Genstat®

Multivariate analysis

www.vsni.co.uk

VSNi

A Guide to Multivariate Analysis in Genstat®
(18th Edition)

by Simon Harding and Roger Payne.

Genstat is developed by VSN International Ltd, in collaboration with practising statisticians at Rothamsted and other organisations in Britain, Australia, New Zealand and The Netherlands.

Published by: VSN International, 2 Amberside, Wood Lane,
Hemel Hempstead, Hertfordshire HP2 4TP, UK
E-mail: info@genstat.co.uk
Website: http://www.genstat.co.uk/

First published 2010, for GenStat *for Windows* 13th Edition
This edition published 2015, for Genstat *for Windows* 18th Edition

Contents

Introduction

Multivariate analysis is useful when you have several different measurements on a set of *n* objects. In Genstat the measurements would usually be stored in separate variates, and these would have a unit for each object. The objects are often regarded as being a set of *n* points in *p* dimensions (*p* being the number of variates).

Many techniques, for example principal components analysis (Chapter 2) and canonical variates analysis (Chapter 3) are aimed at reducing the dimensionality. That is, they aim to find a smaller number of dimensions (usually 2 or 3) that exhibit most of the variation present in the data. This can help you determine patterns or structure in the data, as well as identify the relative importance of individual variables. Genstat has several menus for producing graphical representations, for example principal coordinates analysis (Chapter 4) and multidimensional scaling (Chapter 5). It also has facilities for modelling multivariate data, including multivariate analysis of variance (Chapter 8) and partial least squares.

Another important requirement is to take a set of units and classify them into groups based on their observed characteristics. Hierarchical cluster analysis (Chapter 6) starts with a set of groups each of which contains one of the units. These initial groups are successively merged into larger groups, according to their similarity, until there is just one group containing all the observations. Genstat also provide menus for non-hierarchical classification (Chapter 7), where the aim is to form a single grouping of the observations that optimizes some criterion such as the within-class dispersion, or the Mahalanobis squared distance between the groups, or the between-group sum of squares.

Chapter 9 describes the facilities for constructing classification trees, which allow you to predict the classification of unknown objects using multivariate observations. Regression trees, which predict the value of a response variate from multivariate observations are described in Chapter 10.

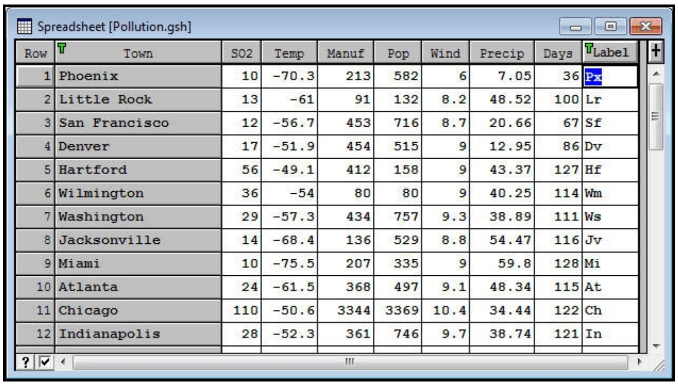
Finally, Chapter 11 describes how generalized Procrustes analysis can be used to obtain a consensus from assessors in activities such as wine tasting.

The book works through a series of straightforward examples, with frequent practicals to allow you to try the methods for yourself. The examples work mainly through the menus of Genstat *for Windows*, so there is no need for prior knowledge of the Genstat command language. Details of the commands for multivariate analysis can be found in Chapter 6 of the *Guide to the Genstat Command Language, Part 2, Statistics*.

1 Exploratory data analysis

Before you begin a multivariate analysis, it is sensible to investigate the properties of the data by looking at some plots and summary statistics.

We illustrate some of the insights that can be obtained by examining seven variables recorded in 41 towns in the USA, stored in the Spreadsheet file Pollution.gsh shown in Figure 1.1:



Spreadsheet [Pollution.gsh]

Row	Town	SO2	Temp	Manuf	Pop	Wind	Precip	Days	Label
1	Phoenix	10	-70.3	213	582	6	7.05	36	Px
2	Little Rock	13	-61	91	132	8.2	48.52	100	Lr
3	San Francisco	12	-56.7	453	716	8.7	20.66	67	Sf
4	Denver	17	-51.9	454	515	9	12.95	86	Dv
5	Hartford	56	-49.1	412	158	9	43.37	127	Hf
6	Wilmington	36	-54	80	80	9	40.25	114	Wm
7	Washington	29	-57.3	434	757	9.3	38.89	111	Ws
8	Jacksonville	14	-68.4	136	529	8.8	54.47	116	Jv
9	Miami	10	-75.5	207	335	9	59.8	128	Mi
10	Atlanta	24	-61.5	368	497	9.1	48.34	115	At
11	Chicago	110	-50.6	3344	3369	10.4	34.44	122	Ch
12	Indianapolis	28	-52.3	361	746	9.7	38.74	121	In

Figure 1.1

SO2	sulphur dioxide
Temp	temperature in degrees Fahrenheit
Manuf	number of enterprises with 20+ staff
Pop	population size in thousands
Wind	average annual wind speed (miles per hour)
Precip	average annual rainfall in inches
Days	average number of days with rain per year

(For more details, see Everitt, 2005, *An R and S-PLUS Companion to Multivariate Analysis*.)

First we open the Summary Statistics menu by clicking on the Summary Statistics sub-option of the Summary Statistics option of the Stats menu on the menu bar (see Figure 1.2).

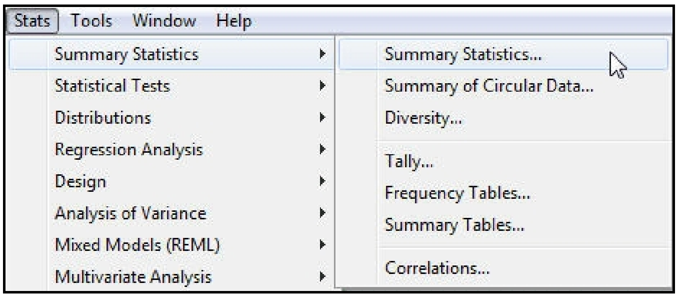
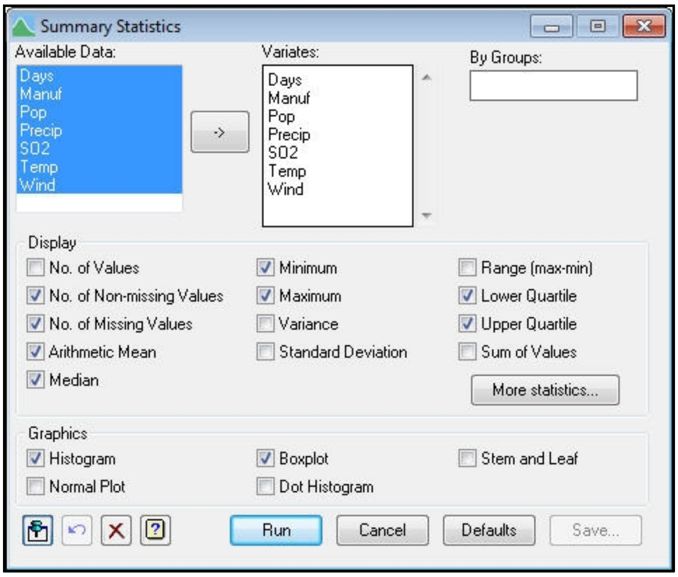


Figure 1.2

In the menu (Figure 1.3), we have entered all seven variates into the Variates box, selected the required summary statistics in the Options box, and checked the Histogram and Boxplot boxes in the Graphics box. The resulting output is shown below, and the plots are in Figures 1.4 - 1.17.



Available Data: Days Manuf Pop Precip SO2 Temp Wind

Variates: Days Manuf Pop Precip SO2 Temp Wind

By Groups:

Display

☐ No. of Values ☑ Minimum ☐ Range (max-min)
☑ No. of Non-missing Values ☑ Maximum ☑ Lower Quartile
☑ No. of Missing Values ☐ Variance ☑ Upper Quartile
☑ Arithmetic Mean ☐ Standard Deviation ☐ Sum of Values
☑ Median

Graphics

☑ Histogram ☑ Boxplot ☐ Stem and Leaf
☐ Normal Plot ☐ Dot Histogram

Figure 1.3

Summary statistics for Days

```
 Number of observations =   41
Number of missing values =   0
                    Mean =   113.9
                  Median =   115
                 Minimum =   36
                 Maximum =   166
          Lower quartile =   102.2
          Upper quartile =   128.2
      Standard deviation =   26.51
```

Summary statistics for Manuf

```
 Number of observations =   41
Number of missing values =   0
                    Mean =   463.1
                  Median =   347
                 Minimum =   35
                 Maximum =   3344
          Lower quartile =   170
          Upper quartile =   488.8
      Standard deviation =   563.5
```

Summary statistics for Pop

```
 Number of observations =   41
Number of missing values =   0
                    Mean =   608.6
                  Median =   515
                 Minimum =   71
                 Maximum =   3369
          Lower quartile =   293.5
          Upper quartile =   723.8
```

```
      Standard deviation =   579.1
```

Summary statistics for Precip

```
  Number of observations =   41
Number of missing values =   0
                    Mean =   36.77
                  Median =   38.74
                 Minimum =   7.05
                 Maximum =   59.8
          Lower quartile =   30.93
          Upper quartile =   43.17
      Standard deviation =   11.77
```

Summary statistics for SO2

```
  Number of observations =   41
Number of missing values =   0
                    Mean =   30.05
                  Median =   26
                 Minimum =   8
                 Maximum =   110
          Lower quartile =   12.75
          Upper quartile =   35.25
      Standard deviation =   23.47
```

Summary statistics for Temp

```
  Number of observations =   41
Number of missing values =   0
                    Mean =   -55.76
                  Median =   -54.6
                 Minimum =   -75.5
                 Maximum =   -43.5
          Lower quartile =   -59.32
          Upper quartile =   -50.55
      Standard deviation =   7.228
```

Summary statistics for Wind

```
  Number of observations =   41
Number of missing values =   0
                    Mean =   9.444
                  Median =   9.3
                 Minimum =   6
                 Maximum =   12.7
          Lower quartile =   8.7
          Upper quartile =   10.6
      Standard deviation =   1.429
```

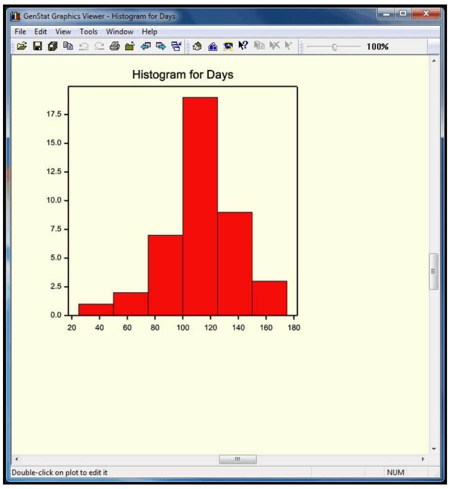

GenStat Graphics Viewer - Histogram for Days
File Edit View Tools Window Help
100%
Histogram for Days
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
20 40 60 80 100 120 140 160 180
Double-click on plot to edit it
NUM

Figure 1.4

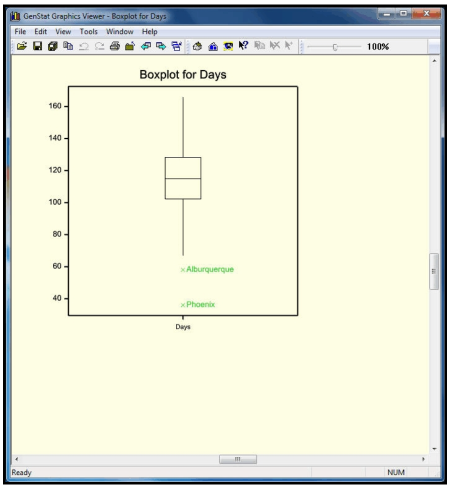

GenStat Graphics Viewer - Boxplot for Days
File Edit View Tools Window Help
100%
Boxplot for Days
160
140
120
100
80
60
40
Alburquerque
Phoenix
Days
Ready
NUM

Figure 1.5

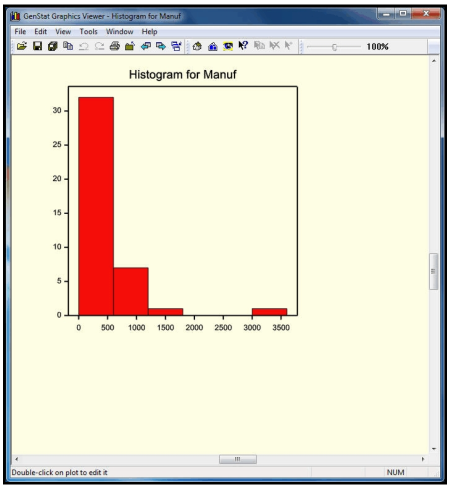

GenStat Graphics Viewer - Histogram for Manuf
File Edit View Tools Window Help
100%
Histogram for Manuf
30
25
20
15
10
5
0
0 500 1000 1500 2000 2500 3000 3500
Double-click on plot to edit it
NUM

Figure 1.6

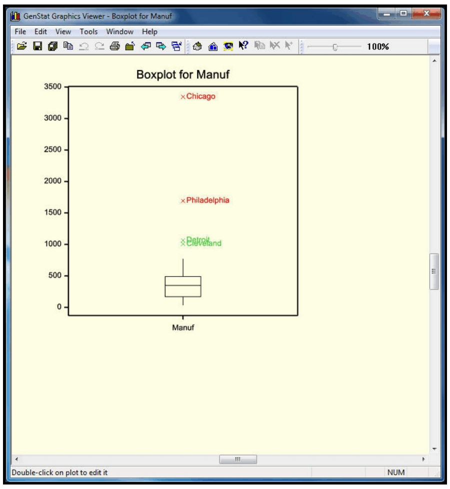

GenStat Graphics Viewer - Boxplot for Manuf
File Edit View Tools Window Help
100%
Boxplot for Manuf
3500
3000
2500
2000
1500
1000
500
0
Chicago
Philadelphia
Detroit
Cleveland
Manuf
Double-click on plot to edit it
NUM

Figure 1.7

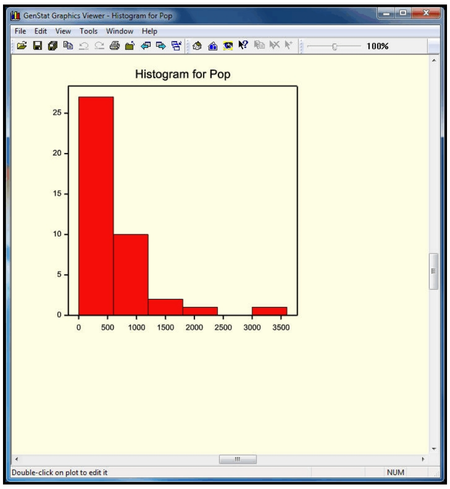

GenStat Graphics Viewer - Histogram for Pop
File Edit View Tools Window Help
100%
Histogram for Pop
25
20
15
10
5
0
0 500 1000 1500 2000 2500 3000 3500
Double-click on plot to edit it
NUM

Figure 1.8

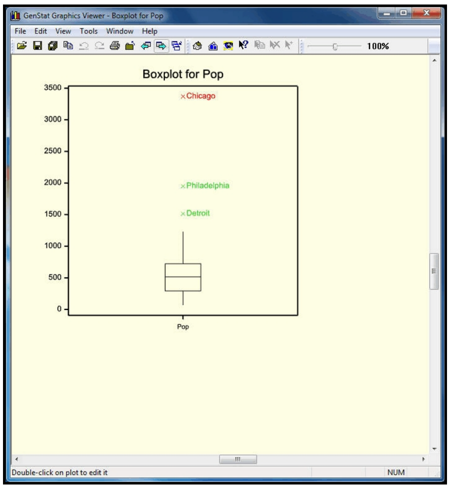

GenStat Graphics Viewer - Boxplot for Pop
File Edit View Tools Window Help
100%
Boxplot for Pop
3500
3000
2500
2000
1500
1000
500
0
Chicago
Philadelphia
Detroit
Pop
Double-click on plot to edit it
NUM

