4-1989

Pattern analysis : a descriptive introduction

J. F. Wallace
Department of Agriculture and Food, Western Australia

Australian Co-operation with the National Agricultural Research Project, Thailand.

AUSTRALIAN CO-OPERATION

WITH THE

NATIONAL AGRICULTURAL RESEARCH PROJECT

THAILAND

PATTERN ANALYSIS:
A DESCRIPTIVE INTRODUCTION

Notes for a workshop on pattern analysis of germplasm data; to be held at the Pathum Thani Rice Research Centre, May-June 1989.

J.F.WALLACE

Western Australian Department of Agricult

April 1989

TABLE OF CONTENTS

1. Introduction

Pattern analysis is a term which covers a very large area of statistical analysis for multivariate data sets. A full understanding of pattern analysis would require years of study of mathematical concepts and computing algorithms. Fortunately, pattern analysis methods can now be used without all this specialist knowledge. This is possible because of the availability of computer programs such as PATN, which is a powerful interactive program running on micro computers. However,such powerful programs are easily misused. If researchers are to make good use of such programs, they should have some understanding of the concepts of pattern analysis, and the options available.

The purpose of these notes is to provide this basic descriptive understanding of the mathematical routines used in pattern analysis. We hope that this will help people to decide when pattern analysis might be useful, and how to interpret the results. Mathematical formulas are avoided as far as possible. However, pattern analysis has a great many 'special words' with specific meanings. Use of these terms is unavoidable. In a way, pattern analysis begins where normal statistical methods end. Some knowledge of basic statistics is assumed.

These notes are far from complete. They are to be used in conjunction with lectures, and hands-on demonstrations of the PATN program on users's data sets. For a fuller coverage of the subject, some of the references should be consulted.

2.0 What is pattern analysis ?

Pattern analysis is a general term to cover methods for analysing multivariate data sets when the aim of the analysis is to discover structure in the data. The methods attempt to analyse all the variables at the same time to give the best information on clustering and separations of individuals. Pattern analysis includes routines for

CLUSTER Analysis - methods for finding group structure in the set of individuals

ORDINATION Methods - optimal methods for the display and assessment of the closeness and distance between individuals in the multivariate data

NETWORK algorithms - methods to display similarity between pairs of individuals

There are several different methods available within these three main analysis types. For example, cluster analysis may begin by grouping close individuals together one at a time; or it may begin with a single large group and attempt to split the group into subclusters in some optimal way. The first method is known as AGGLOMERATIVE clustering, and the second as DIVISIVE.

2.1 Why use pattern analysis ?

Pattern analysis has developed in response to the need to make the best use of particular kinds of complex data sets. The type of data we are talking about are multivariate data where many different measurements are taken on a number of individuals. In our application, the individuals are genetic accessions. We hope to use the data observations to tell us which accessions may be grouped as similar and which are different. Within these 'type' groups we may wish to look at the variation of important characters. We may also wish to summarise the differences and similarities between the type groups. Pattern Analysis is used in problems when 'normal' statistical methods do not apply to the problem (see below).

Multivariate data is really multi-dimensional data. The human mind cannot effectively picture space beyond three dimensions. For this reason, the full information in multivariate data sets just cannot be gained from the raw data.

We can display a single measured variable (or 'attribute') on UNIVARIATE plot - a histogram. Two variables can be looked at together on BIVARIATE plot (also called a SCATTERGRAM). These methods are useful, but mathematical methods must be used to simplify the data before we can look at higher dimensional data. This then is the major aim of pattern analysis; to simplify complex data sets so that we can interpret the relationships between the individuals; and to simplify the data in such a way that the maximum information is kept. Pattern analysis can make use of all the data and are not affected by personal judgement of important types or characters.

2.3 Pattern analysis and statistics

It is useful to compare pattern analysis with the well known statistical methods used in designed experiments. These two approaches are applied in quite different situations; the objectives, the types of data sets, the assumptions, and the interpretation are all different.

Consider, for example a designed experiment to compare two varieties for yield performance. We take repeated measurements for each variety (group) on plots; We know in advance which groups the measurements belong to; and we have a well defined Hypothesis, (Null Hypothesis: the varieties are equal). We also make assumptions about the distribution of the data, and can then calculate a Significance Level for the hypothesis (i.e. a probability value).

In a pattern analysis situation almost everything is different. We know in advance that all the individuals are different (accessions). Therefore there is no Null Hypothesis. We do not have repeated measures on the same population. We have many different types of data observations, and we make no assumptions about the data distributions. We do not test hypotheses, and the idea of a significance test is not relevant. We want the data to tell us about the individuals, and the relationship between individuals. Later, we may want to explain the groupings we discover. Pattern analysis is driven by the data, not by our prior grouping or design. It is used for data understanding and for generating hypotheses, rather than for testing them.

Pattern analysis thus raises questions or ideas, rather than giving an answer to a specific question. There is a large element of 'art' or judgement in the use of pattern analysis. The problem with pattern analysis is that there is no single answer. Different methods give different pictures of the data. This is not surprising when the complexity of the data sets is considered. Pattern analysis is best used when several different results are compared and combined for interpretation by people with knowledge of the original data.

3.0 A Multivariate data set

We will introduce here a sample data set, and describe the general form of the data form expected by PATN and other mutivariate programs. We use the words

INDIVIDUALS - the units on which the different data observations are taken.

ATTRIBUTES - the separate items of information which are recorded for each Individual. The words 'variable' or 'measurement' may be familiar. Attribute is used as a more general term.

DATA MATRIX - The two dimensional array of data; the set of recorded Attribute values on each of the individuals.

We have a set of different attribute measurements on each individual; We can also think of this data as a set of individual measurements for each attribute. It is convenient to represent this data in a rectangular array; usually the ROWS of the DATA MATRIX represent the INDIVIDUALS, and each COLUMN stores the values for an ATTRIBUTE. However, we could just as well change the matrix around and represent the attributes in rows. Some PATN procedures do this if we want to.

3.1 The Pathum Thani rice accession data

The example data set is a group of native rice accessions which were evaluated at Pathum Thani Rice Research Centre. There are 46 different accessions ('Individuals'); 39 different plant characteristics are recorded for each accession, so we have 39 'Attributes'. The standard data matrix has 46 rows and 39 data columns in the form:

		column1	col2	col3	col39
		BLPB	BLCO	LSCO	SNT
row1	2278 BANG GAWK	3	3	1	0
row2	3442 BAHN CHAWNG	3	3	1	0
---		.	.	.	.
---		.	.	.	.
row46	15733BANG GAWK	3	2	1	0

The full data matrix is printed separately. We want to use all the information in the attributes to tell us about the accessions. This data set has 39 attribute dimensions. The attributes are named and listed here.

ATTRIBUTE CODE	DESCRIPTION	RECORDED AS
1.BLPB	blade pubescence	1,2,3 (X for mixture)
2.BLCO	blade colour	1-7,X for colours
3.LSCO	basal leafsheath col.	1-4,X
4.LA	leaf angle	1-9,X
5.LCO	ligule colour	0-3,X
6.LSH	ligule shape	0-3,X
7.CCO	collar colour	1-3,X
8.AUCO	auricle colour	0-2,X
9.HDG	days seed - 50%head	whole number of days
10.CUA	culm angle	1-9,X
11.INCO	internode colour	1-4,X
12.CUST	culm strength	1-9
13.FLA	flag leaf angle	1-7,X
14.PTY	panicle type	1-9,X
15.SBR	secondary branching	0-3,X
16.PEX	panicle exsertion	1-9,X
17.PAX	panicle axis	1,2,X
18.AWPR	awning	1-9,X
19.AWCO	awn colour	1-6,X
20.APCO	apiculus colour	1-7,X
21.STCO	stigma colour	1-5,X
22.SLCO	sterile lemma colour	1-4,X
23.VG	variety group	1,2,3,4 different groups
24.LLT	leaf length	cm (av of 5 leaves)
25.LWDT	leaf width	cm
26.LGLT	ligule length	mm (* for none)
27.CULT	culm length	cm
28.CUNO	culm number	count
29.CUDI	culm diameter	mm
30.PSH	panicle shattering	1-9 (1=<1%,9=>50%)
31.SEN	leaf senescence	1-9,X
32.SPKF	spikelet fertility	1-9,X
33.PTH	panicle threshability	1-9,X
34.LPCO	lemma & palea colour	0-9,w,X
35.LPP	lemma & palea pub.	1-5,X
36.PLT	panicle length	cm
37.ENDO	endosperm type	1,2,3,X
38.SLLT	sterile lemma length	1-9,X
39.SNT	scent	0,1,2

Attributes 1-8 from vegetative stage, 9-29 from reproductive stage, 30-39 from harvest/post-harvest stage.

It is impossible to consider all these 39 dimensions without the use of some multivariate analysis. This is a typical Pattern Analysis type data set; the attributes are of many different TYPES, measured with different SCALES and ACCURACY. It is important to understand the attribute data set before using the multivariate computer programs.

3.2 Data inspection and information

We can also see that some attributes carry more INFORMATION than others. Four of the attributes in the rice data carry no group information at all, because all the values are the same; they are LSH(all 2), AWPR(all zero), AWCO(no values), SNT(all zero). Other attributes appear to carry little information; for example the first attribute (BLPB) has a recorded value of 3 for all except two of the individuals.

