA"
  20 PRINT "NO OF VARIABLES AND SETS";
  25 INPUT N,M
  30 DIM A$(29)
  35 FOR I=1 TO M
  40 INPUT A$(I)
  45 NEXT I
  50 FOR J=-3.0 TO 3.0 STEP 0.5
  55 FOR I=1 TO M
  60 GET 8,10,I
  65 LET I=I-1
  70 LET K=0
  72 LET T=0
  75 LET S=X(K)/(SQR(2))
  80 LET K=K+1
  85 LET T=T+1
  90 IF K=N THEN 120
  95 LET S=S+X(K)*SIN(T*J)
 100 LET K=K+1
 105 IF K=N THEN 120
 110 LET S=S+X(K)*COS(T*J)
 115 GO TO 80
 120 LET S=INT(S*5+40)
 125 PRINT TAB(S);A$(I);
 130 NEXT I
 134 FOR L=1 TO 5
 135 PRINT
 136 NEXT L
 140 NEXT J
 145 STOP
 150 END

READY
```

TABLE 1   BASIC PROGRAM FOR THE PLOTTING OF MULTIDIMENSIONAL DATA.

Table 2. Calculated values of five components of leaf measurements of Betula species

Taxa		Value of component:-				
A.	B. verrucosa	1.95	0.63	-0.36	-1.89	0.17
B.	B. verrucosa	2.00	0.14	-0.55	-1.53	0.08
C.	B. pubescens	1.12	-0.14	0.02	-0.21	0.27
D.	B. pubescens	1.37	-0.23	0.05	-0.34	0.28
E.	B. tortuosa	0.30	0.25	0.07	1.00	-0.11
F.	B. carpatica	0.50	0.31	0.09	-0.62	-0.43
G.	B. oycoviensis	-1.12	0.59	0.44	0.69	-0.05
H.	B. obscura	1.67	0.12	-0.45	-1.38	0.07
I.	B. nana	-5.91	-0.23	-0.77	2.17	0.19
J.	B. humilis	-1.88	-1.46	1.46	2.10	-0.47

The functions for B. verrucosa, B. pubescens, and B. nana are plotted in Figure 1. Clearly, all three species are easily distinguishable, B nana having values consistently less than those for the other two species, and the patterns for B. verrucosa and B. pubescens being distinct. In Figure 2, the functions for B. obscura and for B. tortuosa and B. oycoviensis have been added. B. oscura matches the functions for B. verrucosa almost exactly and is certainly not a distinct species, on the basis of leaf characters at least. B. tortuosa and B. oycoviensis have very similar function shapes, and probably represent a single species, but the species is distinct from B. verrucosa and B. pubescens. Finally, in Figure 3, the functions for B. carpatica and B. humilis have been added. B. carpatica takes values intermediate between B. tortuosa and B. verrucosa along almost the whole range of $-\pi < t < \pi$, and it seems likely that it represents a hybrid between these two latter species. Similarly, the intermediate position of B. humilis between B. nana, the dwarf birch, and the collective species B. alba suggests that it is a hybrid.

The functions plotted in Figure 3 have the added advantage of indicating the optimum discriminator between species. In Figure 3, when t is approximately equal to 2.0, the value of the computed function enables the species to be identified with a fair degree of certainty, and no other value of t enables such good discrimination to be made. At a value of t = 2.5, B. humilis cannot be distinguished from B. nana, or B. carpatica from B. pubescens. However, the best discriminator between B. nana and all other species is at t = 0.5, and hybrids between B. nana and the collective species B. alba might be best identified by values of the function at t = 1.5.