Figure 1.9

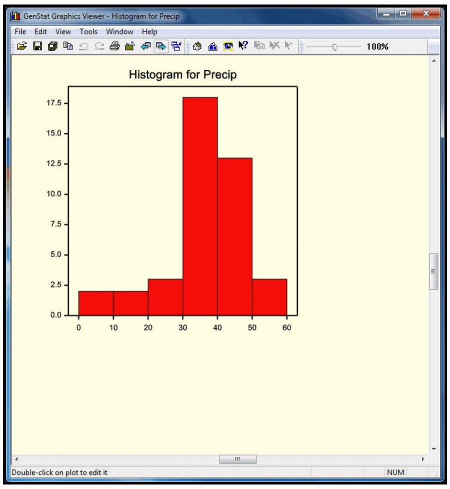

GenStat Graphics Viewer - Histogram for Precip
File Edit View Tools Window Help
100%
Histogram for Precip
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
10
20
30
40
50
60
Double-click on plot to edit it
NUM

Figure 1.10

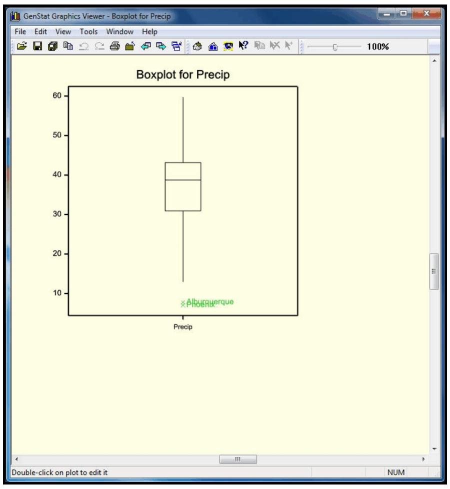

GenStat Graphics Viewer - Boxplot for Precip
File Edit View Tools Window Help
100%
Boxplot for Precip
60
50
40
30
20
10
Albuquerque
Phoenix
Precip
Double-click on plot to edit it
NUM

Figure 1.11

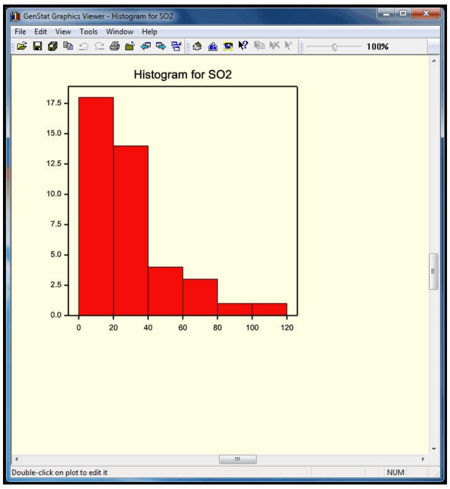

GenStat Graphics Viewer - Histogram for SO2
File Edit View Tools Window Help
100%
Histogram for SO2
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
20
40
60
80
100
120
Double-click on plot to edit it
NUM

Figure 1.12

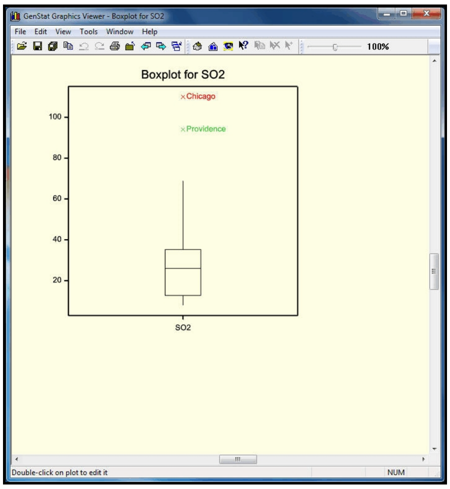

GenStat Graphics Viewer - Boxplot for SO2
File Edit View Tools Window Help
100%
Boxplot for SO2
Chicago
Providence
100
80
60
40
20
SO2
Double-click on plot to edit it
NUM

Figure 1.13

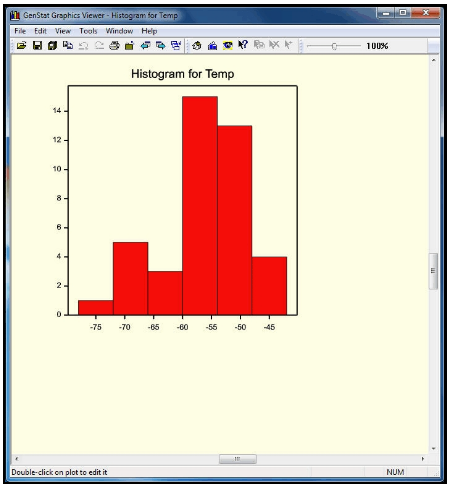

GenStat Graphics Viewer - Histogram for Temp
File Edit View Tools Window Help
100%
Histogram for Temp
14
12
10
8
6
4
2
0
-75
-70
-65
-60
-55
-50
-45
Double-click on plot to edit it
NUM

Figure 1.14

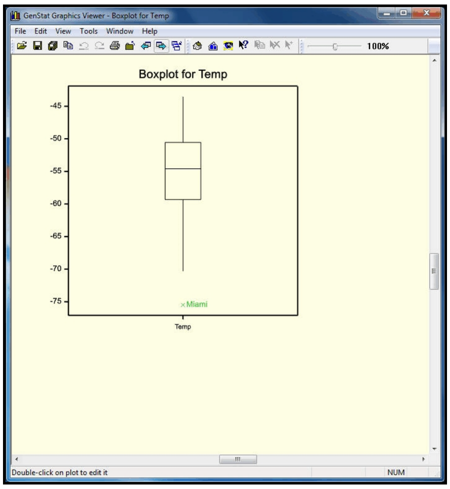

GenStat Graphics Viewer - Boxplot for Temp
File Edit View Tools Window Help
100%
Boxplot for Temp
-45
-50
-55
-60
-65
-70
-75
Miami
Temp
Double-click on plot to edit it
NUM

Figure 1.15

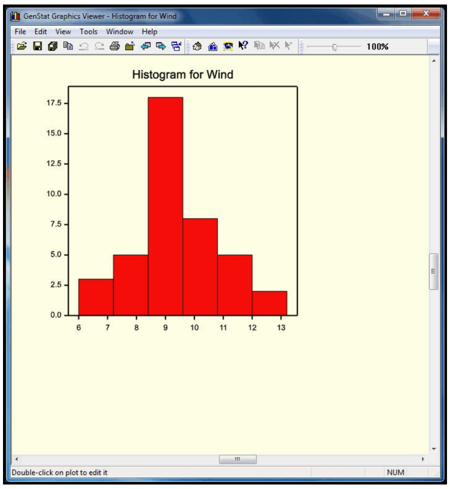


Figure 1.16

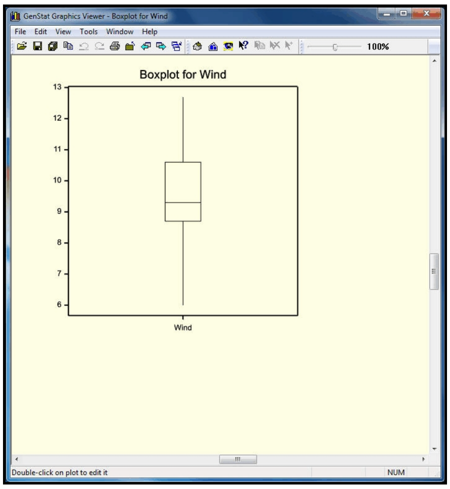
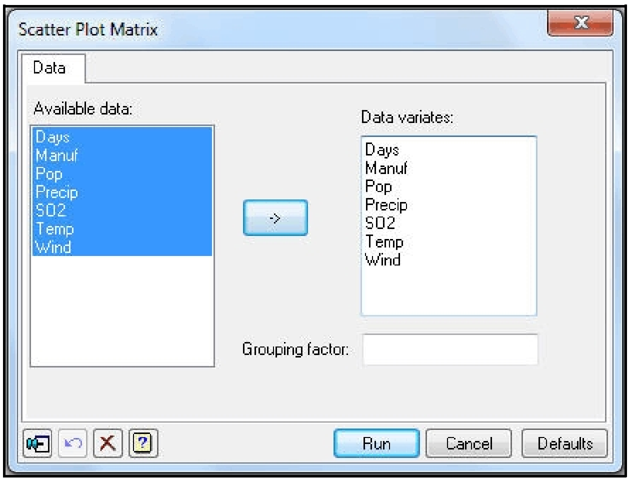


Figure 1.17

The main thing to notice in the plots and statistics show that the variables are on very different scales – and we will need to remember this in the later multivariate analyses.

To study the inter-relationships between the variates, we can select the Scatter Plot Matrix option of the Graphics menu on the menu bar. In the resulting menu (Figure 1.18), we just need to select the variates to plot and click on Run. The plot is useful for showing variates that are positively and negatively correlated, extreme observations and any clusters of the units.

The pollution data are plotted in Figure 1.19).


Scatter Plot Matrix

Data

Available data:
Days
Manuf
Pop
Precip
SO2
Temp
Wind

Data variates:
Days
Manuf
Pop
Precip
SO2
Temp
Wind

Grouping factor:

Figure 1.18

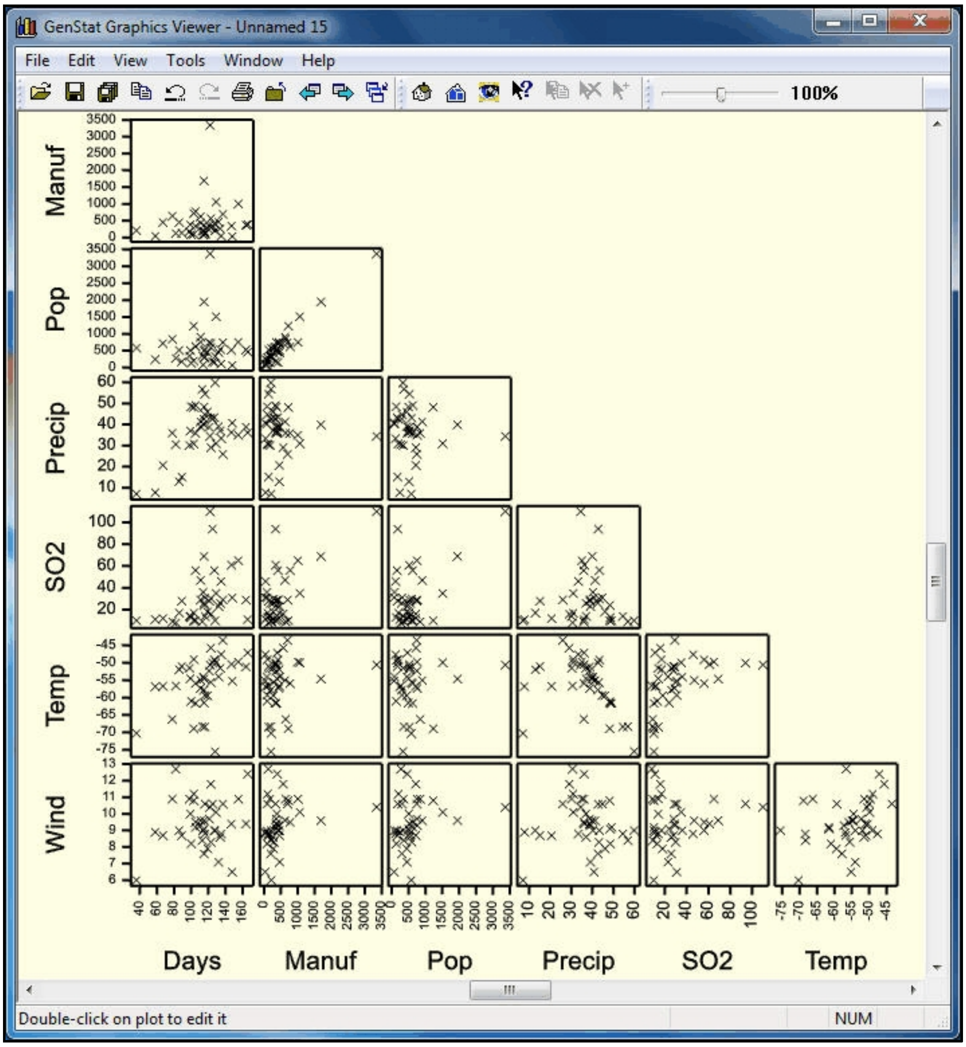
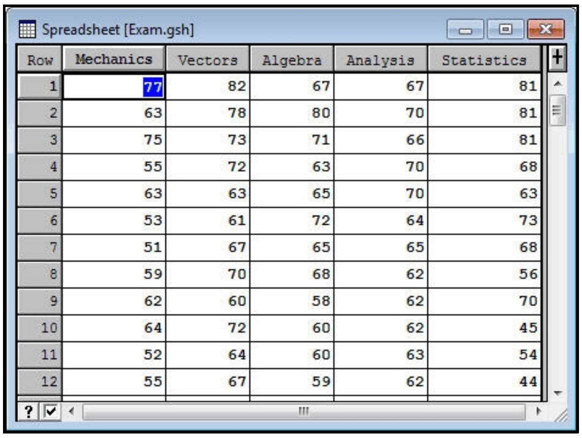


Figure 1.19

1.1 Practical

Genstat spreadsheet file `Exam.gsh` (Figure 1.20) contains examination marks for 88 students in the subjects Mechanics, Vectors, Algebra, Analysis and Statistics. (For details, see Mardia, Kent and Bibby, 1979, *Multivariate Analysis*, Academic Press, London.) Print summary statistics and plot graphs to study the data.

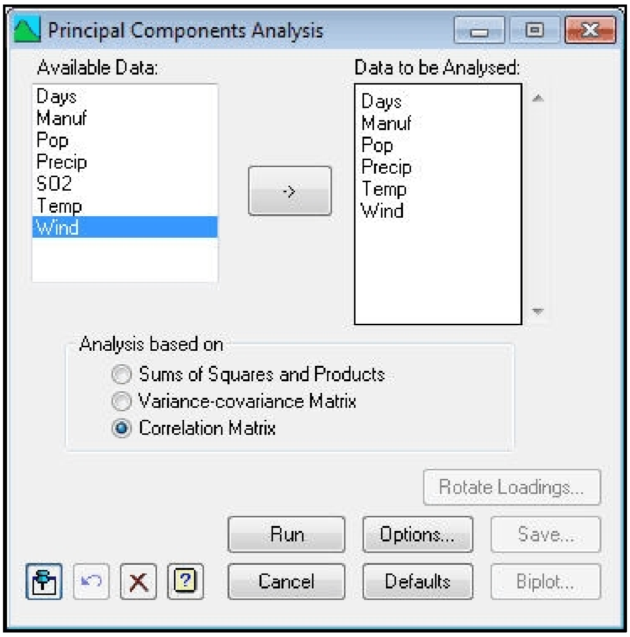


Row	Mechanics	Vectors	Algebra	Analysis	Statistics
1	77	82	67	67	81
2	63	78	80	70	81
3	75	73	71	66	81
4	55	72	63	70	68
5	63	63	65	70	63
6	53	61	72	64	73
7	51	67	65	65	68
8	59	70	68	62	56
9	62	60	58	62	70
10	64	72	60	62	45
11	52	64	60	63	54
12	55	67	59	62	44

Figure 1.20

2 Principal components analysis

A major problem with multivariate data is that there are generally too many variates for you to be able to visualise the properties and inter-relationships of the data units easily. Principal components analysis (or PCP) provides one way to overcome this "curse of dimensionality". It aims to find linear combinations of the data variates that contain most of the variation between the units. The combinations (or *principal components*) indicate relationships between the variates, and also define planes in multi-dimensional space where the relationships between the units can be studied effectively. We shall illustrate this using the pollution data from Chapter 1.



Principal Components Analysis

Available Data:
Days
Manuf
Pop
Precip
SO2
Temp
Wind

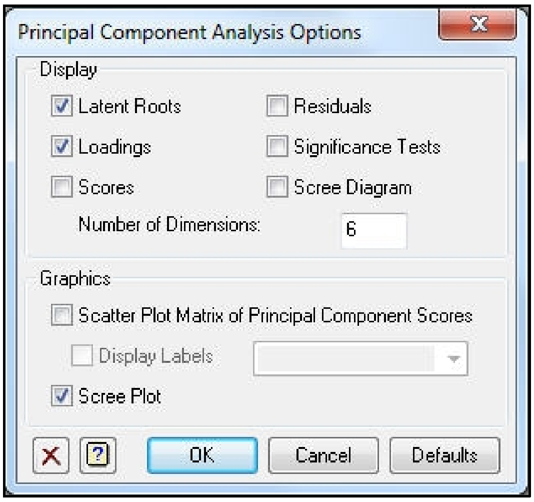
Data to be Analysed:
Days
Manuf
Pop
Precip
Temp
Wind

Analysis based on
○ Sums of Squares and Products
○ Variance-covariance Matrix
● Correlation Matrix

Figure 2.1

The Principal Components Analysis menu (Figure 2.1) is obtained by clicking on the Principal Components line in the Multivariate Analysis section of the Stats menu on the menu bar. You first need to enter the data variates into the Data to be Analysed window. Here we have chosen to enter all of the variates except SO2, which we will be treating as a response variate later in this Guide.