There are mathematical methods for calculating INFORMATION in attributes (see Williams). Easy univariate (HISTOGRAM) and bivariate (SCATTERGRAM) displays can help us assess the different attribute types. These displays are also useful to illustrate what the multivariate routines are doing in higher dimensions.

3.3 Attribute types

It is most important to recognise the different types of data which are recorded; - the way the data can be used is determined by this. Remember, our basic objective is to use the data to measure the similarity between individuals. The main data types are listed below with brief notes; the attributes in the sample data should be discussed in the workshop; for more information and written definitions see (e.g.) Williams,ch 5.

NUMERIC (or QUANTITATIVE, or CONTINUOUS) attributes.
Examples: length, weight, periods of time
The measurements of these data may vary continuously. An important feature is that the scale is constant; e.g. the difference between 1cm and 2cm, is the same as the difference between 101cm and 102cm. In some cases we may wish to transform the data before using it in this way. It is sensible to measure the 'closeness' of individuals by this difference of attribute values.

NOMINAL (or CATEGORICAL) data.
Examples: colour, variety group, leaf shape, soil type
The attribute data record DISCRETE states; the values may be recorded as numbers, but there is no order of states.
Example: for STCO, numbers 1-5 represent different colours but it does not make sense to say that Purple(5) is greater than White(1). Two individuals may either have the same, or different values.
An important type of nominal attribute is referred to as BINARY (or 2-STATE, YES/NO, PRESENCE/ABSENCE, 0/1, DICHOTOMOUS). Only two classes are possible. For comparing individuals, the scientist should decide whether the attribute is symmetric or not.

ORDINAL Attributes
Examples: Culm strength rating, panicle shattering rating, etc
The data are recorded as discrete states which represent some order, e.g. 1,2,3 for early,medium,late. It makes sense to say that 1<2<3. However, the scale is not constant, and these types of attributes are difficult to handle when comparing individuals. Consider four individuals with ratings 1,1,2,3 for example. How can we calculate the relative similarity of these? Unfortunately, this type of attribute is very common in germplasm data, so we must use it somehow.

COUNT data are always only whole numbers. In many cases they may be considered like numerical continuous data. In other cases they are better considered as categorical.

3.4 Univariate Histograms, Bivariate Scatter plots

In this section we will look at graphical displays of the attribute values. These are to be discussed in the workshop sessions. The idea of group structure, METRIC distance, and ORDINATION are to be introduced here. The different two-dimensional plots illustrate the complexity of the multivariate problem.

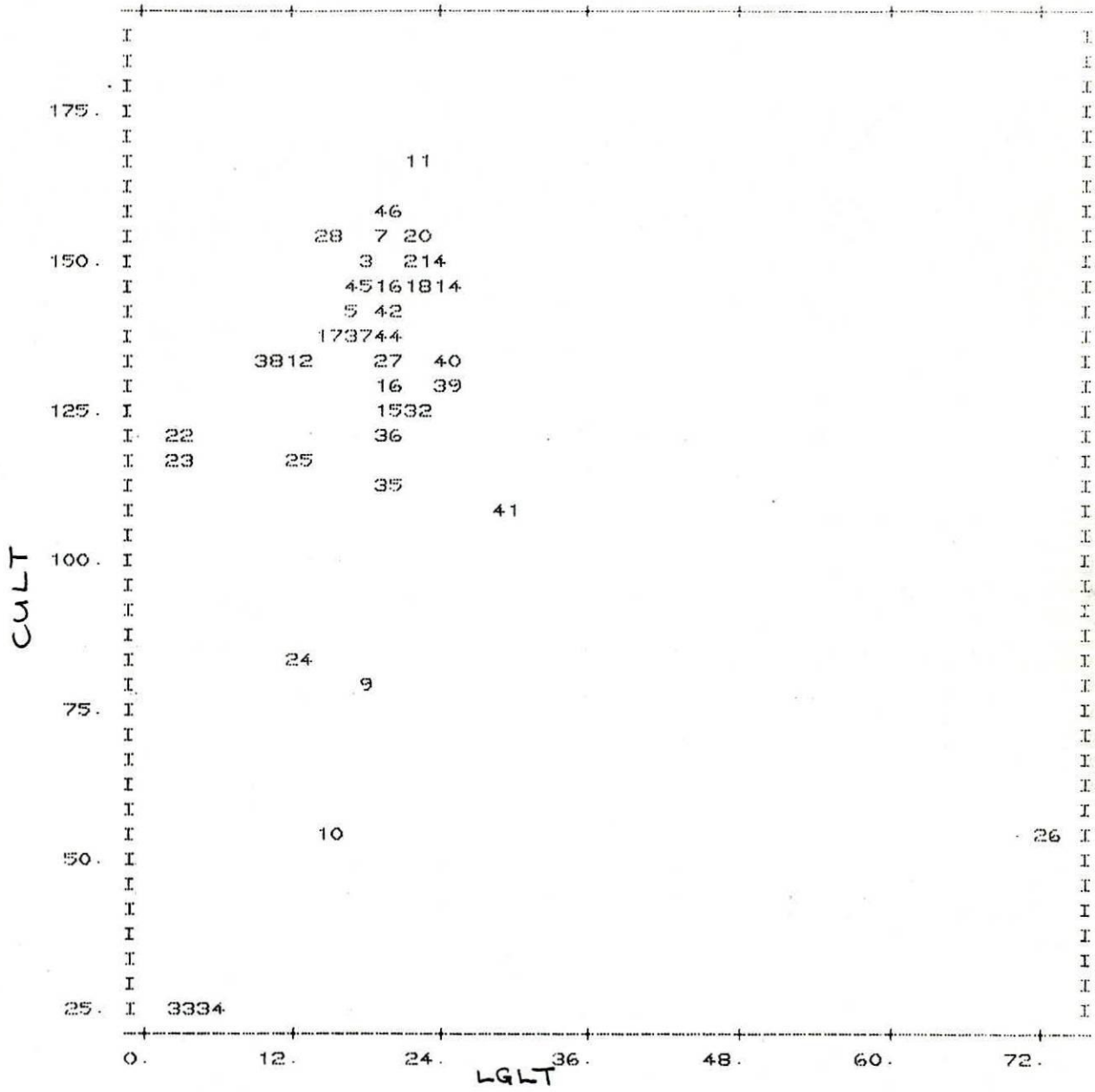
PATN HISTOGRAM OUTPUT

```
 LABEL = col00007          2.700    =>   6.550      4******
                           6.550    =>   10.40      1**
Number..    46.00          10.40    =>   14.25      5********
Minimum.    2.700          14.25    =>   18.10     11******************
1stQuart    15.30          18.10    =>   21.95     15************************
Median..    19.00          21.95    =>   25.80      8*************
Mean....    18.59          25.80    =>   29.65      1**
3rdQuart    21.00          29.65    =>   33.50      0
Maximum.    72.00          33.50    =>   37.35      0
Q3-Q1...    5.700          37.35    =>   41.20      0
Av.Dev..    5.014          41.20    =>   45.05      0
Std.Dev.    9.846          45.05    =>   48.90      0
Variance    96.95          48.90    =>   52.75      0
Range...    69.30          52.75    =>   56.60      0
Sum.....    855.0          56.60    =>   60.45      0
Num > 0.    46.00          60.45    =>   64.30      0
Skewness    3.200          64.30    =>   68.15      0
Kurtosis    16.84          68.15    =>   72.00      1**


 LABEL = col00008          25.40    =>   33.21      2****
                           33.21    =>   41.02      0
Number..    46.00          41.02    =>   48.83      0
Minimum.    25.40          48.83    =>   56.64      2****
1stQuart    122.0          56.64    =>   64.46      0
Median..    136.0          64.46    =>   72.27      0
Mean....    127.7          72.27    =>   80.08      1**
3rdQuart    147.0          80.08    =>   87.89      1**
Maximum.    166.0          87.89    =>   95.70      0
Q3-Q1...    25.00          95.70    =>   103.5      0
Av.Dev..    22.40          103.5    =>   111.3      1**
Std.Dev.    32.66          111.3    =>   119.1      3*******
Variance    1067.          119.1    =>   126.9      4*********
Range...    140.6          126.9    =>   134.8      4*********
Sum.....    5875.          134.8    =>   142.6     11************************
Num > 0.    46.00          142.6    =>   150.4      7***************
Skewness   -1.814          150.4    =>   158.2      9********************
Kurtosis    2.705          158.2    =>   166.0      1**


 LABEL = col00009          2.000    =>   3.944      3****
                           3.944    =>   5.889      1*
Number..    46.00          5.889    =>   7.833      5*******
Minimum.    2.000          7.833    =>   9.778     18************************
1stQuart    8.000          9.778    =>   11.72     11***************
Median..    9.000          11.72    =>   13.67      5*******
Mean....    9.652          13.67    =>   15.61      1*
3rdQuart    11.00          15.61    =>   17.56      1*
Maximum.    37.00          17.56    =>   19.50      0
Q3-Q1...    3.000          19.50    =>   21.44      0
```

TWO DIMENSIONAL SCATTER PLOT (GENSTAT)



D[8] v. D[7] using factor ACC

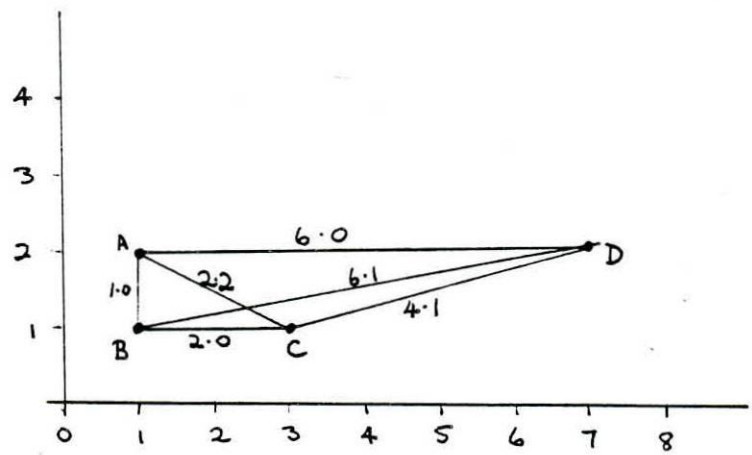
Points coinciding with 6
8

4.0 Measures of Similarity (Association) between Individuals

A basic process in pattern analysis is the calculation of measures of SIMILARITY between individuals using the attribute values. In fact, it is more common to calculate DISSIMILARITY values. To save confusion the the word ASSOCIATION is used. The index is calculated for all pairs of individuals, 1vs2, 1vs3, 2vs3 etc. We may think of the Dissimilarity value as a distance - if two individuals are similar, the index value will be low, if they are very different ('far apart in data space') then the value will be high. It is usual to scale the indices to values between 0 and 1, and to store the values in a diagonal ASSOCIATION MATRIX (also called a SIMILARITY matrix).