__Figure 1.__ __Functions for B. verrucosa, B. pubescens, and B. nana__

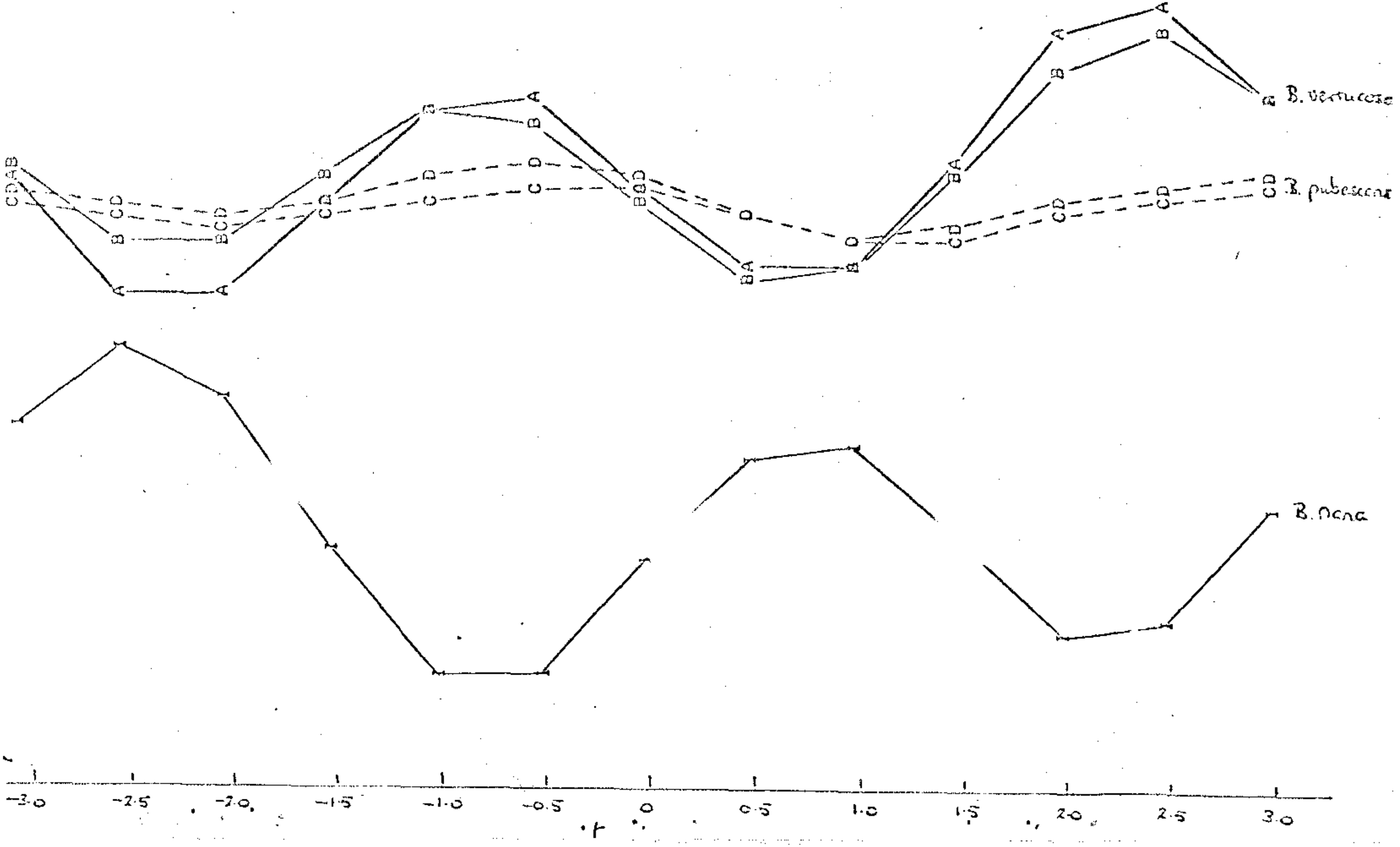


Figure 2. Functions for B. verrucosa, B. pubescens, and B. nana, with addition of