One important issue is to decide whether to base the analysis on sums of squares and products, or variances and covariances or correlations. The first two produce essentially the same analysis (there is just a common scaling of $\sqrt{(n-1)}$ applied to the variates, to convert from sums of squares to variances). The final setting, Correlation Matrix standardizes each variate (by subtracting its mean and dividing by its standard deviation). This can be very useful if the variates do not share a common scale and show very different amounts of variation.



Principal Component Analysis Options

Display
☑ Latent Roots
□ Residuals
☑ Loadings
□ Significance Tests
□ Scores
□ Scree Diagram
Number of Dimensions: 6

Graphics
□ Scatter Plot Matrix of Principal Component Scores
□ Display Labels
☑ Scree Plot

Figure 2.2

In the pollution data set, the variates are not only on different scales (see Chapter1), they are of inherently different types. So we have chosen to use the correlation matrix (which Genstat will calculate for us automatically, from the variates).

Clicking on the Options button produces the Principal Components Analysis Options menu (Figure 2.2), which controls the printed output from the analysis. We have set Display box to print Latent Roots and

Loadings, we have requested a scree plot, and we have set the Number of Dimensions box to 6 which will give all the available latent roots and vectors. If you choose to have less than the full number of dimensions, the Residuals check box can print residuals representing the information in the dimensions that have been excluded. The Number of Dimensions setting also applies to results saved from the Principal Components Save Options menu, which is obtained by clicking on the Save button on the Principal Components Analysis menu.

The output is shown below.

```
Principal components analysis

Latent roots

            1         2         3         4         5         6
        2.196     1.500     1.395     0.760     0.115     0.034

Percentage variation

            1         2         3         4         5         6
        36.60     25.00     23.24     12.67      1.91      0.57

Trace

        6.000

Latent vectors (loadings)

                      1           2           3           4           5

       Days     0.23792     0.70777     0.09309     0.31131     0.58000
      Manuf     0.61154    -0.16806    -0.27289     0.13684    -0.10204
        Pop     0.57782    -0.22245    -0.35037     0.07248     0.07807
     Precip    -0.04081     0.62286    -0.50456    -0.17115    -0.56818
       Temp     0.32965     0.12760     0.67169     0.30646    -0.55806
       Wind     0.35384     0.13079     0.29725    -0.86943     0.11327

                      6

       Days     0.02196
      Manuf     0.70297
        Pop    -0.69464
     Precip    -0.06062
       Temp    -0.13619
       Wind     0.02453
```

The first principal component defines the direction in which the towns exhibit the greatest variation. The second component defines the direction with the greatest variation of the directions orthogonal to the first component. The third component defines the direction with the greatest variation of the directions orthogonal to the first two components, and so on. Here, the first component contains about 37% of the variation, and the first and second components contain about 62%.

It is often interesting to interpret the directions. Those with mainly positive or mainly negative loadings represent "averages" while those with a mixture of signs represent "comparisons". Here, the first component is in the direction

0.23792 × `Days` + 0.61154 × `Manuf` + 0.57782 × `Pop`
– 0.04081 × `Precip` + 0.32965 × `Temp` + 0.35384 × `Wind`

and seems to represent "quality of life". The second component is

0.23792 × `Days` + 0.61154 × `Manuf` + 0.57782 × `Pop`
– 0.04081 × `Precip` + 0.32965 × `Temp` + 0.35384 × `Wind`

and seems to be related to the wetness of the climate.

We have not printed the significance tests for equality of the final *K* roots as these cannot be used when the analysis is based on correlations. When the analysis is based on variances or on sums of squares, they can be useful for deciding how many roots are needed. Asymptotically (that is, as the number of units becomes large) these have chi-square distributions. However, this is not true for analyses based on correlations. To use the tests, we start by testing for equality of all the roots, then all except the first, all except the first and second, and so on, until the test is non-significant. The rationale is that, if we are to omit the final dimension, we should also omit all dimensions that are no more variable than that dimension.

An alternative, visual way of deciding how many roots are needed is to examine the scree plot. The plot for the pollution data, shown in Figure 2.3, shows the pattern that you would hope to find, with a clear jump up from the final roots (with low eigenvalues) to the earlier roots (with larger eigenvalues). This is more in line with the attitude that significance tests are not really relevant if you view principal components analysis mainly as a descriptive technique, where the aim is to find dimensions in which you can most effectively study the inter-relationships of the data units.

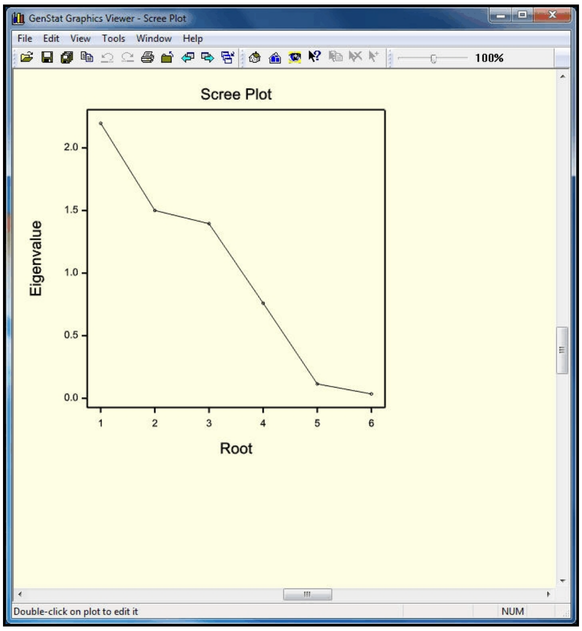
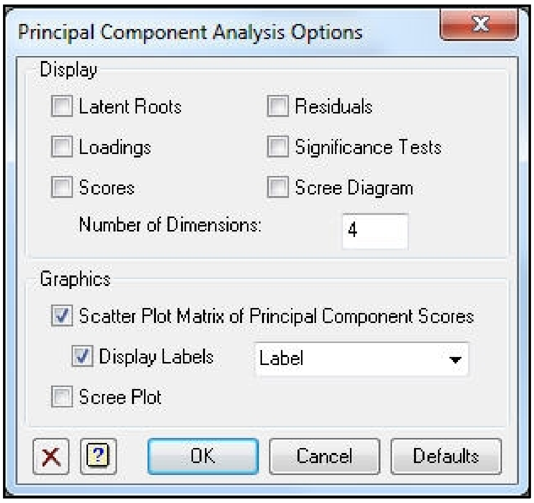

GenStat Graphics Viewer - Scree Plot
File Edit View Tools Window Help
100%
Scree Plot
Eigenvalue
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
Root
Double-click on plot to edit it
NUM

Figure 2.3

Here it seems that four principal components are needed. So, we change the number of dimensions in the Principal Components Analysis Options menu to four, check the box to plot the principal component scores, and select `Label` to label the points. Clicking on OK here, and then Run in the Principal Components Analysis menu, produces the plot in Figure 2.5.



Principal Component Analysis Options

Display

□ Latent Roots □ Residuals

□ Loadings □ Significance Tests

□ Scores □ Scree Diagram

Number of Dimensions: 4

Graphics

☑ Scatter Plot Matrix of Principal Component Scores

☑ Display Labels Label

□ Scree Plot

Figure 2.4

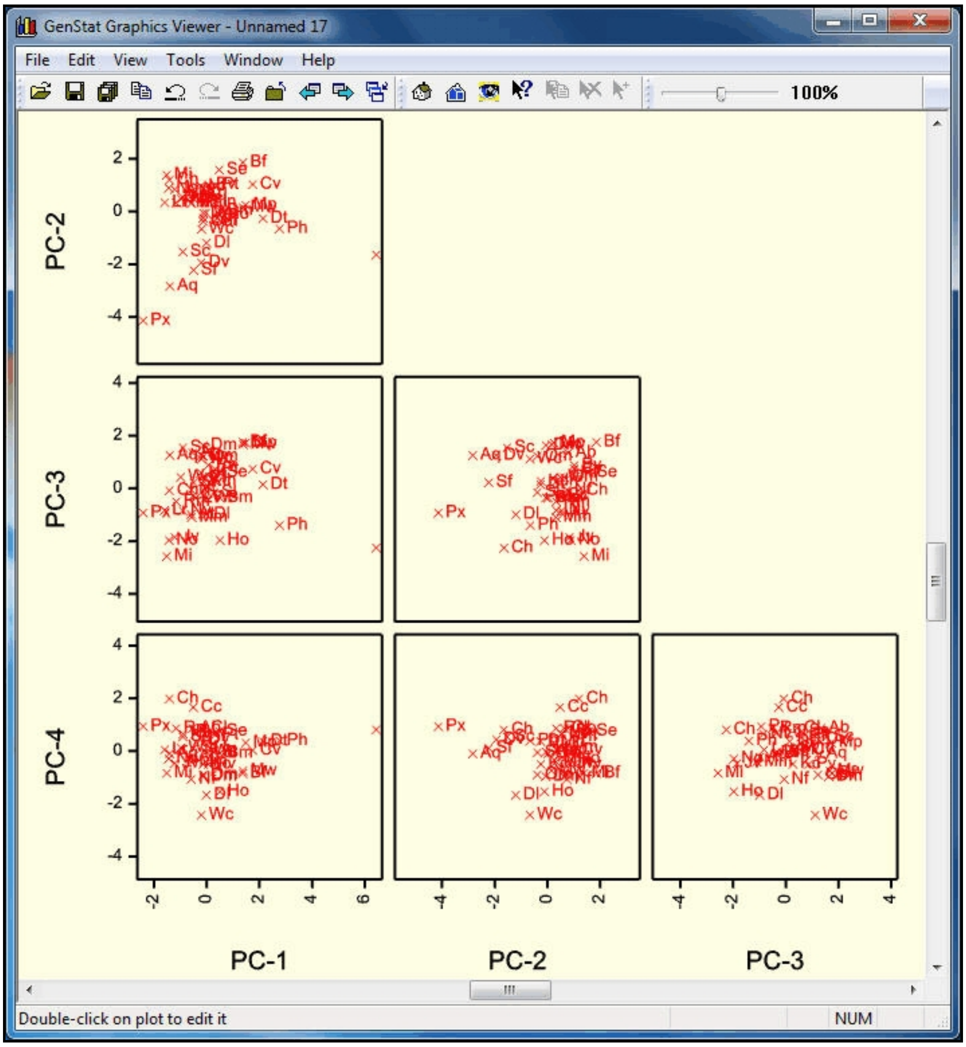


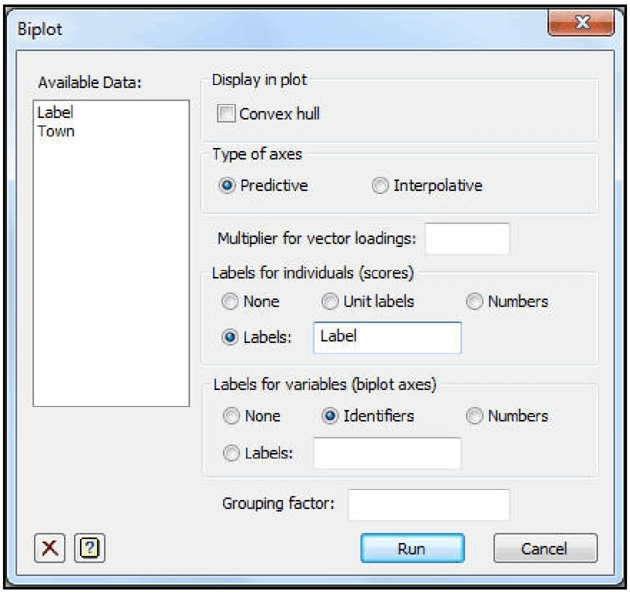
Figure 2.5

The menu uses the `PCP` directive, which is described in Section 6.2.1 of the *Guide to the Genstat Command Language, Part 2 Statistics*.

2.1 Principal-component biplot

Several of the multivariate analysis menus have a Biplot button, that becomes available once the analysis has been run. Biplots provide a convenient way of assessing the relationships between the individual observations in the analysis (here the towns), and their characteristics with respect to the variables in the data.

Clicking on the button in the Principal Components Analysis menu opens the menu in Figure 2.6, which has options to control the labelling of the plot, and the way in which the variables are represented. Here we enter the text vector `Label` to provide labels, and click on Run to produce the graph in Figure 2.7.



Available Data:
Label
Town

Display in plot
☐ Convex hull

Type of axes
● Predictive ○ Interpolative

Multiplier for vector loadings:

Labels for individuals (scores)
○ None ○ Unit labels ○ Numbers
● Labels: Label

Labels for variables (biplot axes)
○ None ● Identifiers ○ Numbers
○ Labels:

Grouping factor:

Figure 2.6

The display plots the individuals in the space defined by the first two dimensions of the multivariate analysis (from a PCP these will be the first two principal components). The plot also contains an "axis" for each variable (its *biplot* axis) that allows you to see how each individual's projection into this plane relates to its value for the variable concerned. Figure 2.7 shows the default, predictive axes. These show the values of the variables that are predicted by the projection into 2-dimensions that is defined for each point by the analysis; essentially this is done by taking an orthogonal projection of the point onto each the biplot axis. Genstat defines a *hot point* at the point for each individual. If you click on the hot point icon at the left-hand end of the Graphics Toolbar, and then click on one of the points, lines will be drawn from the point to the predictions. In Figure 2.7, we have done this for Phoenix, so you can see how this differs from the other towns. The lines can be removed again by clicking on the hot point a second time.

The angles between the biplot axes represent the correlations between the variables, and lines in opposite directions indicate negative correlation. So here we can see that temperature and wind have a strong negative correlation. The % variance of the principal components show the extent to which the plot summarizes the entire data set.

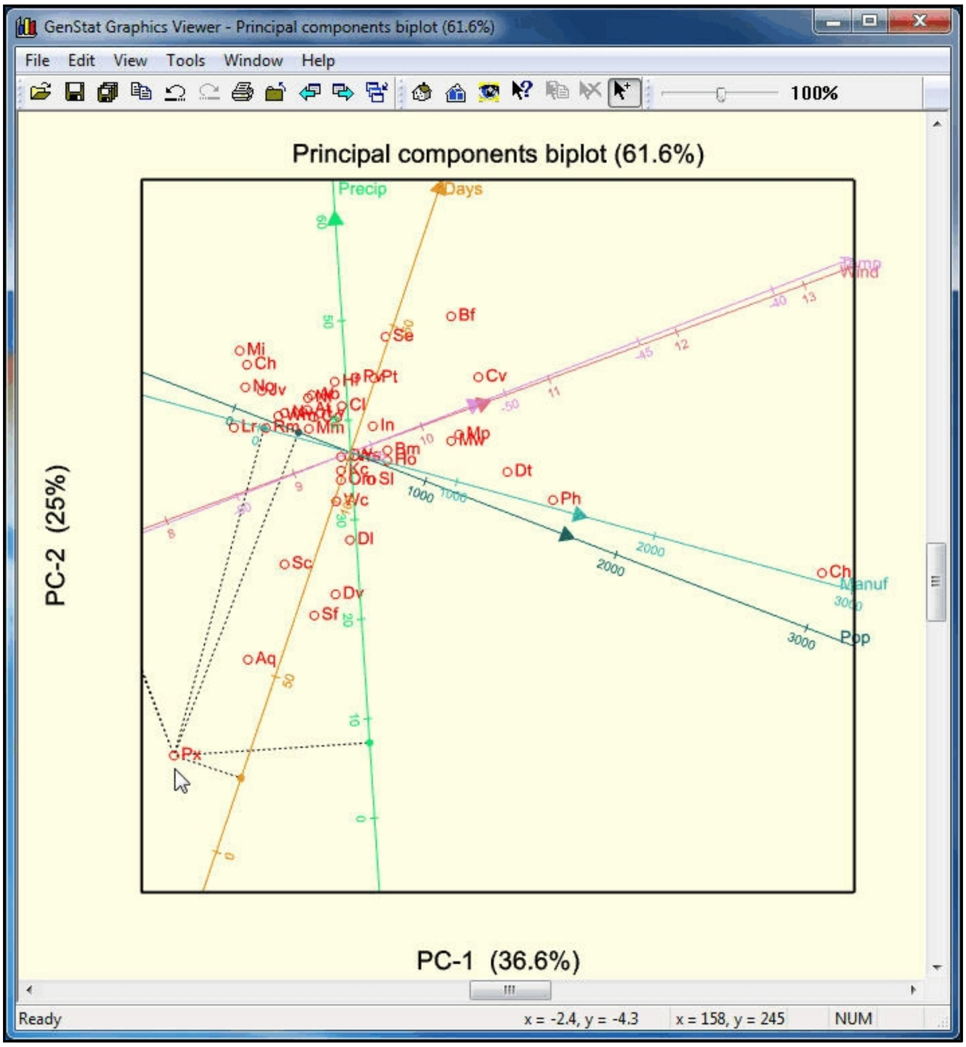

GenStat Graphics Viewer - Principal components biplot (61.6%)
File Edit View Tools Window Help
100%
Principal components biplot (61.6%)
Precip
Days
Temp
Wind
Manuf
Pop
PC-1 (36.6%)
PC-2 (25%)
Ready
x = -2.4, y = -4.3
x = 158, y = 245
NUM

Figure 2.7

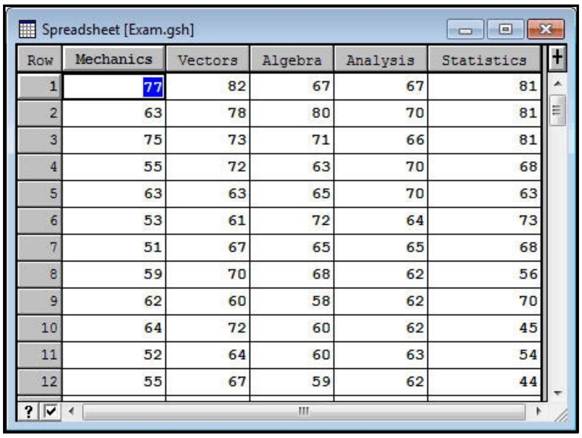
The alternative, interpolative axes show the values of the variables that would lead to a point being placed at the position of the selected point on the graph. So here the point is being predicted by the variables, rather than the variables by the point. This is done by taking the sum of a set of vectors, one in the direction of each variable, with lengths equal to the values of the variables for that point. To obtain interpolative axes, you should select the Interpolative button from the Type of Axes radio buttons in the Biplot menu.