There are many different ways to calculate a dissimilarity index. Different methods are sensible for different types of data. One way is to calculate normal Euclidean distance between pairs of points. This method is not sensible at all for Nominal or most Ordinal type data. Distance may not be sensible for Continuous data either, unless we SCALE or TRANSFORM the data first. However, we present here a simple example using the distance as a measure of Dissimilarity.

Example. Consider 4 individuals, with attribute values in two dimensions as plotted. The distances between the points A,B,C & D are shown.



The ASSOCIATION MATRIX using this distance measure would look like this:

A	0			
B	1.0	0		
C	2.2	2.0	0	
D	6.0	6.1	4.1	0
	A	B	C	D

In fact, the zeros on the diagonal are not usually recorded, and the values are scaled in the range 0-1; if we scaled by dividing by 6.1 (the largest value), we see the matrix as

```
0.164
0.361  0.327
0.984  1.000  0.683
```

This how PATN would store the ASSOCIATION MATRIX.

Note: the size of the association matrix depends on the number of Individuals. For our 46 accessions, the matrix will have 45 rows and 45 columns.

4.1 Exercise: Clustering and Dendrograms from the Association Matrix.

Once we decide how to measure the association between individuals, the ASSOCIATION MATRIX contains all the information we need to perform one type of CLUSTERing procedure. The results can be represented in a figure called a DENDROGRAM. The process is an example of HIERARCHICAL AGGLOMERATIVE CLUSTERING. For the simple example above we can see how it works. In many dimensions, this is impossible; and we need to look at the Dendrogram after the analysis.

The process begins by looking at all the association values. The two 'closest' individuals are combined to form a group. Now, there is one more decision to make. We must calculate an association ('distance') between the new group and all the other individuals; i.e a new Matrix. How should we do this? Use NEAREST NEIGHBOUR first. The smallest value in the new matrix is used to group the next pair of 'Individuals'; and the process is repeated.

Discuss how this works for the example above. At each step, recalculate the association matrix using raw distances. Note the effect of using strategies other than nearest neighbour. Draw the DENDROGRAM which results.

4.2 More on association measures.

4.2.1 Numerical data

We mention above that distance is only one type of measure, and it is only useful for numerical data in some circumstances. Distance values based on raw data are affected by the scale of measurement for the different attributes, and will not treat the different attributes equally. (Consider measurements in mm & m of two attributes). The raw values of all attributes should be STANDARDISED by SCALING before using a distance measure. Prior to this the data may need to be TRANSFORMED as the distance measure may exaggerate the effect of outliers. In classical statistics it

is popular to scale by the standard deviation of the values. However this is not generally optimal for pattern analysis data, and standardisation by range is recommended for EUCLIDEAN distance.

Euclidean distance itself is not so popular because it uses squares of values which tends to increase the distance of outliers. Other types of distance METRIC are used instead:

e.g. MANHATTAN METRIC - sum of absolute values in attribute dims
GOWER METRIC - a range standardised Manhattan metric
BRAY-CURTIS - a different standardisation
CANBERRA METRIC - a different standardisation
(see Williams, ch6 for more details)

Transformation of attribute values will affect the weightings of parts of the scale. This is often useful. Examples are log transforms, transforms to rank order, or transform to 0/1 data at some threshold.

4.2.2 Categorical data

For binary (Presence/absence) data, matching coefficients are calculated between individuals. There are different methods here also, depending on how the similarity is to be calculated; the important decision is whether the distance measure is SYMMETRIC or not (i.e. if 0,0 match = 1,1 match).

Nominal, and ordinal data types present problems for any metric.

4.2.3 In PATN

PATN has a choice of association metrics, most of the metrics will scale the data for the user. The program uses the fact that metrics for numerical data also give suitable information for most categorical data. See the PATN manual, or references for more information.

The user can get PATN to do the work; for different data sets, different metrics can be tried. PATN recommends BRAY-CURTIS in most situations, or GOWER metric. The user should consider whether the data should be transformed before analysis.

5.0 Cluster analysis

There are 3 major strategies for forming clusters from the multivariate data. All are based on the association matrix.

The simple example above gives the idea of how one popular type of cluster analysis works. It was hierarchical agglomerative clustering based on an association matrix of individuals.

HIERARCHICAL means that the clustering is one-way in the sense that once a group is formed, members may be added but not taken from it and added to another group.

AGGLOMERATIVE means that individuals are brought INTO groups, one at a time.

At each clustering point, the individuals which are combined differ by a known value in the association matrix. As the process goes on, the level of 'difference' of the merged pairs increases. A DENDROGRAM is a graphical representation of the groupings and the level at which they occur. The user may decide at what level the groups of interest should be formed.

It should be clear that the groups of individuals which are formed depend on;

1) the attributes used in calculating the associations and any transformation applied
2) the metric used to measure the associations
3) the strategy for calculating the association between new groups and other individuals or groups.

There are other algorithms for DIVISIVE cluster definition. Here we start with a single group and split it into 2 subgroups according to some strategy. The splitting strategy is based on some association measure of all possible subgroup pairs - according to the chosen metric and strategy the two 'most different' subgroups are identified. The measured distance between gives a value to this difference. The subgroups are then split, and the process repeated until the user chooses, or all the individuals are separated. The divisive clustering is HIERARCHICAL if the groups are split successively, i.e. if individuals are separated at any stage, they can never be put in the same group. Dendrograms can be used to display such a process.

The third strategy for clustering is called ALLOCATION. Classes are defined in the first place using SEED individuals. The algorithm then allocates an individual to the closest class, or if it is 'too different' the new individual is made the SEED of a new class. This strategy is fast for a computer, and simple. In PATN, the user can alter the number of seeds, and the threshold for defining new classes. The author of PATN, Lee Belbin, recommends this strategy for data exploration.

6.0 Ordination Methods

We use ORDINATION procedures to help us see what is happening in the multivariate data space. A Scattergram is an example of a two dimensional ordination plot. In fact, if there are only 2 attribute dimensions, the scattergram gives a perfect ordination of the data - the exact location of each individual can be seen.

We cannot see data in multi-dimensional space. Ordination tries to find a simple representation of the space in a few dimensions; and find it so that the maximum amount of information is retained. We can then look at scores or plots in the reduced space, and (we hope) understand what is happening. Many methods of ordination are known, including