B. obscura, B. tortuosa, and B. oycoviensis

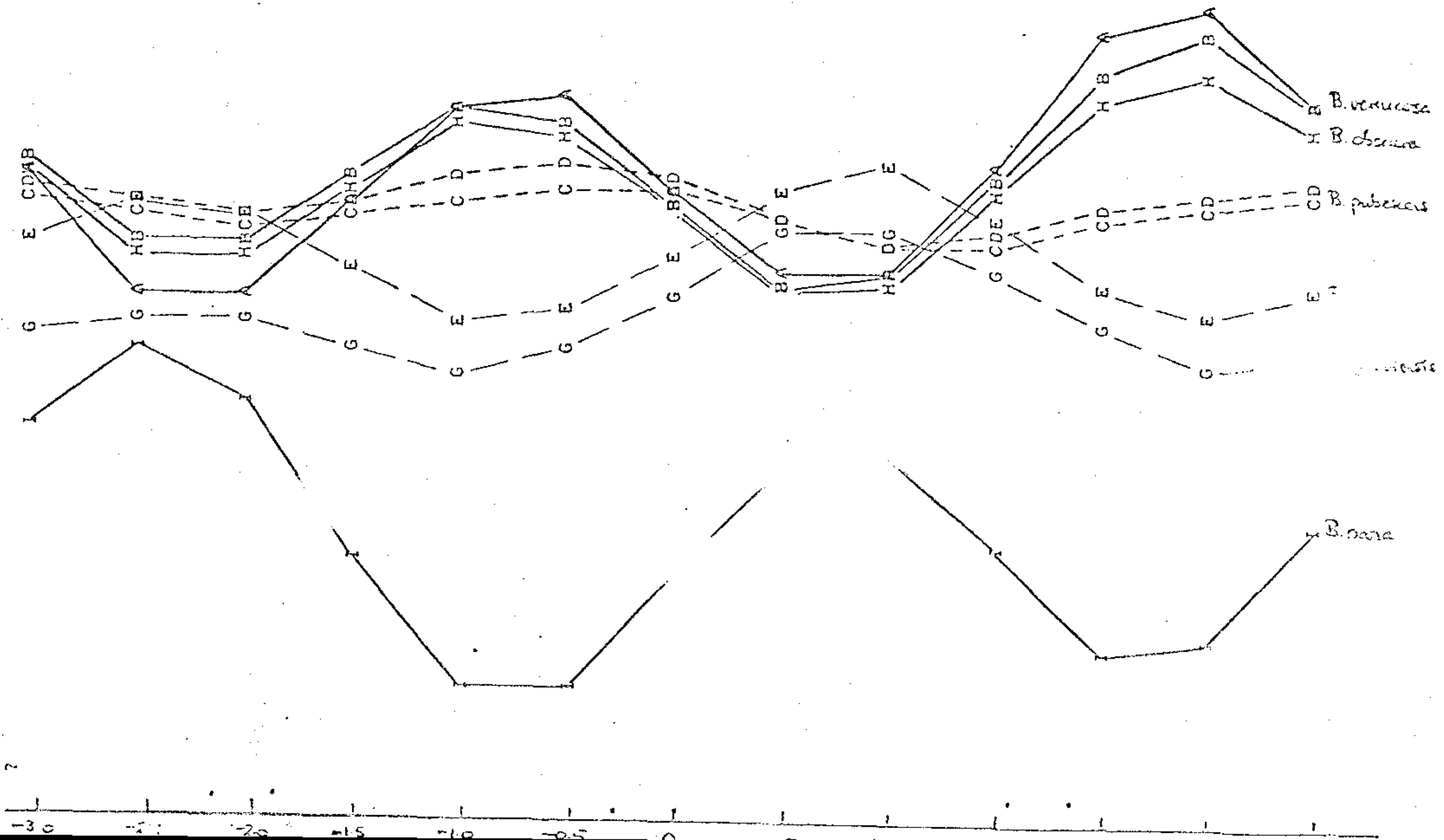
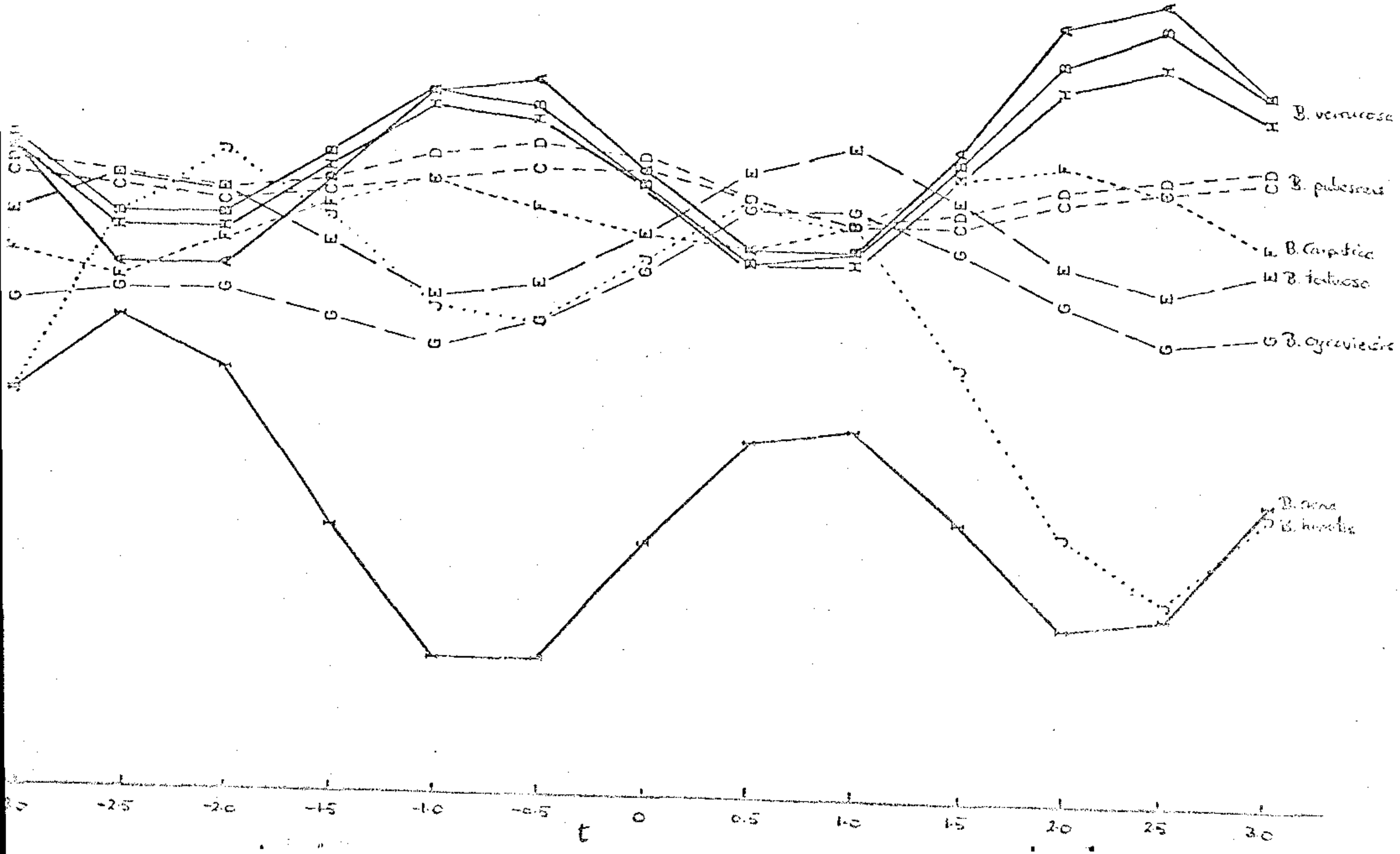


Figure 3. Plotted functions for all species



READY

References

D. F. Andrews, 1972. Plots of high-dimensional data. Biometrics 28 (1) 125-36.

A. S. Gardiner and J. N. R. Jeffers, 1962. Analysis of the collective species Betula alba L. on the basis of leaf measurements. Silvae Genetica 11 (5/6) 156-61.