The Biplot menu uses the `DBIPLOT` procedure, which is described in Section 6.16.1 of the *Guide to the Genstat Command Language, Part 2 Statistics*.

2.2 Practical

Genstat spreadsheet file `Exam.gsh` (Figure 2.8) contains examination marks for 88 students in the subjects Mechanics, Vectors, Algebra, Analysis and Statistics. (See Practical 1.1 and Mardia, Kent and Bibby, 1979, *Multivariate Analysis*, Academic Press, London.)

Perform a principal components analysis. How would you interpret the directions in which the student marks exhibit the greatest variation? How important is Statistics in distinguishing the abilities of the students?



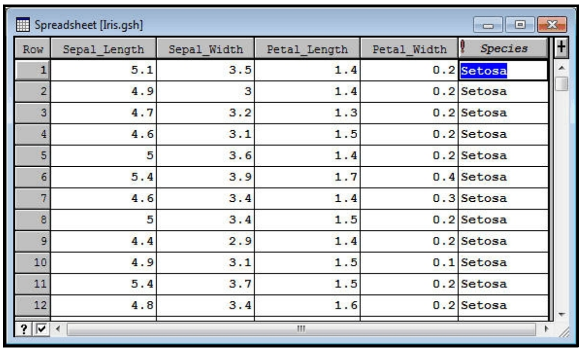
Row	Mechanics	Vectors	Algebra	Analysis	Statistics
1	77	82	67	67	81
2	63	78	80	70	81
3	75	73	71	66	81
4	55	72	63	70	68
5	63	63	65	70	63
6	53	61	72	64	73
7	51	67	65	65	68
8	59	70	68	62	56
9	62	60	58	62	70
10	64	72	60	62	45
11	52	64	60	63	54
12	55	67	59	62	44

Figure 2.8

Display a biplot from the analysis, and use the hotpoints to see how the strongest and weakest students differ from the other students.

3 Canonical variates analysis

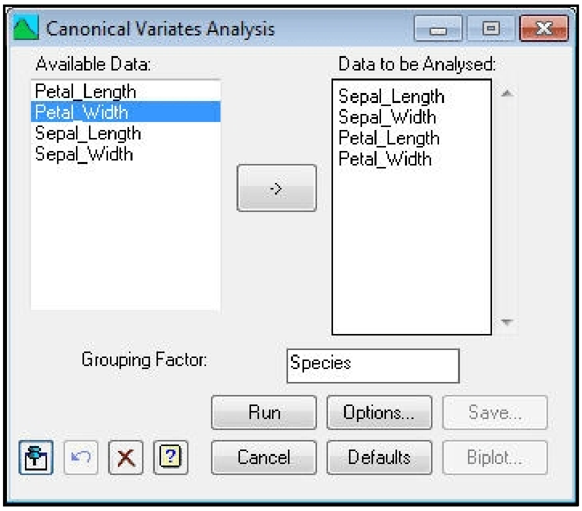
Canonical variates analysis is appropriate when the units are classified into groups. The aim is to find linear combinations of the data variates that represent most of the variation between the groups (rather than between the individual units, as in principal components analysis; Chapter 2). We illustrate the analysis using a classic data set, Fisher's Iris Data, which consists of measurements of sepal and petal lengths and widths on iris plants of three different species. This is available in Genstat spreadsheet `Iris.gsh` (Figure 3.1).



Row	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Species
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3	1.4	0.2	Setosa
3	4.7	3.2	1.3	0.2	Setosa
4	4.6	3.1	1.5	0.2	Setosa
5	5	3.6	1.4	0.2	Setosa
6	5.4	3.9	1.7	0.4	Setosa
7	4.6	3.4	1.4	0.3	Setosa
8	5	3.4	1.5	0.2	Setosa
9	4.4	2.9	1.4	0.2	Setosa
10	4.9	3.1	1.5	0.1	Setosa
11	5.4	3.7	1.5	0.2	Setosa
12	4.8	3.4	1.6	0.2	Setosa

Figure 3.1

The Canonical Variates Analysis menu (Figure 3.2) is obtained by clicking on the Canonical Variates line in the Multivariate Analysis section of the Stats menu on the menu bar. You need to enter the data variates into the Data to be Analysed window, and the factor defining the groups into the Grouping Factor window. Clicking on the Options button produces the Canonical Variates Analysis Options menu, which controls the printed output from the analysis.

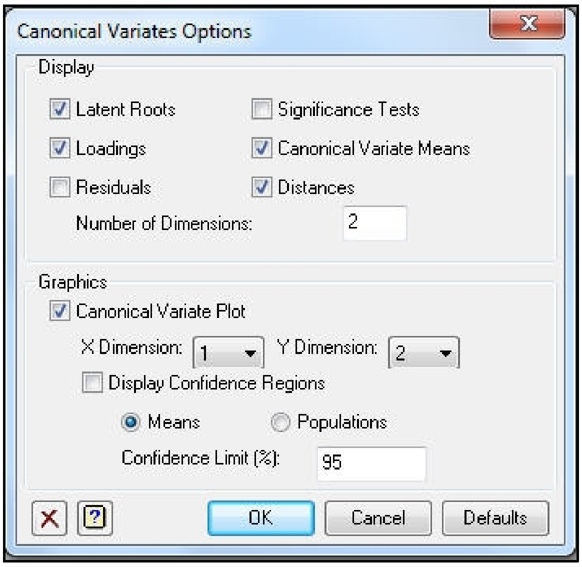


Available Data:
Petal_Length
Petal_Width
Sepal_Length
Sepal_Width

Data to be Analysed:
Sepal_Length
Sepal_Width
Petal_Length
Petal_Width

Grouping Factor: Species

Figure 3.2



Display

☑ Latent Roots
□ Significance Tests
☑ Loadings
☑ Canonical Variate Means
□ Residuals
☑ Distances

Number of Dimensions: 2

Graphics

☑ Canonical Variate Plot

X Dimension: 1
Y Dimension: 2

□ Display Confidence Regions

● Means ○ Populations

Confidence Limit (%): 95

Figure 3.3

In the Options menu (Figure 3.3), we have set the Display box to print Latent Roots, Loadings, Canonical Variate Means and Distances. The Number of Dimensions box is set to 2, which is the maximum possible here as there are only three species of iris in the data set. If you choose to have less than the full number of dimensions, the Residuals check box can print residuals representing the information in the dimensions that have been excluded. The Number of Dimensions setting also applies to results saved from the Canonical Variates Save Options menu, which is obtained by clicking on the Save button on the

Canonical Variates Analysis menu. The Graphics section of the menu is set to plot the data with the first canonical variate along the x-axis, and the second along the y-axis.

The output from the analysis is shown below.

Canonical variate analysis

Latent roots

```
              1              2
          32.19           0.29
```

Percentage variation

```
              1              2
          99.12           0.88
```

Trace

```
          32.48
```

Latent vectors (loadings)

```
                               1              2
               1           0.829          0.024
               2           1.534          2.165
               3          -2.201         -0.932
               4          -2.810          2.839
```

Canonical variate means

```
                               1              2
               1           7.608          0.215
               2          -1.825         -0.728
               3          -5.783          0.513
```

Adjustment terms

```
                               1              2
               1          -2.105          6.661
```

Inter-group distances

```
               1           0.000
               2           9.480          0.000
               3          13.393          4.147          0.000
```

```
              1              2              3
```

The results show that 99% of the between-group variation is in the direction of the first canonical variate:

0.829 × Sepal-Length + 1.534 × Sepal-Width − 2.201 × Petal-Length − 2.810 × Petal-Width

(using the coefficients in column 1 of the latent-vectors matrix). This is confirmed by the plot in Figure 3.4.

The matrix of canonical variate means presents the coordinates (or *scores*) for each group in the direction of each canonical variate. These are adjusted so that the centroid of the points, weighted by sizes of the groups, is at the origin. The adjustment term for each canonical original variates in order to achieve this. (See *Guide to the Genstat Command Language*, Part 2 Section 6.3.1 for more details.)

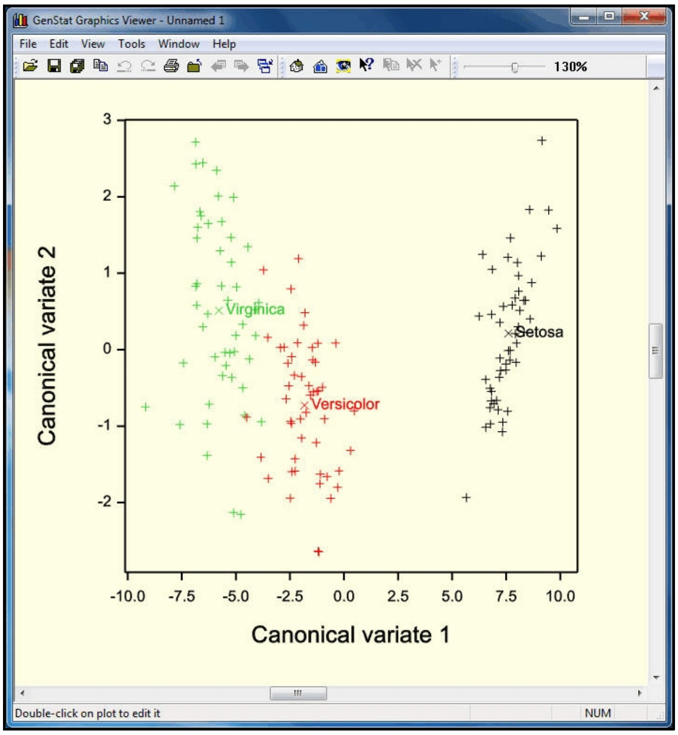
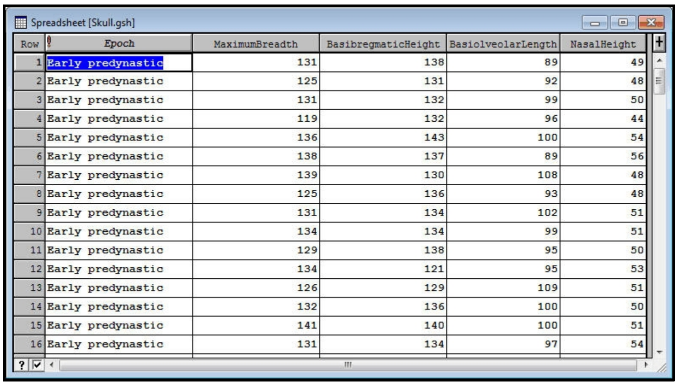


Figure 3.4

3.1 Practical

Genstat spreadsheet file `Skull.gsh` (Figure 3.5) contains data on 150 male Egyptian skulls from five different epochs (see pages 4 and 5 of Manly, 1986, *Multivariate Statistical Methods a Primer*, Chapman & Hall, London).

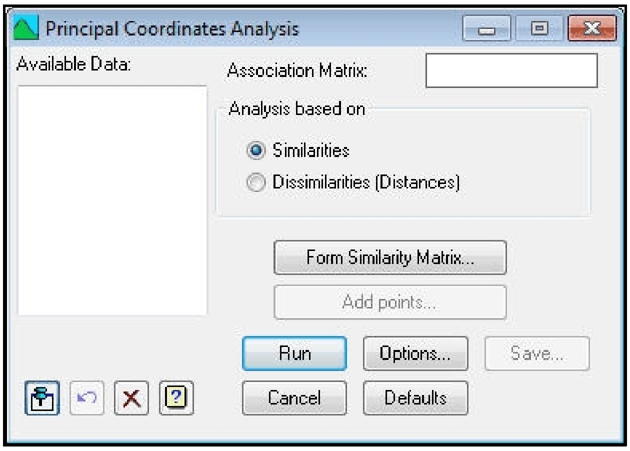
Perform a canonical variates analysis. Plot the first two canonical variates and study how the skulls differ between epochs.


Spreadsheet [Skull.gsh]

Row	Epoch	MaximumBreadth	BasibregmaticHeight	BasiolveolarLength	NasalHeight
1	Early predynastic	131	138	89	49
2	Early predynastic	125	131	92	48
3	Early predynastic	131	132	99	50
4	Early predynastic	119	132	96	44
5	Early predynastic	136	143	100	54
6	Early predynastic	138	137	89	56
7	Early predynastic	139	130	108	48
8	Early predynastic	125	136	93	48
9	Early predynastic	131	134	102	51
10	Early predynastic	134	134	99	51
11	Early predynastic	129	138	95	50
12	Early predynastic	134	121	95	53
13	Early predynastic	126	129	109	51
14	Early predynastic	132	136	100	50
15	Early predynastic	141	140	100	51
16	Early predynastic	131	134	97	54

Figure 3.5

4 Principal coordinates analysis



Analysis based on

● Similarities

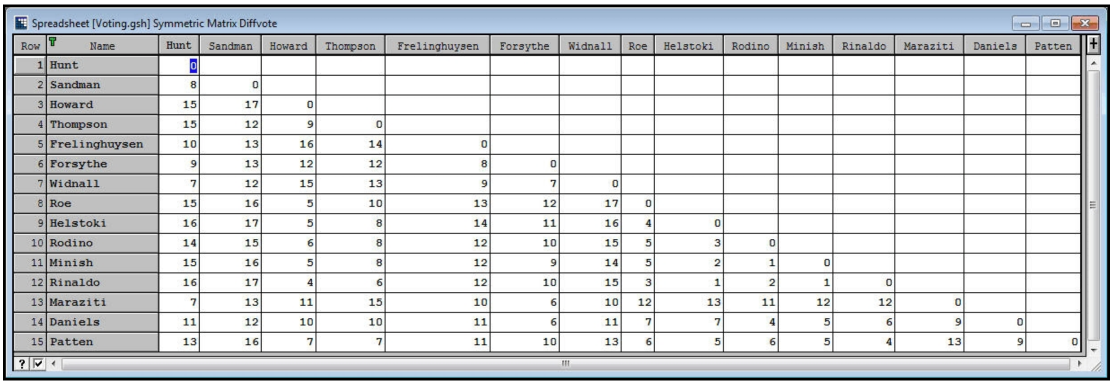
○ Dissimilarities (Distances)

Figure 4.1

Principal coordinates analysis differs from principal components and canonical variates analysis in that the focus is more on the data units than the data variables. So the basic input is a symmetric matrix representing the "associations" between the data units. The menu (Figure 4.1) has radio buttons that you can use to specify how the the associations are supplied. Similarities are on a range from zero (completely different) to one (absolutely identical). The alternative is to specify dissimilarities or distance which are zero when the two units are identical. Distances d are converted automatically to similarities s by the menu, using the transformation

$$s = -d^2$$

(See the *Guide to the Genstat Command Language*, Part 2, Section 6.10 for more information.)