PRINCIPAL COMPONENTS ANALYSIS
PRINCIPAL COORDINATE ANALYSIS
MULTIDIMENSIONAL SCALING

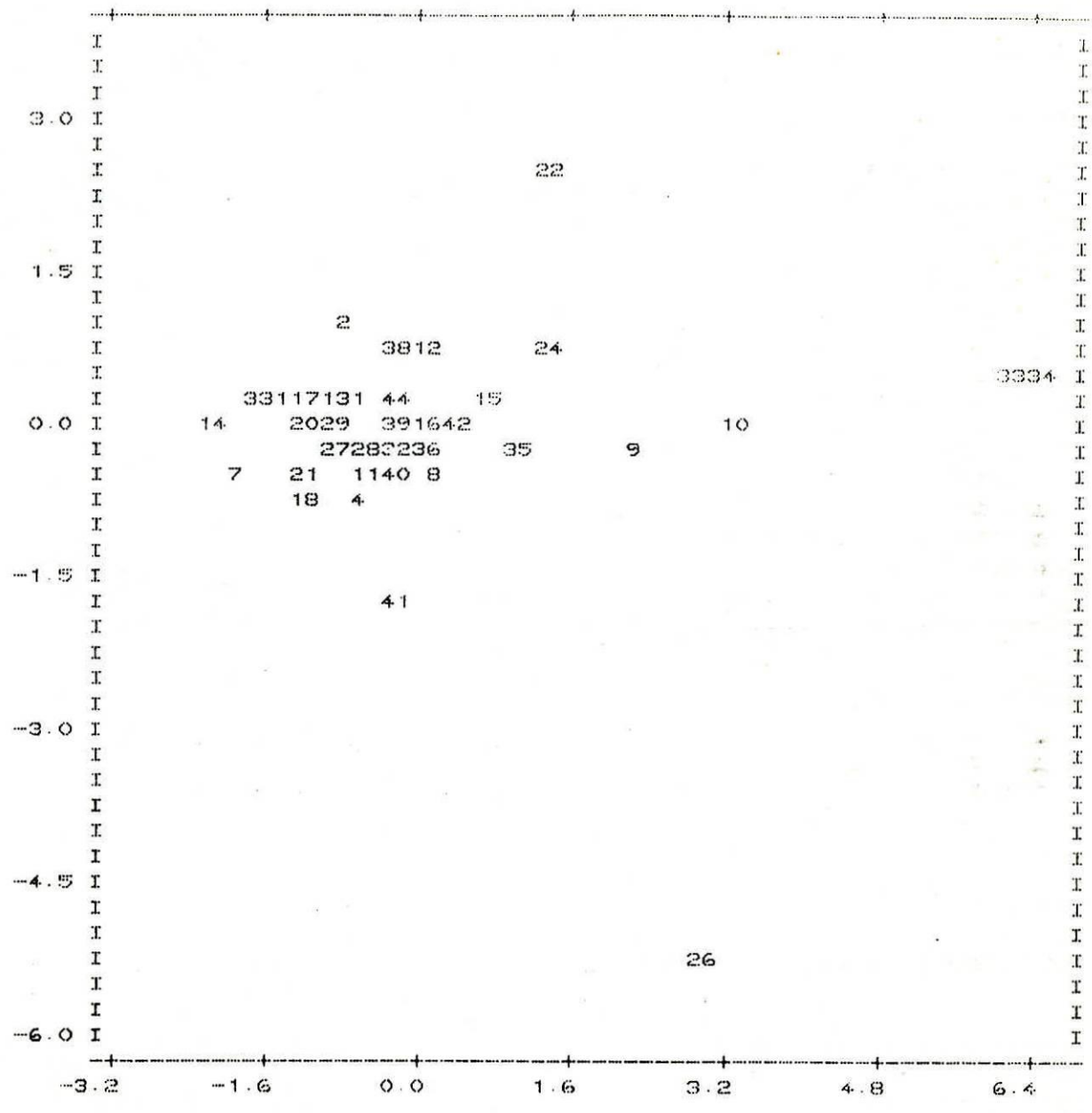
Multidimensional scaling (MDS)is the most interesting in pattern analysis. The mathematics is not simple, but the idea is simple. MDS begins with a matrix of associations based on many attributes; the user tells the program how many dimensions the information is to be reduced to (2 say). MDS then goes to work to produce a 2-d plot in such a way that the distances between all pairs of points are in proportion to the association between the points. It is impossible to do this exactly, but the maximum possible amount of association information which can be seen in 2 dimensions is displayed. If reduction is to 3 dimensions, the scores can be looked at in 3 scatterplots.
MDS is the recommended Ordination procedure in PATN.

The other methods use different criteria to find projections in the multidimensional space which carry the most information. Principal Coordinate Analysis is more generally applicable than the Principal Components method (which really relies on normally distributed data). These methods may be discussed in the workshop. Both methods produce useful plots at the data exploration stage.

Discussion Exercise: What will MDS do given the trivial 4-point example of section 4 above ?

Note: Ordination takes a lot of time on the computer, and the results may be useless if there are too many points. If there are many individuals (>100 say), it is recommended to use some clustering process first; then use ordination on the cluster centroids (centres) to see how the clusters are related.

PCA ORDINATION OF 46 ACCESSION USING 8 NUMERICAL ATTRIBUTES

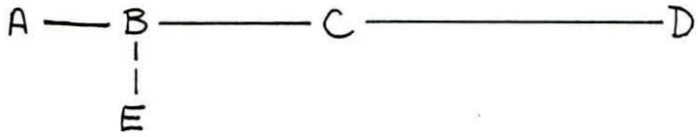


PCSC[2] v. PCSC[1] using factor ACC

Points coinciding with 12
19 25
Points coinciding with 22
23
Points coinciding with 29
30
Points coinciding with 17
37
Points coinciding with 16
43
Points coinciding with 13
45
Points coinciding with 32
46

7.0 NETWORKS

Network methods also use association measures as a starting point. They produce outputs which show linkages between close individuals in the multivariate space. They focus on the local neighbourhoods in the space. MINIMUM SPANNING TREES (MST) is the name of one type of NETWORK method which is available in PATN. It produces output, and from this we can draw figures like the one below, with numbers representing individuals, and distances (very roughly) the association between pairs.



Networks and Ordination methods describe different things about the data. Networks are true locally - that is close objects on the network are really close. Not much can be said about distant pairs. With ordination, the opposite is true - if objects are far apart on the simplified ordination, then they are truly distant; however close objects in the ordination may not really be close - the points may be separated in some of the higher dimensions.

8.0 So What Do We Do ?

Try it; that is on a data set which we know and understand, try many of the alternatives in the methods above, and combine and compare the results. Remember pattern analysis does not give one answer, it is a tool for data exploration and data reduction. The methods of CLUSTER, ORDINATION and NETWORKS all show different aspects of the data set, and all should be tried.

For the rice data set it may be sensible first to throw out the useless attributes, and perhaps to recode some others. We may then split up the attribute set into three smaller sets for the Vegetative, Reproductive, and Harvest stages. The pattern analysis results for the separate stages can be compared with an analysis using all the attributes. The analysis may identify a reduced set of useful attributes, and groups of individuals.

REFERENCES

Williams W.T (ed).1976. Pattern Analysis in Agricultural Science. CSIRO, Elsevier. Melbourne.

Belbin L.1989. The PATN Manual. CSIRO Division of Wildlife and Ecology. Canberra.

Payne R.W. 1987. The GENSTAT 5 Reference Manual. Clarendon Press. Oxford.

These books contain many further references for specific methods used in pattern analysis.

PART 2: Use of the Pattern Analysis Program PATN

These notes give a step-by-step guide to the use of the PATN program. The data set used is the rice accession data set of 46 accessions, 39 attributes measured (35 of these are useful). The notes cover only a few of the PATN routines - the program has a great range of procedures and options. The notes are not complete; they are to be used with demonstration and practice sessions.

The only way to discover what a computer program like this can do for you is to try it. Of course, we first need some background to understand what the program should be able to do. But we also need to try the program, to make mistakes, and to look at what happens with different options.

1.0 What You Need To Know First:

Before trying to use the PATN program, we need:

(1) Some knowledge of the computer system commands; directories, our data files, how to copy, print and display files, how to turn it on & off, etc.

(2) A data set which is suitable for Pattern Analysis: we need to store this data in a computer file and know the FORMAT of the file.
We use the rice accession data.

(3) Some knowledge of what Pattern Analysis can do for the data set. We should decide on what we want to do BEFORE we sit down at the computer.

(4) The PATN MANUAL. Before trying any new procedure we should look in the manual to discover the commands for the program. When running the program, the manual is also needed to explain some of the options.

The PATN manual gives a lot of background and references for Pattern Analysis in general, and for each procedure. It is useful to read the manual. Computer manuals usually make more sense after you have used the program. I recommend reading enough of a manual section to try a procedure, then read it again after looking at the results.

1.1 The PATN Program and the Computer

This version of PATN runs on a computer with an 80286 processor fitted with an 80287 maths coprocessor. (IBM PC AT or compatible). The coprocessor must be fitted or the program will not run.

The program is INTERACTIVE; this means that the user drives the program by answering questions on the screen, one at a time. PATN has sensible DEFAULTS for most answers. The program is therefore 'friendly' in a sense, but the questions will only make sense if the user knows what the routine is doing; the manual should be studied first for each procedure.

PATN produces a lot of working FILES, and output FILES. For this reason, users should run PATN from a user directory. A different directory could be set up for each data set. It may be convenient to place the user's working directory on floppy disk in drive A:. However, the program will run faster from hard disk.

The PATN programs themselves should all be loaded into a single directory ('PATN') on hard disk. If the user's PATH includes access to this directory, then the user can work in his own directory.

1.2 Starting the Program; the PATN prompts

To start PATN with the set-up above, the user needs to

Turn on the computer
Change to the user working directory (cd ...)
the data files should be copied there.
type PATN (return)
(if the program command is not recognised, the user path must be extended).

After a short time, the screen will show the PATN title banner. The first time, you will see a Warning at the bottom of the screen

Warning:Parameter file PATN.PRM not found.

This is a warning that PATN does not know about any data yet.

PATN expects two types of input from the user, it needs COMMANDS to start PROCEDURES; within procedures it asks the user to set PARAMETERS, usually from a choice of OPTIONS.

When the program starts, you will see at the bottom of the screen the PATN COMMAND PROMPT

```
*PATN<
```

The user should decide which procedure he wants. If you want histograms, the procedure is called HIST, and you type

```
*PATN< HIST  (return)
```

the screen then shows

```
==========================================>HIST
```

while PATN loads the procedure.

Of course, the procedure will not run without data !
PATN has many procedure commands, we will look at some below.

Within a PROCEDURE, PARAMETERS are chosen and set. Sometimes a menu of numbered choices is presented; the user chooses the OPTION by responding to the OPTION PROMPT. This prompt tells the user what type of answer is expected, and what DEFAULT will be used. For example, if the HIST procedure is started, you need to tell the program which attributes are to be plotted, the prompt is :

```
OPTIONS
 1: operate on ROWS
 2: operate on COLUMNS
 3: stop  (I, D:2)              ?:
```

The (I,D:2) tells the user that the expected answer is an integer and that the default value is 2, i.e. if we hit return, the program will choose the default value for us and make histograms of the columns of data.

1.3 PATN files, data files, output files

PATN uses files, and produces files. Some of these files are Patn WORKING FILES, and some are OUTPUT files for the user. You cannot print or TYPE the working files, they are binary files which can only be read by the computer. PATN uses the work files because they are faster for the program.

PATN actually changes our text (ASCII) data file into its own working file format before it can do anything. The procedures PRAM and DATN are used to do this.

This file management is quite difficult for new users, we will try to explain the files for each procedure in the workshop.

2.0 First Session - Starting PATN, some procedures

The source data file we will use is called SNRGB2.TXT, it is part of the rice accession data : 46 rows by 20 attribute columns. The listing of this file is attached. Note the packed FORMAT of the file. This file is a normal ASCII file, we can print, type and edit the file. Note, the 'X' and missing values have been replaced by numbers.

The commands we will use in the first session just start PATN, and read in the data. This is the hardest part for new computer users.

The procedures we use will be (in order)

PRAM - this tells patn about our data, how many rows, the filename we want for the working data file etc.
Suppose we choose the name RICE.DAT.
It produces file RICE.PRM which stores the details.

DATN - to read in our source data to the working data file.
Produces patn work file RICE.DAT.

LABN - to give names (labels) to the rows and columns.
Produces patn work files RICE.RLB & RICE.CLB

HIST - to create histograms of the attributes.
Produces output file RICE.HIS which stores the histograms,
This file can be printed, or typed to the screen.

So, let us begin with a directory which has only the data file in it; type dir to check.

To start PATN, type

```
PATN
```

you will see the banner and warning about no PRM file (no data), so, to tell patn about our data, type

```
PATN< PRAM
```

You will see some details and a menu with many options, at first patn knows nothing so choose option 0 ('ALTER ALL'). We then see a series of questions asking for title, answer as follows;

```
title : Rice data file 20 characters
file name: RICE.DAT                    (this is work file name)
number rows : 46
number columns: 20
row groups : 0
column groups : 0
```

The details are then displayed, and the menu appears again. We

have finished for now so choose option 11 ('STOP PRAM')
PATN now knows what sort of data to expect, to read it in, type

```
PATN< DATN
```

again you see a menu: choose option 1 (ascii -->PATN).
input filename: SNRGB2.TXT
output filename: RICE.DAT (the work file we defined above)
input format:
This is difficult, it must match our ascii file - if the file is free format it is easy, we just type (*). In our case the format is : (5x,4i1,f4.0,f3.0,f4.0,f5.0,i2,f4.0,6i1,f4.0,i1,i2,i1)

The data should then be read, if PATN has any problems, an error message will be displayed. The file RICE.DAT is created.
A warning message is given that no labels are known. type

```
PATN< LABN      to label the rows and columns
```

We see a series of menus again. We could type in all the accession names (for rows), and attribute names for columns.
At this stage choose option 1 ('AUTOGENERATION') and type in 'ACC' for the row BASE name; the rows will be named ACC0001 - ACC0046 automatically. The base name for columns could be COL.

Now, to see what we have done, type

```
PATN< $dir      this shows the current directory contents
                (the $ allows us to use DOS commands at the
                 PATN< prompt without leaving the program)
```

We should see our directory has grown to

SNRGB2.TXT	- the original file
RICE.PRM	- the data parameter description file
PATN.PRM	- patn's current parameter file = RICE.PRM
RICE.DAT	- patn working data file
RICE.RLB	- the row label file
RICE.CLB	- the column label file

All the new files are working files - you cannot print them.

Now (at last), we can get PATN to do something; try Histograms by typing

```
PATN< HIST
```

A menu appears, enter 2 for columns (this is the default); then you see

```
Enter column number (I,D:ALL) ? :
```

Just type return, and patn will use its default - it will produce histograms for all the different attributes.

BUT, there will be an ERROR MESSAGE ! PATN has a problem with the last attribute - all the values are the same so it has tried to divide by zero. The attribute is in fact useless. All the other histograms are produced, to see this, type

```
PATN< $dir
```

The file RICE.HIS has been produced. This is a user output file, you can look at it or print it. Try

```
PATN< $type RICE.HIS
```

or

```
PATN< $print RICE.HIS
```

ERROR TRAPPING
The error above is a rare case - usually if PATN has a problem it will give a helpful warning message and you can try again. Try typing something wrong like

```
PATN< HIPP    - there is no such command in PATN
                the program will tell you that it does not
                understand, and display a table of
                available commands
```

Now, we have done enough for a first session. Stop the program by typing

```
PATN< EXIT
```

The files remain in the directory, so we can start again where we finished. You can print the RICE.HIS file like any ascii file using normal DOS commands.

(PATN OUTPUT FILE RICE.HIS - part only)

04/25/89 07:57:30.89 HIST Rice data file characters 20-29

LABEL = col00001

Number..	46.00
Minimum.	1.000
1stQuart	1.000
Median..	1.000
Mean....	2.043
3rdQuart	4.000
Maximum.	6.000
Q3-Q1...	3.000
Av.Dev..	1.543
Std.Dev.	1.849
Variance	3.420
Range...	5.000
Sum.....	94.00
Num > 0.	46.00
Skewness	1.328
Kurtosis	0.3656E-01

1.000	=>	1.278	34**********
1.278	=>	1.556	0
1.556	=>	1.833	0
1.833	=>	2.111	0
2.111	=>	2.389	0
2.389	=>	2.667	0
2.667	=>	2.944	0
2.944	=>	3.222	0
3.222	=>	3.500	0
3.500	=>	3.778	0
3.778	=>	4.056	6****
4.056	=>	4.333	0
4.333	=>	4.611	0
4.611	=>	4.889	0
4.889	=>	5.167	0
5.167	=>	5.444	0
5.444	=>	5.722	0
5.722	=>	6.000	6****

LABEL = col00002

Number..	46.00
Minimum.	1.000
1stQuart	3.000
Median..	3.000
Mean....	3.348
3rdQuart	3.000
Maximum.	5.000
Q3-Q1...	0.0000
Av.Dev..	0.7183
Std.Dev.	0.9711
Variance	0.9430
Range...	4.000
Sum.....	154.0
Num > 0.	46.00
Skewness	0.4316
Kurtosis	0.3932

1.000	=>	1.222	2*
1.222	=>	1.444	0
1.444	=>	1.667	0
1.667	=>	1.889	0
1.889	=>	2.111	0
2.111	=>	2.333	0
2.333	=>	2.556	0
2.556	=>	2.778	0
2.778	=>	3.000	0
3.000	=>	3.222	34**********
3.222	=>	3.444	0
3.444	=>	3.667	0
3.667	=>	3.889	0
3.889	=>	4.111	0
4.111	=>	4.333	0
4.333	=>	4.556	0
4.556	=>	4.778	0
4.778	=>	5.000	10*******

LABEL = col00003

Number..	46.00
Minimum.	1.000
1stQuart	1.000
Median..	1.000
Mean....	1.196
3rdQuart	1.000
Maximum.	4.000
Q3-Q1...	0.0000
Av.Dev..	0.3573

1.000	=>	1.167	42**********
1.167	=>	1.333	0
1.333	=>	1.500	0
1.500	=>	1.667	0
1.667	=>	1.833	0
1.833	=>	2.000	0
2.000	=>	2.167	0
2.167	=>	2.333	0
2.333	=>	2.500	0
2.500	=>	2.667	0
2.667	=>	2.833	0

3.0 Session 2. Association and Clustering.

In this session we will use the following procedures.

MASK - to reduces the number of attribute columns, we have seen that col 20 is all zero and is no use. In these notes we reduce the data to 10 columns - the reproductive data.
This routine produces a new data (.DAT) working file with label files (.RLB, .CLB) and a .PRM file. These are produced automatically, you only have to give a name to the .DAT file. (we use XTEN.DAT) USE A DIFFERENT NAME FROM THE ORIGINAL FILE.

ASO - this routine is very important, it produces the association matrix between individuals.
Produces working file .ASO (e.g. XTEN.ASO).

FUSE - a clustering algorithm which performs hierarchical agglomerative clustering (or FUSION).
It uses the .ASO file and produces an output file called (e.g.) XTEN.FUS. We can print this file.

DEND - draws a dendrogram, using the .FUS file. Produces an output file XTEN.DEN we can see or print.

The data handling procedures, such as MASK, are often more difficult to use than the analysis procedures. If you are not familiar with computing language you will need to read the manual carefully or get some advice.

So to restart, type PATN. When it starts, type <PRAM. You see the details of the current file (it will be RICE.DAT from the first session). Patn remembers the last file used, we will create a new data file, and then use this.
Choose option 11 to 'STOP PRAM'. Then type

```
PATN< MASK          we want to keep all rows,
                    and keep columns 1-10
```

To do this, read the screen prompts
ROW MASK - Choose option 3 (Select all 1-46)
COL MASK - Choose option 1 (enter from terminal)
Patn asks you to list the columns you want, type
1 -10 <ctrl-z> this is short for 1 2 3 4 ... 10<ctrl-z>
There are more questions, just DEFAULT them all, type return.
Mask then asks you for the file name for the new working file, type in XTEN.DAT - do not use RICE.DAT again !!