Spreadsheet [Voting.gsh] Symmetric Matrix Diffvote

Row	Name	Hunt	Sandman	Howard	Thompson	Frelinghuysen	Forsythe	Widnall	Roe	Helstoki	Rodino	Minish	Rinaldo	Maraziti	Daniels	Patten
1	Hunt	0														
2	Sandman	8	0													
3	Howard	15	17	0												
4	Thompson	15	12	9	0											
5	Frelinghuysen	10	13	16	14	0										
6	Forsythe	9	13	12	12	8	0									
7	Widnall	7	12	15	13	9	7	0								
8	Roe	15	16	5	10	13	12	17	0							
9	Helstoki	16	17	5	8	14	11	16	4	0						
10	Rodino	14	15	6	8	12	10	15	5	3	0					
11	Minish	15	16	5	8	12	9	14	5	2	1	0				
12	Rinaldo	16	17	4	6	12	10	15	3	1	2	1	0			
13	Maraziti	7	13	11	15	10	6	10	12	13	11	12	12	0		
14	Daniels	11	12	10	10	11	6	11	7	7	4	5	6	9	0	
15	Patten	13	16	7	7	11	10	13	6	5	6	5	4	13	9	0

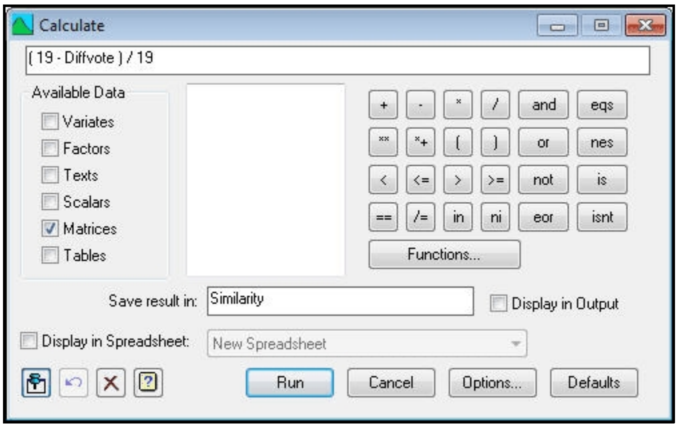
Figure 4.2

As an example, spreadsheet file `Voting.gsh` contains the number of times that 15 congressmen from New Jersey voted differently in 19 environmental bills (see Table 10.3 of Manly, 1986, *Multivariate Statistical Methods a Primer*, Chapman & Hall, London).

We shall analyse these as similarities, and so we first use the Calculations menu to convert the differences to proportions of times that they congressmen voted in the same way i.e.

```
(19-Diffvote)/19
```

(see Figure 4.3). We then enter the resulting symmetric matrix, Similarity, into the Association Matrix box in Figure 4.1.


Calculate

(19 - Diffvote) / 19

Available Data

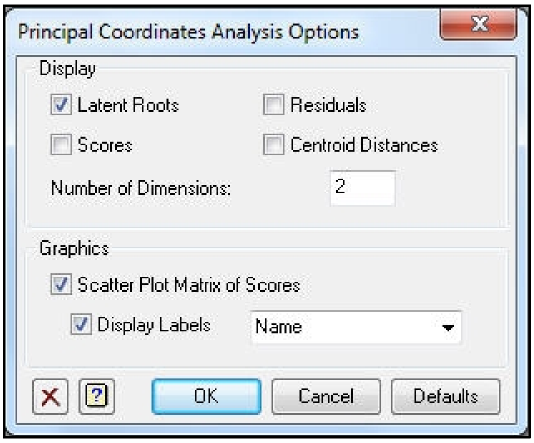
☐ Variates
☐ Factors
☐ Texts
☐ Scalars
☑ Matrices
☐ Tables

Save result in: Similarity ☐ Display in Output

☐ Display in Spreadsheet: New Spreadsheet

Figure 4.3

The Principal Coordinates Analysis Options menu (Figure 4.4) allows you to select the output to display, and specify the number of dimensions to fit. Here we have chosen to fit only two dimensions.


Principal Coordinates Analysis Options

Display

☑ Latent Roots ☐ Residuals
☐ Scores ☐ Centroid Distances

Number of Dimensions: 2

Graphics

☑ Scatter Plot Matrix of Scores
☑ Display Labels Name

Figure 4.4

Principal coordinates analysis

Latent roots

```
          1           2
      2.824       0.972
```

Percentage variation

```
          1           2
      38.22       13.15
```

Trace

```
      7.389
```

In interpreting the plot (Figure 4.5), it is interesting to note that congressmen Daniels, Helstoki, Howard, Minish, Patten, Rodino, Roe and Thompson were Democrats, while congressmen Forsythe, Frelinghuysen, Hunt, Maraziti, Rinaldo, Sandman, Widnall were Republicans.

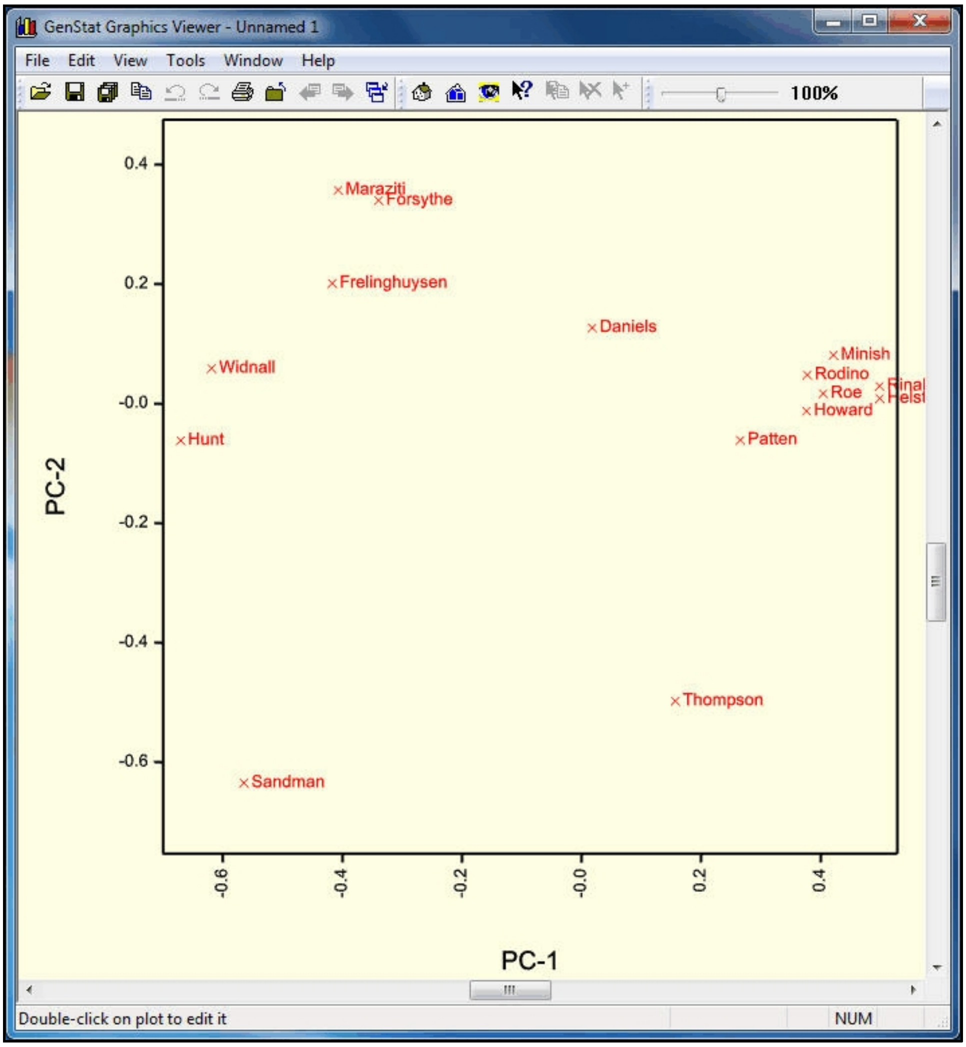
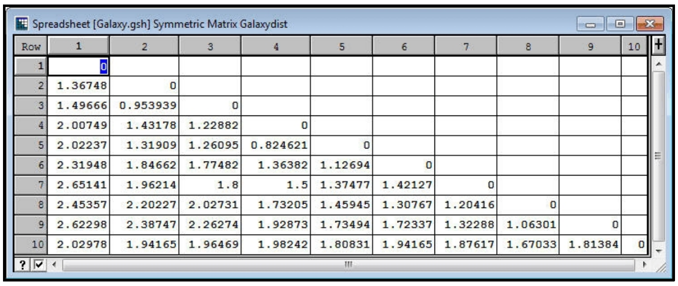


Figure 4.5

4.1 Practical

Genstat spreadsheet file `Galaxy.gsh` (Figure 4.6) contains distances between ten types of Galaxy. Use principal coordinates analysis to represent them in three dimensions.

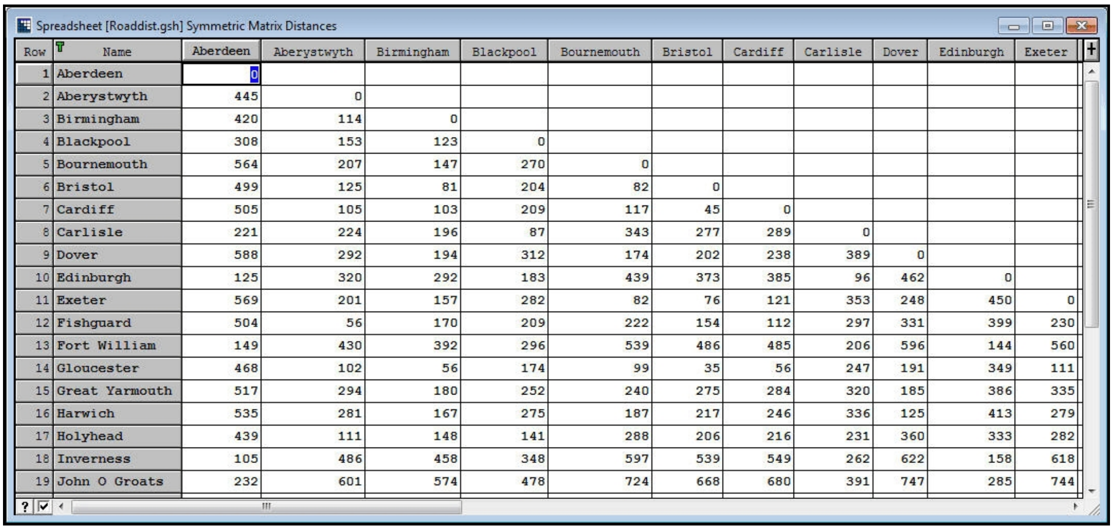


Row	1	2	3	4	5	6	7	8	9	10
1	0									
2	1.36748	0								
3	1.49666	0.953939	0							
4	2.00749	1.43178	1.22882	0						
5	2.02237	1.31909	1.26095	0.824621	0					
6	2.31948	1.84662	1.77482	1.36382	1.12694	0				
7	2.65141	1.96214	1.8	1.5	1.37477	1.42127	0			
8	2.45357	2.20227	2.02731	1.73205	1.45945	1.30767	1.20416	0		
9	2.62298	2.38747	2.26274	1.92873	1.73494	1.72337	1.32288	1.06301	0	
10	2.02978	1.94165	1.96469	1.98242	1.80831	1.94165	1.87617	1.67033	1.81384	0

Figure 4.6

5 Multidimensional scaling

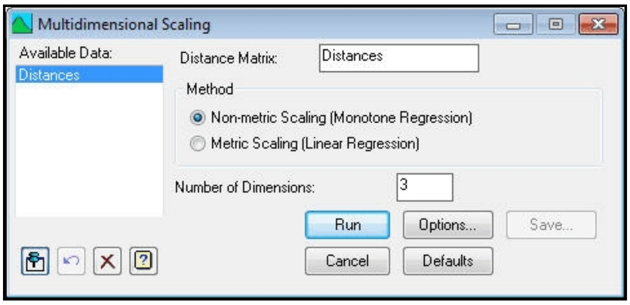
Multidimensional scaling operates on a symmetric matrix which is assumed to represent distances between a set of units. It aims to construct coordinates of points, in a defined number of dimensions, whose distances are approximately the same as those in the original matrix. To illustrate the analysis we will try to recreate the locations of some British towns, based on figures for the shortest distances between each of them by road. The data are in the Genstat spreadsheet Roaddist.gsh (Figure 5.1). This is a symmetric matrix spreadsheet (as shown by the blanks above the diagonal).



Spreadsheet [Roaddist.gsh] Symmetric Matrix Distances

Row	Name	Aberdeen	Aberystwyth	Birmingham	Blackpool	Bournemouth	Bristol	Cardiff	Carlisle	Dover	Edinburgh	Exeter
1	Aberdeen	0										
2	Aberystwyth	445	0									
3	Birmingham	420	114	0								
4	Blackpool	308	153	123	0							
5	Bournemouth	564	207	147	270	0						
6	Bristol	499	125	81	204	82	0					
7	Cardiff	505	105	103	209	117	45	0				
8	Carlisle	221	224	196	87	343	277	289	0			
9	Dover	588	292	194	312	174	202	238	389	0		
10	Edinburgh	125	320	292	183	439	373	385	96	462	0	
11	Exeter	569	201	157	282	82	76	121	353	248	450	0
12	Fishguard	504	56	170	209	222	154	112	297	331	399	230
13	Fort William	149	430	392	296	539	486	485	206	596	144	560
14	Gloucester	468	102	56	174	99	35	56	247	191	349	111
15	Great Yarmouth	517	294	180	252	240	275	284	320	185	386	335
16	Harwich	535	281	167	275	187	217	246	336	125	413	279
17	Holyhead	439	111	148	141	288	206	216	231	360	333	282
18	Inverness	105	486	458	348	597	539	549	262	622	158	618
19	John O Groats	232	601	574	478	724	668	680	391	747	285	744

Figure 5.1



Multidimensional Scaling

Available Data: Distances

Distance Matrix: Distances

Method

● Non-metric Scaling (Monotone Regression)

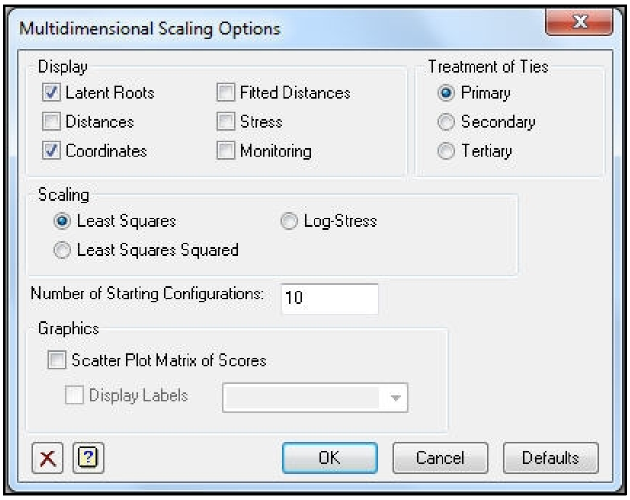
○ Metric Scaling (Linear Regression)

Number of Dimensions: 3

Figure 5.2

The Multidimensional Scaling menu is obtained by clicking on the Multidimensional Scaling line in the Multivariate Analysis section of the Stats menu. In Figure 5.2, we have entered Distances as the distance matrix to use, and set the required number of dimensions to 3.

The algorithm starts with an initial configuration of points which it then modifies using a method known as *steepest descent*, until no further improvements are possible (see the *Guide to the Genstat Command Language*, Part 2 Section 6.12). To evaluate the configuration, it does a regression of the inter-point distances, calculated from the current configuration, against the supplied distances. The Method setting on the menu controls whether this is a “monotone regression” (which corresponds to what is known as *non-metric scaling*) or an ordinary linear regression (corresponding to *metric scaling*). It then compares the fitted distances from the regression with the original distances using a quantity known as the *stress*.


Multidimensional Scaling Options

Display
☑ Latent Roots □ Fitted Distances
□ Distances □ Stress
☑ Coordinates □ Monitoring

Treatment of Ties
● Primary
○ Secondary
○ Tertiary

Scaling
● Least Squares ○ Log-Stress
○ Least Squares Squared

Number of Starting Configurations: 10

Graphics
□ Scatter Plot Matrix of Scores
□ Display Labels

Figure 5.3

The Scaling section of the Multidimensional Scaling Options menu (Figure 5.3) allows you to specify what to display from the analysis. Here we have asked to display the latent roots and the coordinates. It also controls whether the stress is calculated on a least-squares scale, a least-squares-squared scale or a logarithmic scale. The Treatment of Ties section of the options menu allows you to vary the way in which tied values in the supplied distances are treated. With the Primary setting, no restrictions are placed on the inter-point distances corresponding to tied distances. In the Secondary setting, the inter-point distances corresponding to tied distances are required to be as nearly equal as possible. The Tertiary setting is a compromise between the primary and secondary approaches to ties: the block of ties corresponding to the smallest distance are handled by the secondary method, and the remaining blocks of ties are handled by the primary method. This is particularly useful when the supplied distance matrix contains only a distinct values. Further information is given in the *Guide to the Genstat Command Language*, Part 2 Section 6.12, which describes the MDS directive that is used by these menus. The directive also has some additional facilities, for example the ability to try several automatically-generated initial configurations, or to supply your own.