Then it asks 'UPDATE PRM FILE??' , answer Y, and our new reduced file of 10 columns will be the active data file.
At the prompt PATN< , type PRAM again to check this, then 11 to exit the PRAM procedure.

Now, XTEN.DAT is the active file. We run the ASO procedure.

```
PATN< ASO
```

You see a menu of metric measures, choose option 1 (the default). That is all you have to do; you see messages, and the file XTEN.ASO is created. This contains the association matrix for the 46 accessions.

Next, we will try some clustering; type

```
PATN< FUSE
```

You see a menu of fusion strategies (Nearest neighbour etc).
Choose the default this time. (option 5, UPGMA).
Choose default for the next questions - these all do special things (see manual for more information).
This program produces an output file XTEN.FUS which you can see and print.

XTEN. FUS contains the fusion results, but it is not easy to see the groups. We will produce a DENDROGRAM display, type

```
PATN< DEND
```

This procedure goes straight to work on the .FUS file, it asks
  Number of groups to print (I,D 46):
The default is 46, ie all individuals appear, choose this.
Then set suitable display or print file option.

The idea of using the program should be clear now; you need to decide in advance what you want to do, choose the procedure with the help of the manual, and try it. Reading the manual section is necessary to understand all the choices. The defaults are often useful, especially in the analysis procedures.

We can stop PATN now ('exit') and discuss the results. It would be useful to try FUS with a different strategy (e.g. nearest) and compare the cluster results on the dendrograms.
Also, we could calculate a new .ASO matrix with a different metric and see what happens.

Discuss the clusters on the dendrogram, and other output files

(PATN OUTPUT FILE XTEN.FUS)

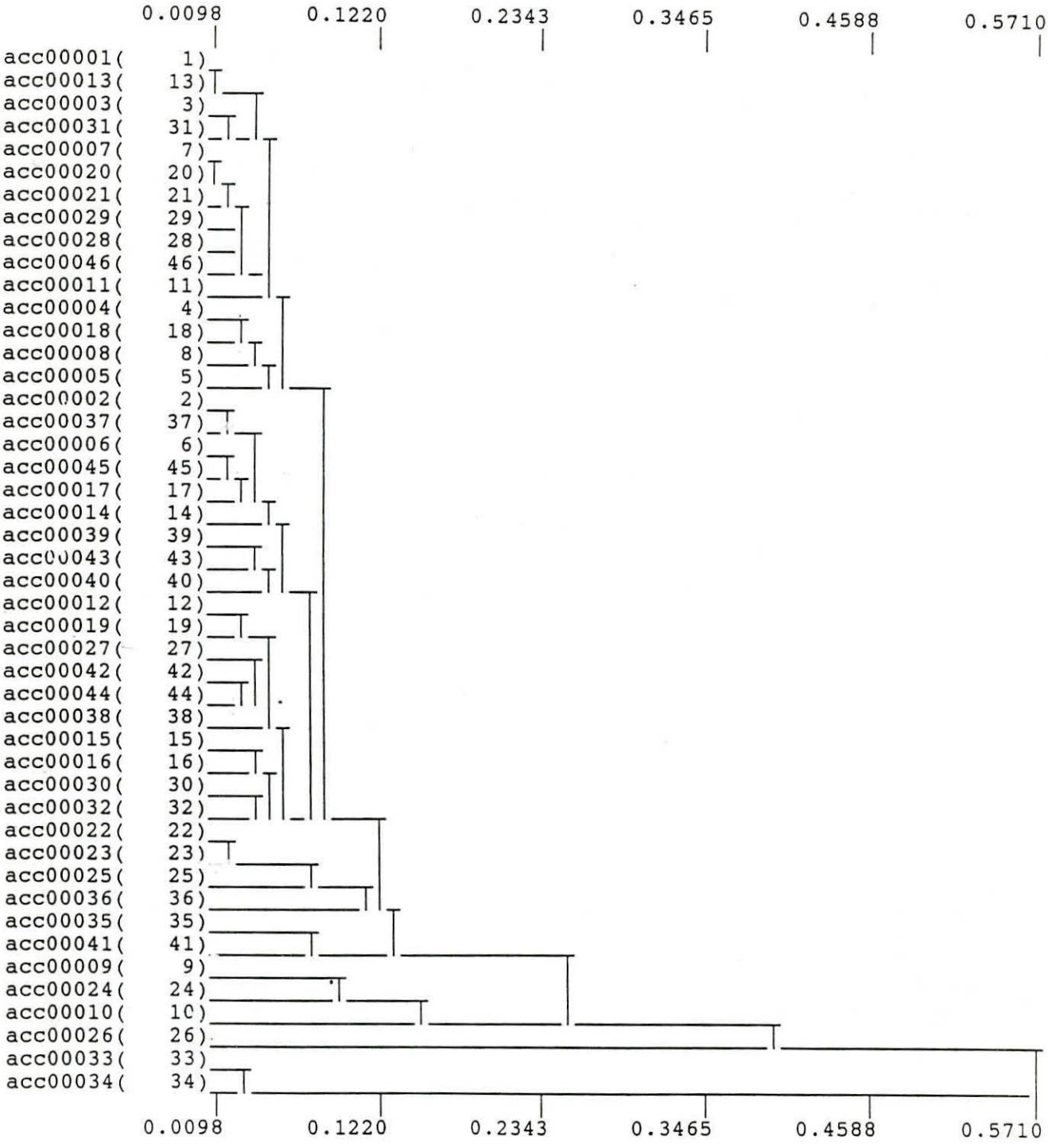
04/25/89 08:06:20.26 FUSE Rice data file characters 20-29

FLEXIBLE UPGMA OR GROUP AVERAGE FUSION WITH BETA = 0.00

GROUPS	FUSION GROUPS	NEW GROUP	LEVEL	INCREMENT	STRESS
45	acc00007(7)+acc00020(20)	=GP(7)-	0.978E-02	0.000	0.000
44	acc00001(1)+acc00013(13)	=GP(1)-	0.140E-01	0.426E-02	0.000
43	acc00002(2)+acc00037(37)	=GP(2)-	0.185E-01	0.448E-02	0.000
42	acc00006(6)+acc00045(45)	=GP(6)-	0.197E-01	0.117E-02	0.000
41	acc00007(7)+acc00021(21)	=GP(7)-	0.210E-01	0.132E-02	0.000
40	acc00022(22)+acc00023(23)	=GP(22)-	0.226E-01	0.157E-02	0.000
39	acc00003(3)+acc00031(31)	=GP(3)-	0.228E-01	0.208E-03	0.000
38	acc00033(33)+acc00034(34)	=GP(33)-	0.248E-01	0.204E-02	0.000
37	acc00007(7)+acc00029(29)	=GP(7)-	0.257E-01	0.892E-03	0.000
36	acc00042(42)+acc00044(44)	=GP(42)-	0.271E-01	0.138E-02	0.000
35	acc00028(28)+acc00046(46)	=GP(28)-	0.274E-01	0.274E-03	0.000
34	acc00012(12)+acc00019(19)	=GP(12)-	0.294E-01	0.207E-02	0.000
33	acc00007(7)+acc00028(28)	=GP(7)-	0.307E-01	0.123E-02	0.000
32	acc00004(4)+acc00018(18)	=GP(4)-	0.312E-01	0.501E-03	0.000
31	acc00006(6)+acc00017(17)	=GP(6)-	0.312E-01	0.799E-06	0.000
30	acc00027(27)+acc00042(42)	=GP(27)-	0.344E-01	0.327E-02	0.000
29	acc00039(39)+acc00043(43)	=GP(39)-	0.345E-01	0.778E-05	0.000
28	acc00001(1)+acc00003(3)	=GP(1)-	0.352E-01	0.759E-03	0.000
27	acc00004(4)+acc00008(8)	=GP(4)-	0.363E-01	0.107E-02	0.000
26	acc00015(15)+acc00016(16)	=GP(15)-	0.375E-01	0.118E-02	0.000
25	acc00030(30)+acc00032(32)	=GP(30)-	0.375E-01	0.312E-04	0.000
24	acc00002(2)+acc00006(6)	=GP(2)-	0.388E-01	0.127E-02	0.000
23	acc00004(4)+acc00005(5)	=GP(4)-	0.427E-01	0.392E-02	0.000
22	acc00007(7)+acc00011(11)	=GP(7)-	0.430E-01	0.306E-03	0.000
21	acc00012(12)+acc00027(27)	=GP(12)-	0.431E-01	0.723E-04	0.000
20	acc00002(2)+acc00014(14)	=GP(2)-	0.431E-01	0.239E-04	0.000
19	acc00001(1)+acc00007(7)	=GP(1)-	0.467E-01	0.359E-02	0.000
18	acc00015(15)+acc00030(30)	=GP(15)-	0.478E-01	0.108E-02	0.000
17	acc00039(39)+acc00040(40)	=GP(39)-	0.482E-01	0.424E-03	0.000
16	acc00012(12)+acc00038(38)	=GP(12)-	0.495E-01	0.128E-02	0.000
15	acc00002(2)+acc00039(39)	=GP(2)-	0.546E-01	0.514E-02	0.000
14	acc00001(1)+acc00004(4)	=GP(1)-	0.590E-01	0.443E-02	0.000
13	acc00012(12)+acc00015(15)	=GP(12)-	0.600E-01	0.970E-03	0.000
12	acc00002(2)+acc00012(12)	=GP(2)-	0.707E-01	0.107E-01	0.000
11	acc00035(35)+acc00041(41)	=GP(35)-	0.756E-01	0.494E-02	0.000
10	acc00022(22)+acc00025(25)	=GP(22)-	0.763E-01	0.715E-03	0.000
9	acc00001(1)+acc00002(2)	=GP(1)-	0.810E-01	0.465E-02	0.000
8	acc00009(9)+acc00024(24)	=GP(9)-	0.880E-01	0.699E-02	0.000
7	acc00022(22)+acc00036(36)	=GP(22)-	0.113	0.254E-01	0.000
6	acc00001(1)+acc00022(22)	=GP(1)-	0.122	0.916E-02	0.000
5	acc00001(1)+acc00035(35)	=GP(1)-	0.129	0.662E-02	0.000
4	acc00009(9)+acc00010(10)	=GP(9)-	0.151	0.221E-01	0.000
3	acc00001(1)+acc00009(9)	=GP(1)-	0.253	0.102	0.000
2	acc00001(1)+acc00026(26)	=GP(1)-	0.391	0.138	0.000
1	acc00001(1)+acc00033(33)	=GP(1)-	0.571	0.180	0.000