If we click on OK here, and on Run in the Multidimensional Scaling menu itself, Genstat produces the output below.

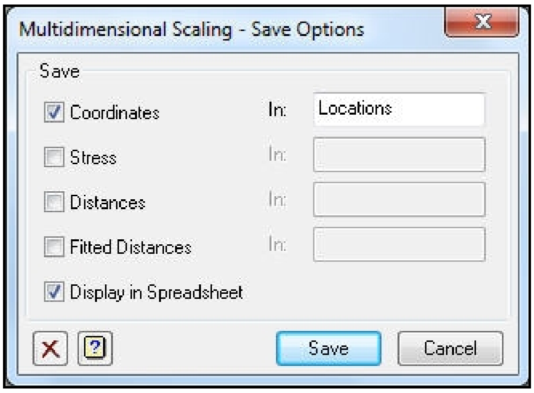
```
Message: Default seed for random number generator used with value 33633

Multidimensional scaling

Least-squares scaling criterion

Distances fitted using monotonic regression (non-metric MDS).
Primary treatment of ties.

Coordinates

                            1           2           3
          Name
      Aberdeen        -1.5498     -0.1754     -0.0455
   Aberystwyth         0.3087      0.3597     -0.3365
    Birmingham         0.3114     -0.0462     -0.0424
     Blackpool        -0.2214      0.0591     -0.1727
   Bournemouth         0.8882     -0.0407      0.2126
       Bristol         0.6224      0.1741      0.0695
```

Cardiff	0.6422	0.3121	-0.0901
Carlisle	-0.5888	0.0634	-0.0696
Dover	0.9364	-0.7146	-0.0893
Edinburgh	-1.0006	-0.0158	0.0541
Exeter	0.8894	0.3092	0.3292
Fishguard	0.5061	0.5383	-0.4924
Fort William	-1.4431	0.3719	0.2468
Gloucester	0.5245	0.0843	0.0032
Great Yarmouth	0.3566	-0.9353	-0.0210
Harwich	0.5698	-0.7701	0.1975
Holyhead	0.1291	0.2137	-0.7365
Inverness	-1.6910	0.1129	0.2851
John O Groats	-2.2088	0.1106	0.4087
Hull	-0.1110	-0.4433	-0.1190
Lands End	1.2693	0.6009	0.6692
Lincoln	0.0800	-0.3966	0.0124
Liverpool	-0.0530	0.1262	-0.1838
Newcastle	-0.5933	-0.2389	0.0496
Plymouth	1.0069	0.4264	0.4607
Portsmouth	0.9079	-0.2205	0.1607
Sheffield	0.0046	-0.1912	-0.0319
Stranraer	-0.9726	0.2432	-0.4059
Swansea	0.6916	0.3303	-0.3322
York	-0.2118	-0.2478	0.0095

Latent roots

1	23.36
2	4.15
3	2.49

The Multidimensional Scaling Options menu has a check box to plot the coordinates (or scores) in a scatter-plot matrix showing all the pairs of dimensions. If you want to plot a single pair of dimensions, you first need to save the coordinates, using the Multidimensional Scaling Save Options menu (Figure 5.4) which is obtained by clicking on the Save button on the Multidimensional Scaling menu. Here we have asked to save the coordinates in a matrix called `Locations`, and to display these in a spreadsheet.



Multidimensional Scaling - Save Options

Save

☑ Coordinates In: Locations

□ Stress In:

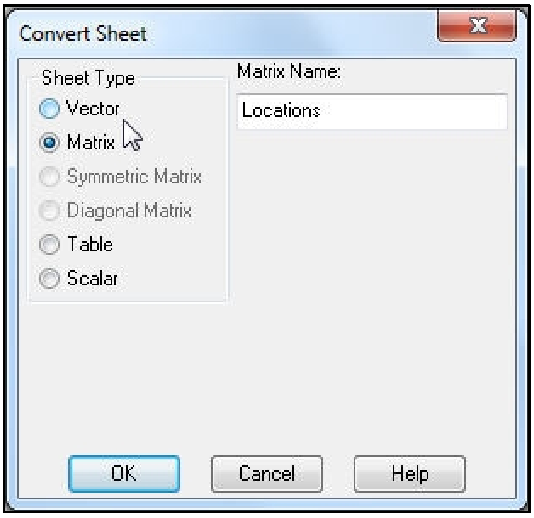
□ Distances In:

□ Fitted Distances In:

☑ Display in Spreadsheet

Figure 5.4

To plot the points, we need to convert the rectangular matrix spreadsheet of locations to a vector spreadsheet, by making this the active window and then selecting Convert in the Sheet section of the Spread menu on the menu bar. In the resulting Convert Sheet menu (Figure 5.5), we change the radio button from Matrix to Vector, and click on OK to make the change. The columns will then be become variates, probably with names `C1`, `C2` and `C3`, which can be used in the graphics menus in the usual way.

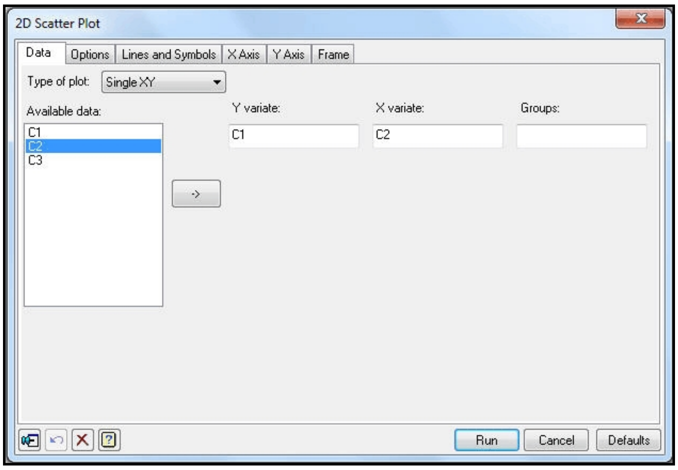


Sheet Type
○ Vector
● Matrix
○ Symmetric Matrix
○ Diagonal Matrix
○ Table
○ Scalar

Matrix Name: Locations

Figure 5.5

We can then plot the points using the 2D Scatter Plot wizard in the usual way. First we use the initial Data menu (Figure 5.6) to select `C1` for the y-coordinates, and `C2` for the x-coordinates.



Type of plot: Single XY

Available data: C1, C2, C3

Y variate: C1

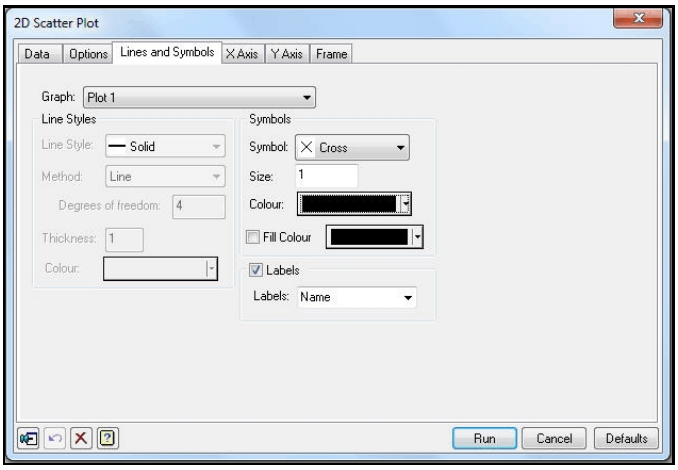
X variate: C2

Groups:

Figure 5.6

Then we select the Lines and Symbols tab of the Attributes menu (Figure 5.7), and arrange to label the points using the text `Name`. We also cancel the key on the Options tab.

Figure 5.8 shows the resulting plot of the first two dimensions, and Figure 5.9 shows a similar plot of the second dimension against the third dimension (showing some of the distortion in the data from a 2-dimensional solution).



Graph: Plot 1

Line Styles

Line Style: Solid

Method: Line

Degrees of freedom: 4

Thickness: 1

Colour:

Symbols

Symbol: Cross

Size: 1

Colour:

☐ Fill Colour

☑ Labels

Labels: Name

Figure 5.7

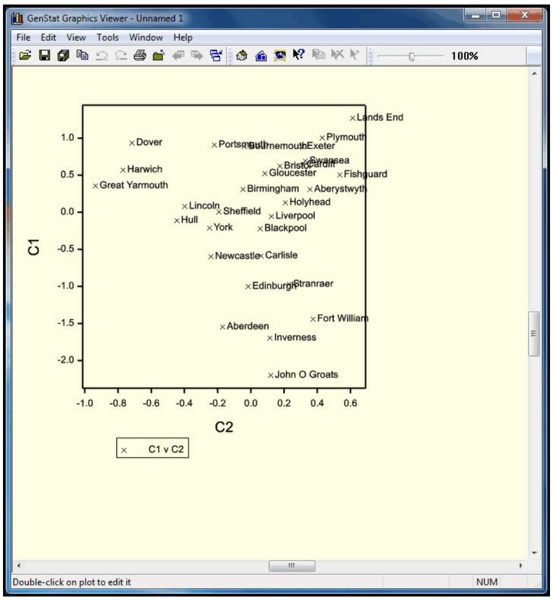


Figure 5.8

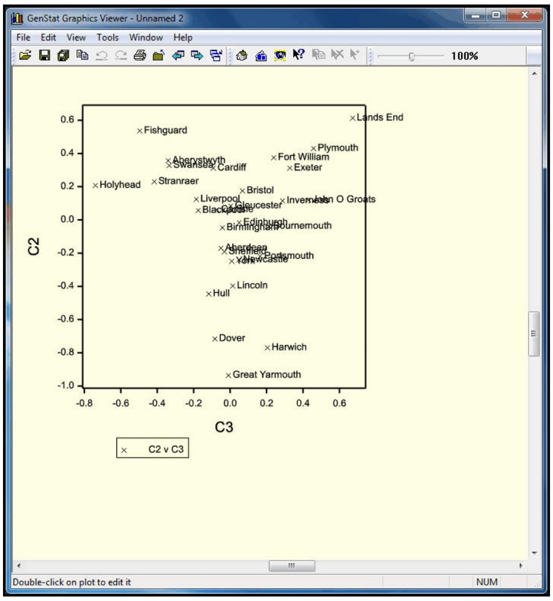
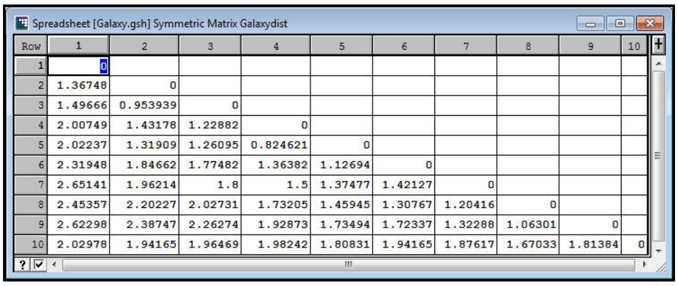


Figure 5.9

5.1 Practical

Genstat spreadsheet file Galaxy.gsh (Figure 5.10) contains distances between ten types of Galaxy. Use multidimensional scaling to represent them in three dimensions.



Spreadsheet [Galaxy.gsh] Symmetric Matrix Galaxydist

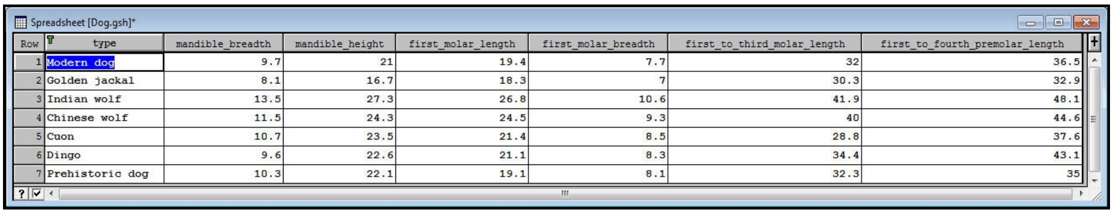
Row	1	2	3	4	5	6	7	8	9	10
1	0									
2	1.36748	0								
3	1.49666	0.953939	0							
4	2.00749	1.43178	1.22882	0						
5	2.02237	1.31909	1.26095	0.824621	0					
6	2.31948	1.84662	1.77482	1.36382	1.12694	0				
7	2.65141	1.96214	1.8	1.5	1.37477	1.42127	0			
8	2.45357	2.20227	2.02731	1.73205	1.45945	1.30767	1.20416	0		
9	2.62298	2.38747	2.26274	1.92873	1.73494	1.72337	1.32288	1.06301	0	
10	2.02978	1.94165	1.96469	1.98242	1.80831	1.94165	1.87617	1.67033	1.81384	0

Figure 5.10

6 Hierarchical cluster analysis

The hierarchical cluster analysis facilities in Genstat provide ways of grouping *n* objects into classes according to their similarity. It starts with a set of *n* clusters (or groups), each containing a single object. These initial clusters are successively merged into larger clusters, according to their similarity, until there is just one cluster (containing all the objects).

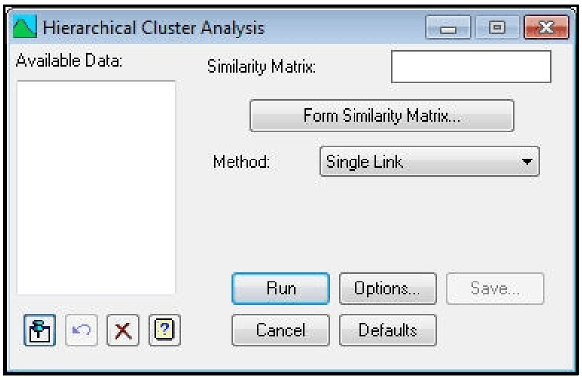
We shall use a set of data concerning mean mandible measurements of various types of modern and prehistoric dog (Higham, Kijngam & Manly, 1980, An analysis of prehistoric canid remains from Thailand, *Journal of Archaeological Science*, 7, 149-165). This data set is also discussed by Manly (1986, *Multivariate Statistical Methods a Primer*, Chapman & Hall, London). The data are in the Genstat spreadsheet `Dog.gsh` (Figure 6.1).


Spreadsheet [Dog.gsh]*

Row	type	mandible_breadth	mandible_height	first_molar_length	first_molar_breadth	first_to_third_molar_length	first_to_fourth_premolar_length
1	Modern dog	9.7	21	19.4	7.7	32	36.5
2	Golden jackal	8.1	16.7	18.3	7	30.3	32.9
3	Indian wolf	13.5	27.3	26.8	10.6	41.9	48.1
4	Chinese wolf	11.5	24.3	24.5	9.3	40	44.6
5	Cuon	10.7	23.5	21.4	8.5	28.8	37.6
6	Dingo	9.6	22.6	21.1	8.3	34.4	43.1
7	Prehistoric dog	10.3	22.1	19.1	8.1	32.3	35

Figure 6.1

The Hierarchical Cluster Analysis menu is obtained by clicking on the Hierarchical line in the Cluster Analysis subsection in the Multivariate Analysis section of the Stats menu on the menu bar. If you have already formed a similarity matrix, you can enter its name straight into the Similarity Matrix field in the menu (Figure 6.2).

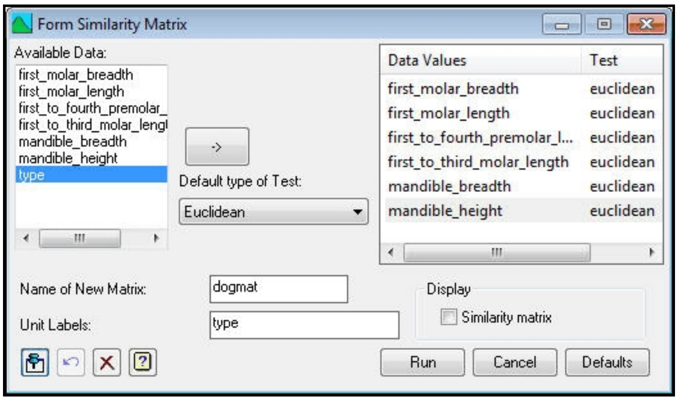

Hierarchical Cluster Analysis

Similarity Matrix:

Method: Single Link

Figure 6.2

Alternatively, you can click on the Form Similarity Matrix button and use the Form Similarity Matrix menu (Figure 6.3). The names of the variates need to be entered into the Data Values window, and you need to define a name (here `dogmat`) for the resulting symmetric matrix. You must also select the way in which the similarities are to be calculated from each variate. Here we have chosen the default


Form Similarity Matrix

Available Data:
first_molar_breadth
first_molar_length
first_to_fourth_premolar_
first_to_third_molar_lengt
mandible_breadth
mandible_height
type

Data Values	Test
first_molar_breadth	euclidean
first_molar_length	euclidean
first_to_fourth_premolar_l...	euclidean
first_to_third_molar_length	euclidean
mandible_breadth	euclidean
mandible_height	euclidean

Default type of Test: Euclidean

Name of New Matrix: dogmat

Unit Labels: type

Display
□ Similarity matrix

Figure 6.3

type to be “euclidean”, which uses the geometric distance between the points representing each pair of objects. (A formal definition of this, and the other possibilities is in the *Guide to the Genstat Command Language*, Part 2, Section 6.1.2, where it describes the METHOD option of the FSIMILARITY directive, used by the menu.) You can change the types of individual data variables by double-clicking on their lines in the right-hand box. Finally, you can specify a vector (here the text called type from the first column of the spreadsheet) to label rows and columns of the matrix. When you click Run, the name of the matrix is automatically entered into the Similarity Matrix field in the Hierarchical Cluster Analysis menu.

The Method field in the Hierarchical Cluster Analysis menu (Figure 6.2) contains a drop-down list box to specify the method of clustering to use. These differ according to the way in which they define the similarity between clusters containing more than one object:

Single Link defines the similarity to be the maximum similarity between any pair of objects (taken one from each cluster);

Nearest Neighbour is a synonym for Single Link;

Complete Link defines the similarity between two clusters as the minimum similarity between any pair of objects;

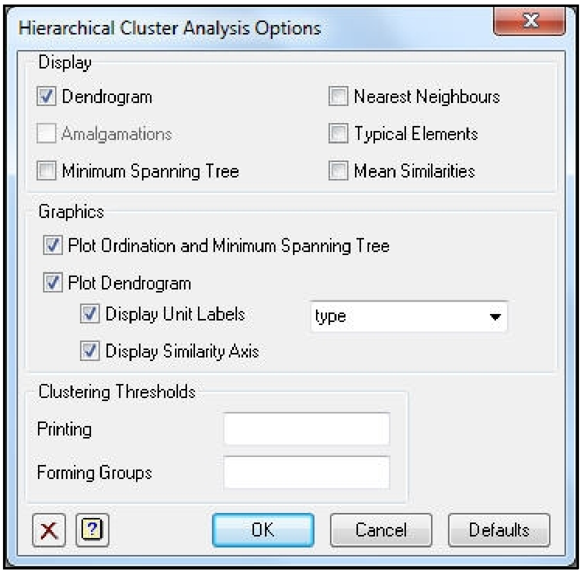
Furthest Neighbour is a synonym for Complete Link;

Average Link defines the similarity, between a cluster and a new cluster formed by merging two clusters, as the average of the similarities with each of the original clusters;

Group Average is similar to Average Link, except that the average is over all the objects in the two merging clusters;

Median Sorting if we regard the clusters as points in a multidimensional space, when two clusters join the new cluster is represented by the midpoint of the original cluster points.

Output from the analysis is controlled by the Hierarchical Cluster Analysis Options menu (Figure 6.4). For the dog example, we will simply print, and plot, the dendrogram. This displays the points at which the various clusters combine, allowing you to assess the relationships between the objects. If you specify a threshold in the Forming Groups field of the options menu, Genstat will form a factor grouping all the objects that have been combined into a single cluster at that level of similarity. You can arrange to save the factor using the Hierarchical Cluster Analysis Save Options menu, obtained by clicking on the Save button in the Hierarchical Cluster Analysis menu (Figure 6.2).



☑ Dendrogram
□ Amalgamations
□ Minimum Spanning Tree
□ Nearest Neighbours
□ Typical Elements
□ Mean Similarities
☑ Plot Ordination and Minimum Spanning Tree
☑ Plot Dendrogram
☑ Display Unit Labels: type
☑ Display Similarity Axis
Printing:
Forming Groups:

Figure 6.4

Single linkage cluster analysis

Dendrogram

```
       ** Levels   100.0  90.0

Modern dog          1  ..
Prehistoric dog     7  ..)
Cuon                5  ..)..
Dingo               6  .....)
Golden jackal       2  .....)
Chinese wolf        4  .....)
Indian wolf         3  .....)...........
```

The dendrogram for the dogs, printed above and plotted (with better resolution) in Figure 6.5, shows that the modern and prehistoric dogs are most closely related, and that both of these are related to the Cuon and to the Dingo and Golden jackal. The Indian and Chinese wolves are related to each other more than any of the other dogs, but the similarity is not close.

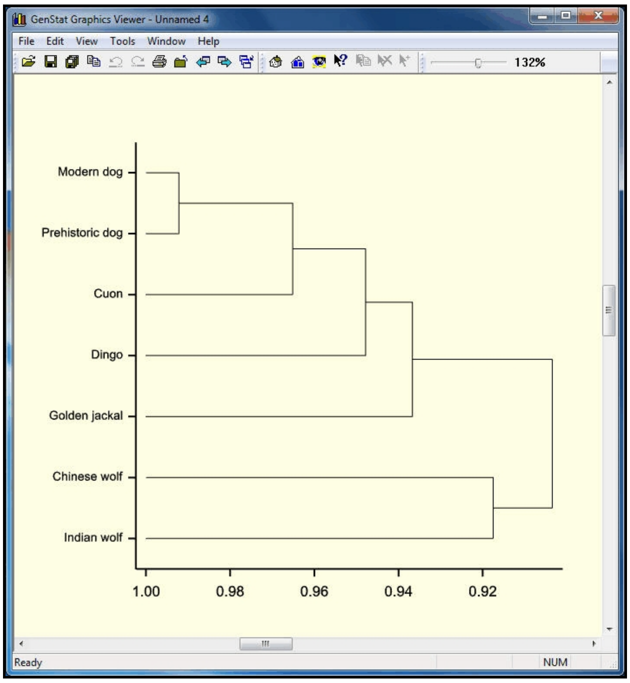
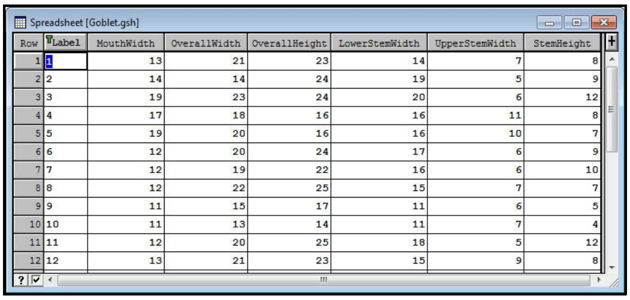


Figure 6.5

6.1 Practical

Genstat spreadsheet file Goblet.gsh (Figure 6.6) contains data on 25 goblets from prehistoric sites in Thailand (see page 147 of Manly, 1986, *Multivariate Statistical Methods a Primer*, Chapman & Hall, London). Perform a principal components analysis to study the relationships between the goblets. Then perform a cluster analysis of the goblets. How does the dendrogram reflect the closeness of the goblets in the principal-component plot?


Spreadsheet [Goblet.gsh]

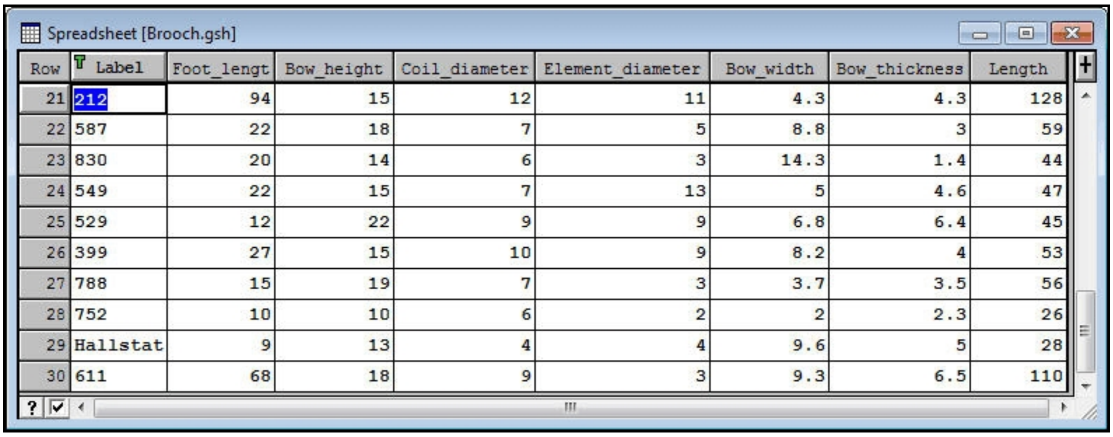
Row	Label	MouthWidth	OverallWidth	OverallHeight	LowerStemWidth	UpperStemWidth	StemHeight
1	1	13	21	23	14	7	8
2	2	14	14	24	19	5	9
3	3	19	23	24	20	6	12
4	4	17	18	16	16	11	8
5	5	19	20	16	16	10	7
6	6	12	20	24	17	6	9
7	7	12	19	22	16	6	10
8	8	12	22	25	15	7	7
9	9	11	15	17	11	6	5
10	10	11	13	14	11	7	4
11	11	12	20	25	18	5	12
12	12	13	21	23	15	9	8

Figure 6.6

7 Non-hierarchical cluster analysis

Non-hierarchical cluster analysis aims to find a single grouping of a set of *n* objects by optimizing a criterion, for example by maximizing the between-group sum of squares. Other criteria in Genstat include maximizing the total between-groups Mahalanobis distance, minimizing the within-class dispersion or a criterion known as *maximal predictive classification*, which is designed specifically for binary data. For full definitions, see the *Guide to the Genstat Command Language*, Part 2 Section 6.20. This form of clustering includes the technique known as *K-means clustering*, where the criterion is usually the within-class dispersion.

To illustrate the menus we shall use some measurements taken on 30 bronze brooches (Doran & Hodson, 1975, *Mathematics and Computers in Archaeology*, Edinburgh University Press, Table 9.1). These are stored in Genstat spreadsheet Brooch.gsh (Figure 7.1).



Spreadsheet [Brooch.gsh]

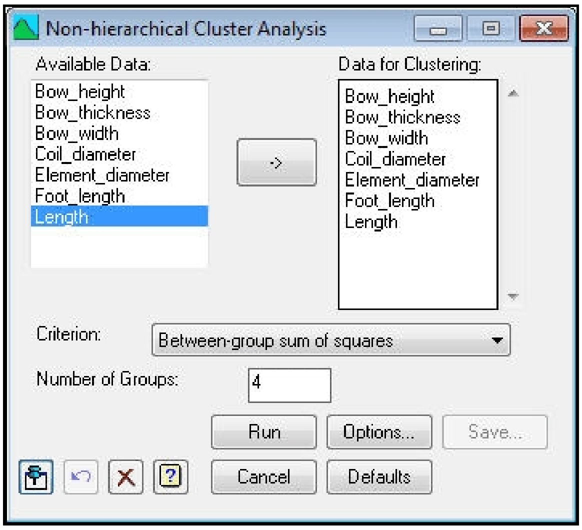
Row	Label	Foot_lengt	Bow_height	Coil_diameter	Element_diameter	Bow_width	Bow_thickness	Length
21	212	94	15	12	11	4.3	4.3	128
22	587	22	18	7	5	8.8	3	59
23	830	20	14	6	3	14.3	1.4	44
24	549	22	15	7	13	5	4.6	47
25	529	12	22	9	9	6.8	6.4	45
26	399	27	15	10	9	8.2	4	53
27	788	15	19	7	3	3.7	3.5	56
28	752	10	10	6	2	2	2.3	26
29	Hallstat	9	13	4	4	9.6	5	28
30	611	68	18	9	3	9.3	6.5	110

Figure 7.1

Before doing the cluster analysis, to counteract skewness in the variables, we transform each column of measurements *x* to log10(*x*+1). This can be done using the Calculate menu in the usual way. Alternatively, to save time, the transformed data are available in spreadsheet Logbrooch.gsh.

To obtain the Non-hierarchical Cluster Analysis menu you click on the Non-hierarchical line in the Cluster Analysis subsection in the Multivariate Analysis section of the Stats menu on the menu bar.

Figure 7.2 shows the menu set to use all the measurements to form four groups using the between-group sum of squares criterion.



Non-hierarchical Cluster Analysis

Available Data:
Bow_height
Bow_thickness
Bow_width
Coil_diameter
Element_diameter
Foot_length
Length

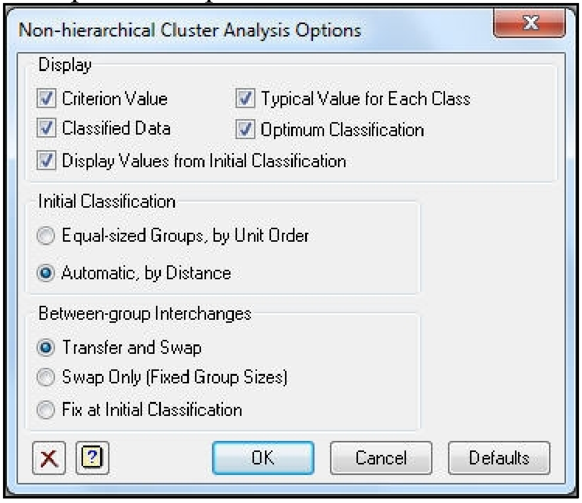
Data for Clustering:
Bow_height
Bow_thickness
Bow_width
Coil_diameter
Element_diameter
Foot_length
Length

Criterion: Between-group sum of squares

Number of Groups: 4

Figure 7.2

The Non-hierarchical Cluster Analysis Options menu controls the way in which the search for the best grouping is carried out, and the output that is produced.



Non-hierarchical Cluster Analysis Options

Display

☑ Criterion Value ☑ Typical Value for Each Class

☑ Classified Data ☑ Optimum Classification

☑ Display Values from Initial Classification

Initial Classification

○ Equal-sized Groups, by Unit Order

● Automatic, by Distance

Between-group Interchanges

● Transfer and Swap

○ Swap Only (Fixed Group Sizes)

○ Fix at Initial Classification

Figure 7.3

In Figure 7.3, we have asked Genstat to form the initial classification by finding the four objects that are furthest apart in the 7-dimensional space defined by the measurements, using these objects as the "cores" of the initial groups, and allocating the other objects to the group with the nearest core. (Note: this is feasible only if the number of groups does not exceed the number of variates.) The Between-group Interchanges box controls how Genstat generates new groupings from the initial classification. Here we are allowing objects both to be swapped between groups, and to be transferred from group to group. The setting Swap Only would constrain the group sizes to remain the same throughout the search (which might be useful, for example, if you wanted groups of equal sizes, and chosen the Equal-sized Groups option for the Initial Classification), and the setting Fix at Initial Configuration makes no changes.

Output from the clustering is of the brooches is shown below.