STRESS THRESHOLD= 0.001E-37 AVERAGE INCREMENT & STRESS : 0.125E-01 0.000

(PATN OUTPUT FILE XTEN.DEN)



4.0 Session 3. Ordination and Network.

In this session we continue to use the data file XTEN.DAT, of 46 rows by ten columns. The association matrix XTEN.ASO has been calculated already and is stored. Now we will use the procedures;

MDS - Multidimensional Scaling. An ordination procedure based on the Association matrix.
This produces two files, XTEN.MDS contains the ordination scores, it is a printable outptut file.

MST - Network procedure, also based on the association Matrix.
Produces an output file XTEN.MST, which displays the links in the network. The actual network is not drawn, you have to print this file and draw it yourself if you wish to see the whole network.

Alternative ordination procedure PCA will be used if there is time.

To begin, start PATN, and type PRAM to check that the XTEN.DAT file is active (type 11 to STOP PRAM); then type

```
PATN< MDS
```

You will see questions appear one after another - the defaults will produce good answers; we will discuss these in the workshop in turn; the first is

```
Number of Dimensions (I,D:2)   the default is a 2-d ordination
```

then other parameters are advised
Maximum iterations etc
Just use the defaults each time (i.e. type return), down to
Source of Initial Configuration
- choose option 1 ('random starts')
But for Number of random starts, the default is 10, this is recommended for real work but it takes time, so put in 5 for a first try.

What happens is this. MDS tries to get the best 2-d picture of the associations. It remembers the best fit and stores it in the output file.
When you have answered all the questions, MDS goes to work.
It takes some time, WAIT !!.

Then type the output file XTEN.MDS to the screen,
or print it. Discuss.

The Network Method MST is next; this is very easy to run; type

```
PATN< MST
```

That is all, no options at all. MST goes straight to work to produce the network links. Look at the output file XTEN.MST, it displays which individuals are linked ('close'), and the measure of association of that link.

Print the file XTEN.MST, and draw the links, or discuss them.

It is now a suitable stage to discuss what we have learnt about computers, multivariate data, and Pattern Analysis and interpretation

Other procedures should be tried, but it is now important to look at the results from FUSION (XTEN.FUS and XTEN.DEN); from ORDINATION (XTEN.MDS and plot); from NETWORK (XTEN.MST). All these results depend on the data we have used (10 columns), and the association options we used.

Can we see any groups or separate outlying accessions? If so there are other PATN routine to tell us about the important attribute in defining the structure. Perhaps we should try to use all the data to confirm our ideas. Or, we could use the data from a different growth stage, say vegetative.

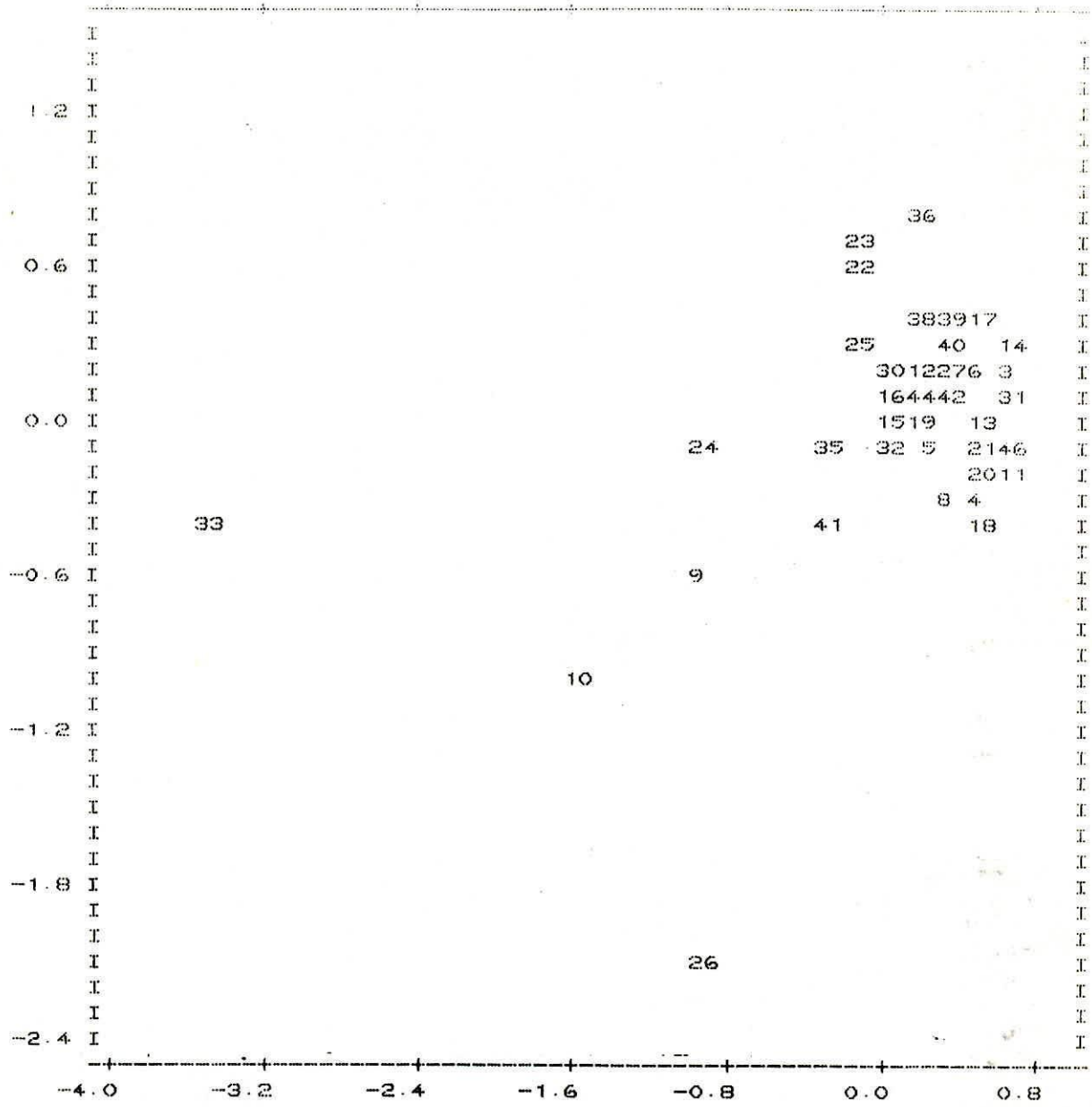
(PATN OUTPUT FILE XTEN.MDS)

04/25/89 10:30:06.04 MDSM Rice data file characters 20-29

LABEL	VECTORS 1-> 2	
acc00001(1)	0.4302	0.0803
acc00002(2)	0.3851	0.4181
acc00003(3)	0.6317	0.1917
acc00004(4)	0.4474	-0.2799
acc00005(5)	0.2766	-0.1306
acc00006(6)	0.4841	0.2020
acc00007(7)	0.5463	-0.1822
acc00008(8)	0.3384	-0.3372
acc00009(9)	-0.9290	-0.5936
acc00010(10)	-1.5686	-0.9738
acc00011(11)	0.6947	-0.1881
acc00012(12)	0.1322	0.1553
acc00013(13)	0.5095	0.0457
acc00014(14)	0.6904	0.3270
acc00015(15)	-0.0635	-0.0189
acc00016(16)	0.0237	0.1475
acc00017(17)	0.4479	0.3818
acc00018(18)	0.4231	-0.3633
acc00019(19)	0.1481	0.0218
acc00020(20)	0.5072	-0.1689
acc00021(21)	0.4931	-0.0831
acc00022(22)	-0.1955	0.6102
acc00023(23)	-0.2132	0.7193
acc00024(24)	-0.9403	-0.1477
acc00025(25)	-0.2342	0.3352
acc00026(26)	-0.9883	-2.1413
acc00027(27)	0.2758	0.2439
acc00028(28)	0.5312	-0.1270
acc00029(29)	0.5007	-0.2296
acc00030(30)	0.0551	0.1855
acc00031(31)	0.6748	0.1170
acc00032(32)	0.0283	-0.1031
acc00033(33)	-3.5376	-0.4239
acc00034(34)	-3.4527	-0.3570
acc00035(35)	-0.3224	-0.0980
acc00036(36)	0.0864	0.7605
acc00037(37)	0.4892	0.3792
acc00038(38)	0.1610	0.4447
acc00039(39)	0.3029	0.4498
acc00040(40)	0.2990	0.3458
acc00041(41)	-0.2606	-0.3875
acc00042(42)	0.2993	0.1159
acc00043(43)	0.1052	0.3583
acc00044(44)	0.2195	0.1160
acc00045(45)	0.4832	0.2352
acc00046(46)	0.5845	-0.0530

RANDOM NUMBER SEED = 4.500000000
INT/RATIO-ORDINAL CUT = 0.8000000000
STRESS (1) = 0.7356687000E-01
NUMBER OF ITERATIONS = 44
1=INTERVAL__2=RATIO = 2

(ORDINATION PLOT OF MDS SCORES IN XTEN.MDS)
(NUMBERS ARE ACCESSIONS - PRODUCED BY GENSTAT)



MDS2 v. MDS1 using factor ACC

Points coinciding with 21
28
Points coinciding with 20
29
Points coinciding with 33
34
Points coinciding with 17
37
Points coinciding with 38
43
Points coinciding with 6
45

(PATN OUTPUT FILE XTEN.MST)

04/25/89 10:55:04.73 MST Rice data file characters 20-29

```
INDENT   LABEL (SEQ.#)  -  LABEL (SEQ.#): ASSOCIATION
------   -------------     ------------  -----------

  1  acc00001(     1) - acc00006(     6) :    0.0258
  2   acc00006(     6) - acc00031(    31) :    0.0289
  3    acc00031(    31) - acc00003(     3) :    0.0228
  2   acc00006(     6) - acc00037(    37) :    0.0324
  3    acc00037(    37) - acc00002(     2) :    0.0185
  4     acc00002(     2) - acc00039(    39) :    0.0357
  5      acc00039(    39) - acc00043(    43) :    0.0345
  6       acc00043(    43) - acc00030(    30) :    0.0372
  7        acc00030(    30) - acc00025(    25) :    0.0633
  8         acc00025(    25) - acc00022(    22) :    0.0740
  9          acc00022(    22) - acc00023(    23) :    0.0226
  7         acc00030(    30) - acc00032(    32) :    0.0375
  6        acc00043(    43) - acc00036(    36) :    0.0721
  3    acc00037(    37) - acc00014(    14) :    0.0350
  2   acc00006(     6) - acc00045(    45) :    0.0197
  3    acc00045(    45) - acc00017(    17) :    0.0257
  1  acc00001(     1) - acc00013(    13) :    0.0140
  2   acc00013(    13) - acc00021(    21) :    0.0253
  3    acc00021(    21) - acc00004(     4) :    0.0277
  4     acc00004(     4) - acc00005(     5) :    0.0385
  4     acc00004(     4) - acc00018(    18) :    0.0312
  5      acc00018(    18) - acc00008(     8) :    0.0353
  3    acc00021(    21) - acc00020(    20) :    0.0163
  4     acc00020(    20) - acc00007(     7) :    0.0098
  4     acc00020(    20) - acc00011(    11) :    0.0325
  4     acc00020(    20) - acc00029(    29) :    0.0216
  3    acc00021(    21) - acc00028(    28) :    0.0252
  2   acc00013(    13) - acc00046(    46) :    0.0256
  1  acc00001(     1) - acc00042(    42) :    0.0269
  2   acc00042(    42) - acc00027(    27) :    0.0324
  2   acc00042(    42) - acc00044(    44) :    0.0271
  3    acc00044(    44) - acc00016(    16) :    0.0336
  4     acc00016(    16) - acc00015(    15) :    0.0375
  5      acc00015(    15) - acc00035(    35) :    0.0535
  6       acc00035(    35) - acc00009(     9) :    0.1302
  7        acc00009(     9) - acc00010(    10) :    0.1389
  8         acc00010(    10) - acc00026(    26) :    0.2088
  8         acc00010(    10) - acc00034(    34) :    0.3099
  9          acc00034(    34) - acc00033(    33) :    0.0248
  7        acc00009(     9) - acc00024(    24) :    0.0880
  6       acc00035(    35) - acc00041(    41) :    0.0756
  3    acc00044(    44) - acc00019(    19) :    0.0308
  4     acc00019(    19) - acc00012(    12) :    0.0294
  5      acc00012(    12) - acc00038(    38) :    0.0348
  3    acc00044(    44) - acc00040(    40) :    0.0345
```

NUMBER OF MST LINKAGES/OBJECT:

```
acc00033(   33) :   1.     acc00026(   26) :   1.     acc00003(    3) :   1.
acc00036(   36) :   1.     acc00005(    5) :   1.     acc00038(   38) :   1.
acc00007(    7) :   1.     acc00008(    8) :   1.     acc00041(   41) :   1.
acc00014(   14) :   1.     acc00011(   11) :   1.     acc00028(   28) :   1.
acc00029(   29) :   1.     acc00046(   46) :   1.     acc00023(   23) :   1.
acc00032(   32) :   1.     acc00017(   17) :   1.     acc00024(   24) :   1.
acc00027(   27) :   1.     acc00040(   40) :   1.     acc00045(   45) :   2.
acc00002(    2) :   2.     acc00015(   15) :   2.     acc00018(   18) :   2.
acc00025(   25) :   2.     acc00034(   34) :   2.     acc00019(   19) :   2.
acc00012(   12) :   2.     acc00031(   31) :   2.     acc00022(   22) :   2.
acc00039(   39) :   2.     acc00016(   16) :   2.     acc00037(   37) :   3.
acc00010(   10) :   3.     acc00001(    1) :   3.     acc00004(    4) :   3.
acc00013(   13) :   3.     acc00030(   30) :   3.     acc00035(   35) :   3.
acc00042(   42) :   3.     acc00009(    9) :   3.     acc00043(   43) :   3.
acc00020(   20) :   4.     acc00044(   44) :   4.     acc00021(   21) :   4.
```

5.0 Final comments

The user of such a big and powerful program will always face problems in understanding all the possible choices. It is also difficult to handle computers and computer files if you are not used to it.

The user should always try to decide in advance what he wants the program to do; and what sort of output picture he would like to get from the program. PATN will probably be able to do it somehow, so look in the manual for a suitable procedure and set of options. It is better to understand a few procedures at first. With confidence, you can try more later. But remember, always have an idea of what you want, and make sure that you drive the program - not the other way round !

APPENDIX

THE DATA FILE SNRGB2.TXT - 46 ACCESSIONS (ROWS), 20 ATTRIBUTES
FORMAT (5X,4I1,F4.1,F3.1,F4.1,F5.1,I2,F4.1,6I1,F4.1,I1,I2,I1)

```
2278 131166.01.319.0147.0 9 8.349550427.21 30
3442 131180.01.418.0136.0 9 5.034380429.01 30
3535 131177.81.718.0152.0 9 9.333380426.81 50
3727 131151.01.424.0150.0 810.034380427.61 30
4396 131152.81.617.0142.0 8 8.345350424.41 30
6338 131174.01.720.0144.0 8 8.334380428.41 30
6490 131156.22.619.0156.0 7 8.364380425.81 30
6574 131144.61.720.0147.0 9 6.634250424.21 30
8117 131143.21.218.0 80.013 5.053380423.61 30
8190 131142.81.014.0 53.0 4 5.063350422.41 X0
9149 131158.41.622.0166.011 8.364380423.81 30
9534 131161.01.312.0135.010 6.645350426.21 30
9804 131165.81.219.0153.010 8.328350429.01 30
9928 111184.01.524.0145.0 910.038390430.41 30
9954 651154.41.318.0127.011 6.04835X426.21 30
10006453162.21.318.0129.013 8.07848X426.01 30
10012111174.41.415.0139.01111.648380425.81 30
10022453148.02.022.0147.0 810.038350428.81 30
10049651159.81.414.0137.0 9 6.04835X427.41 30
10073131156.21.621.0155.0 8 8.337380431.61 30
10075131159.81.721.0152.0 810.035350426.81 30
11450131165.81.1 2.8122.011 1.164350425.42 30
11455131170.01.2 2.7118.612 0.464350426.22 30
11487131154.01.211.0 83.0 6 6.623380127.62 X0
11537634166.01.613.0116.6 8 6.045388527.02 10
11549131141.41.272.0 56.0 2 5.038750417.42 50
12355131168.41.219.0135.01011.645380427.61 30
12640451158.61.615.3153.0 810.045350423.01 30
12674451156.41.822.0153.0 7 6.635380429.21 30
12691131163.41.618.0126.0 810.0 2250428.81 30
12861131174.41.320.0156.0 710.035450429.01 30
12884131155.21.322.0127.0 6 8.358250428.41 30
15078131223.40.5 3.2 25.4 3 2.613310415.01 30
15277131225.20.5 4.0 25.8 3 2.313310512.01 30
15283451151.21.419.0113.014 6.654353423.21 X0
15291131167.81.420.0122.037 8.323380424.61 30
15293131182.01.717.0139.0 9 8.358280425.01 30
15311131166.81.410.0135.017 8.364380425.01 30
1534413 178.21.123.0131.012 8.365380426.11 20
15353653172.01.724.0135.010 7.674280421.61 30
15354631157.41.028.0109.01215.064380427.02 30
15373131164.21.220.0140.011 6.634380424.21 30
15374131173.01.221.0124.0 9 8.323380424.21 10
15382651163.41.120.0136.011 8.384351428.81 30
15410131172.81.316.0145.01010.063380425.81 30
15733451162.61.619.0157.0 9 8.339380421.21 30
```