Non-hierarchical clustering

Sums of squares criterion

Initial classification

Number of classes = 4

Class contributions to criterion

```
                 1           2           3           4
            0.5471      0.2434      1.4189      0.3416
```

Criterion value = 2.55101

Classification of units

```
1    4    3    3    3    4    3    3    2    3    1
1    2    3    3    3    3    3    2    3    1    2
2    3    3    3    4    4    3    1
```

Class mean values

```
           Bow_height  Bow_thickness    Bow_width  Coil_diameter
     1          1.303          0.768        0.828          1.093
     2          1.217          0.514        1.102          0.899
     3          1.243          0.728        0.818          0.912
     4          1.112          0.564        0.619          0.874

     Element_diameter    Foot_length       Length
     1          0.992          1.819        1.994
     2          0.756          1.402        1.710
     3          0.959          1.298        1.674
     4          0.540          1.259        1.623
```

Units rearranged into class order

```
Group 1
                 1            2            3            4            5
             1.398        0.653        0.653        1.230        1.146
             1.380        0.833        0.919        1.079        1.176
             1.255        0.756        0.833        1.041        0.954
             1.204        0.724        0.724        1.114        1.079
             1.279        0.875        1.013        1.000        0.602

                 6            7
             1.973        2.061
             1.623        1.857
             1.681        1.898
             1.978        2.111
             1.839        2.045

Group 2
                 1            2            3            4            5
             1.146        0.380        1.270        0.903        0.778
             1.204        0.690        1.104        0.845        0.845
             1.279        0.519        0.959        1.000        0.778
             1.279        0.602        0.991        0.903        0.778
             1.176        0.380        1.185        0.845        0.602

                 6            7
             1.505        1.740
             1.477        1.681
             1.342        1.699
             1.362        1.778
             1.322        1.653

Group 3
                 1            2            3            4            5
             1.204        0.623        0.690        0.903        0.954
             1.431        0.940        0.857        1.000        1.114
             1.380        0.792        0.940        0.903        0.954
             1.230        0.708        0.851        1.000        0.903
             1.279        0.653        0.653        1.041        0.778
             1.255        0.881        1.009        0.845        0.845
             1.146        0.568        0.792        0.845        1.041
             1.204        0.763        0.756        0.778        1.114
             1.230        0.653        0.785        0.903        0.954
```

```
          1.204          0.681          0.813          0.903          0.845
          1.146          0.732          0.732          0.778          1.041
          1.255          0.556          0.544          1.041          0.954
          1.204          0.748          0.778          0.903          1.146
          1.362          0.869          0.892          1.000          1.000
          1.204          0.699          0.964          1.041          1.000
          1.146          0.778          1.025          0.699          0.699

              6              7
          1.531          1.785
          1.380          1.875
          1.322          1.839
          1.041          1.663
          1.204          1.613
          1.301          1.602
          1.380          1.623
          1.322          1.591
          1.255          1.653
          1.322          1.708
          1.322          1.568
          1.462          1.732
          1.362          1.681
          1.114          1.663
          1.447          1.732
          1.000          1.462

Group 4
              1              2              3              4              5
         0.9031         0.4314         0.6532         0.8451         0.4771
         1.2041         0.6532         0.6721         0.9031         0.6021
         1.3010         0.6532         0.6721         0.9031         0.6021
         1.0414         0.5185         0.4771         0.8451         0.4771

              6              7
         1.3424         1.5563
         1.4472         1.7482
         1.2041         1.7559
         1.0414         1.4314
```

Optimum classification

```
Number of classes = 4
```

Class contributions to criterion

```
              1              2              3              4
         0.5471         0.2434         0.9901         0.7254

Criterion value = 2.50611
```

Classification of units

```
1       4       3       3       3       4       3       4       2       3       1
1       2       3       3       3       3       3       2       3       1       2
2       3       3       3       4       4       4       1
```

Class mean values

```
                Bow_height   Bow_thickness       Bow_width   Coil_diameter
          1          1.303           0.768           0.828           1.093
          2          1.217           0.514           1.102           0.899
          3          1.247           0.730           0.815           0.917
          4          1.146           0.615           0.692           0.873

           Element_diameter    Foot_length          Length
          1          0.992           1.819           1.994
          2          0.756           1.402           1.710
          3          0.991           1.326           1.694
          4          0.606           1.207           1.594
```

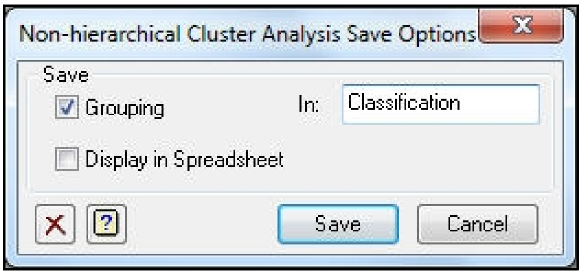
Units rearranged into class order

```
Group 1
               1               2               3               4               5
           1.398           0.653           0.653           1.230           1.146
           1.380           0.833           0.919           1.079           1.176
           1.255           0.756           0.833           1.041           0.954
           1.204           0.724           0.724           1.114           1.079
           1.279           0.875           1.013           1.000           0.602

               6               7
           1.973           2.061
           1.623           1.857
           1.681           1.898
           1.978           2.111
           1.839           2.045

Group 2
               1               2               3               4               5
           1.146           0.380           1.270           0.903           0.778
           1.204           0.690           1.104           0.845           0.845
           1.279           0.519           0.959           1.000           0.778
           1.279           0.602           0.991           0.903           0.778
           1.176           0.380           1.185           0.845           0.602

               6               7
           1.505           1.740
           1.477           1.681
           1.342           1.699
           1.362           1.778
           1.322           1.653
```

```
Group 3
              1             2             3             4             5
          1.204         0.623         0.690         0.903         0.954
          1.431         0.940         0.857         1.000         1.114
          1.380         0.792         0.940         0.903         0.954
          1.230         0.708         0.851         1.000         0.903
          1.255         0.881         1.009         0.845         0.845
          1.146         0.568         0.792         0.845         1.041
          1.204         0.763         0.756         0.778         1.114
          1.230         0.653         0.785         0.903         0.954
          1.204         0.681         0.813         0.903         0.845
          1.146         0.732         0.732         0.778         1.041
          1.255         0.556         0.544         1.041         0.954
          1.204         0.748         0.778         0.903         1.146
          1.362         0.869         0.892         1.000         1.000
          1.204         0.699         0.964         1.041         1.000

              6             7
          1.531         1.785
          1.380         1.875
          1.322         1.839
          1.041         1.663
          1.301         1.602
          1.380         1.623
          1.322         1.591
          1.255         1.653
          1.322         1.708
          1.322         1.568
          1.462         1.732
          1.362         1.681
          1.114         1.663
          1.447         1.732

Group 4
              1             2             3             4             5
         0.9031        0.4314        0.6532        0.8451        0.4771
         1.2041        0.6532        0.6721        0.9031        0.6021
         1.2788        0.6532        0.6532        1.0414        0.7782
         1.3010        0.6532        0.6721        0.9031        0.6021
         1.0414        0.5185        0.4771        0.8451        0.4771
         1.1461        0.7782        1.0253        0.6990        0.6990

              6             7
         1.3424        1.5563
         1.4472        1.7482
         1.2041        1.6128
         1.2041        1.7559
         1.0414        1.4314
         1.0000        1.4624
```

The output gives details of the initial classification and of the final (optimal) classification, showing the criterion value, how the objects are allocated to the groups and the mean values of the measurements in each group. In this example, the initial classification has been very successful. The optimum classification differs only in that the eighth object has been transferred from group 3 to group 4, and the 29th object from group 4 to group 3.

If you want, you can now save the final classification by using the Non-hierarchical Cluster Analysis Save Options menu (Figure 7.4), which is obtained by clicking on the Save button in the Non-hierarchical Cluster Analysis menu (Figure 7.2). Check the Grouping box, type the name of a factor to store the information into the In box, and then click on Save.



Non-hierarchical Cluster Analysis Save Options

Save

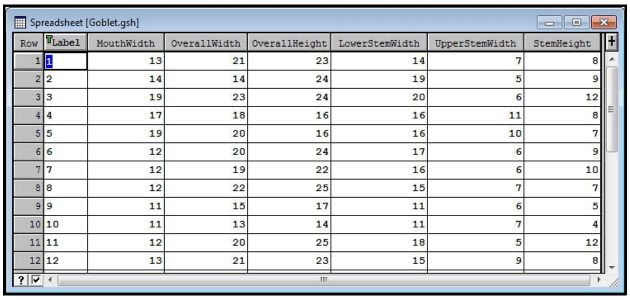
☑ Grouping In: Classification

□ Display in Spreadsheet

Figure 7.4

7.1 Practical

Genstat spreadsheet file Goblet.gsh (Figure 7.5) contains data on 25 goblets from prehistoric sites in Thailand (see page 147 of Manly, 1986, *Multivariate Statistical Methods a Primer*, Chapman & Hall, London). Perform a non-hierarchical classification into five groups. How does this compare with the dendrogram produced in Practical 6.1?



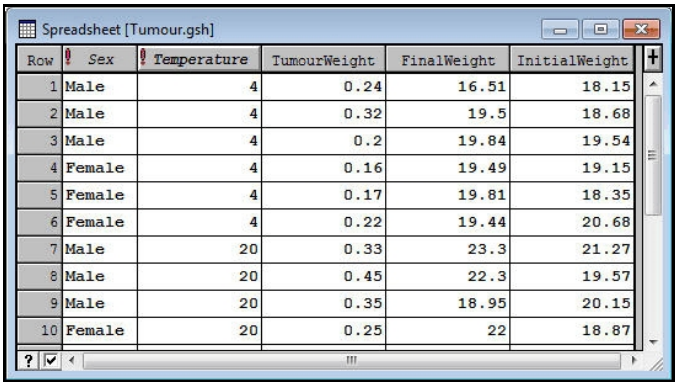
Spreadsheet [Goblet.gsh]

Row	Label	MouthWidth	OverallWidth	OverallHeight	LowerStemWidth	UpperStemWidth	StemHeight
1	1	13	21	23	14	7	8
2	2	14	14	24	19	5	9
3	3	19	23	24	20	6	12
4	4	17	18	16	16	11	8
5	5	19	20	16	16	10	7
6	6	12	20	24	17	6	9
7	7	12	19	22	16	6	10
8	8	12	22	25	15	7	7
9	9	11	15	17	11	6	5
10	10	11	13	14	11	7	4
11	11	12	20	25	18	5	12
12	12	13	21	23	15	9	8

Figure 7.5

8 Multivariate analysis of variance

Multivariate analysis of variance can be viewed as the extension of ordinary analysis of variance (as in Chapter 6) to handle several response variates at once. So, for example, instead of making assumptions of Normality for the residuals from a single response variate, we are now assuming multivariate Normality of residuals from several response variates.

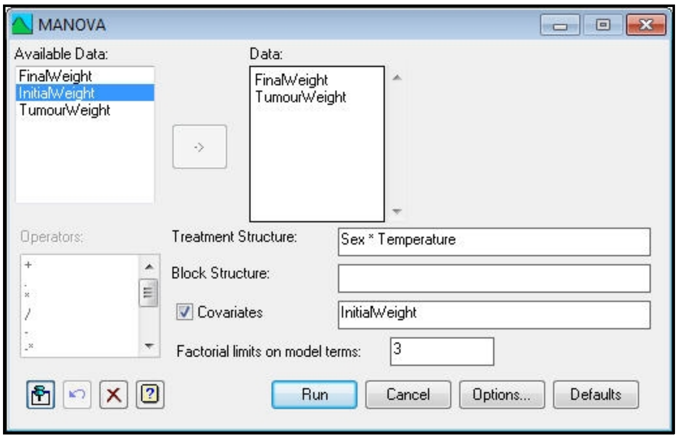

Spreadsheet [Tumour.gsh]

Row	Sex	Temperature	TumourWeight	FinalWeight	InitialWeight
1	Male	4	0.24	16.51	18.15
2	Male	4	0.32	19.5	18.68
3	Male	4	0.2	19.84	19.54
4	Female	4	0.16	19.49	19.15
5	Female	4	0.17	19.81	18.35
6	Female	4	0.22	19.44	20.68
7	Male	20	0.33	23.3	21.27
8	Male	20	0.45	22.3	19.57
9	Male	20	0.35	18.95	20.15
10	Female	20	0.25	22	18.87

Figure 8.1

We illustrate the analysis using data from an experiment to investigate sex and temperature effects on the growth of tumours in rats (see page 143 of Chatfield & Collins, 1986, *Introduction to Multivariate Analysis*, Chapman and Hall, London). Three rats of each sex were reared in each of three temperatures (4, 20 and 34). There was no blocking (i.e. this is a completely randomized design). The weights of the rats were measured (prior to sub-cutaneous seeding of the tumours). The response variates, taken at the end of the experiment, are the tumour weight and the final weight (excluding the tumour). The data are available in spreadsheet `Tumour.gsh` (Figure 8.28).

The MANOVA menu (Figure 8.2) is obtained by clicking on the MANOVA line in the Multivariate Analysis section of the Stats menu on the menu bar.


MANOVA

Available Data:
FinalWeight
InitialWeight
TumourWeight

Data:
FinalWeight
TumourWeight

Treatment Structure: Sex * Temperature

Block Structure:

☑ Covariates: InitialWeight

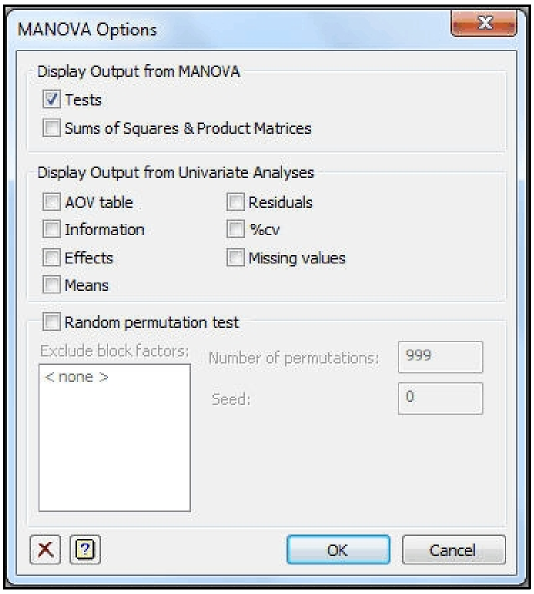
Factorial limits on model terms: 3

Figure 8.2

In Figure 8.2, we have specified a treatment structure of `Sex*Temperature`, to fit the main effects of sex and temperature, and their interaction (see Section 6.6). There is no block structure but we want to treat the variate `InitialWeight` as a *covariate* (so we check the Covariates box, and enter its name into the adjacent field). Covariates are included in the treatment model like variates in a linear regression. So, Genstat estimates a regression coefficient for them, and adjusts the other estimates and sums of squares to take account of their presence in the model (see *Guide to the Genstat Command Language*, Part 2 Section 4.3).

The MANOVA Options menu, shown in Figure 8.30, allows you to control the output from the multivariate analysis, and also to display output from the univariate anova’s of the individual response variates. Here we have asked just to print the various tests from the multivariate analysis, and omitted the sums and squares and products of the treatment effects and residuals (which are involved in calculating the tests).

The output, shown below, indicates that there are sex and temperature effects, but no interaction and no effect of the covariate.



MANOVA Options

Display Output from MANOVA

☑ Tests

□ Sums of Squares & Product Matrices

Display Output from Univariate Analyses

□ AOV table □ Residuals

□ Information □ %cv

□ Effects □ Missing values

□ Means

□ Random permutation test

Exclude block factors: < none >

Number of permutations: 999

Seed: 0

Figure 8.3

Multivariate analysis of covariance

Y-variates: FinalWt, TumourWt.
Covariate: InitWt.

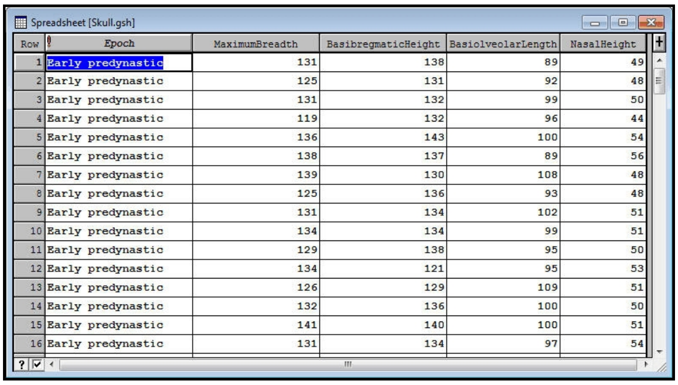
Test statistics

```
           Term    d.f.    Wilks' lambda   Rao F     n.d.f.    d.d.f.   F prob.
            Sex       1           0.3485    9.35         2        10     0.005
    Temperature       2           0.3269    3.75         4        20     0.020
Sex.Temperature       2           0.7830    0.65         4        20     0.633
      Covariate       1           0.8219    1.08         2        10     0.375

           Term    d.f.     Pillai-Bartlett    Roy's maximum    Lawley-Hotelling
                                      trace        root test               trace
            Sex       1              0.6515           0.6515              1.8697
    Temperature       2              0.8477           0.4949              1.5249
Sex.Temperature       2              0.2278           0.1605              0.2634
      Covariate       1              0.1781           0.1781              0.2167
```

The analysis uses the MANOVA procedure (see *Guide to the Genstat Command Language*, Part 2 Section 6.6.1). This uses the ANOVA directive, which requires the design to be balanced (see Section 6.7 or *Guide to the Genstat Command Language*, Part 2 Section 4.7). For unbalanced data, you can use the RMULTIVARIATE procedure, but this is not currently accessible through the menus.

8.1 Practical

Genstat spreadsheet file `Skull.gsh` (Figure 8.4) contains data on 150 male Egyptian skulls from five different epochs (see Practical 3.1 and pages 4 and 5 of Manly, 1986, *Multivariate Statistical Methods a Primer*, Chapman & Hall, London). Perform a multivariate analysis of variance. Are there any epoch differences?



Spreadsheet [Skull.gsh]

Row	Epoch	MaximumBreadth	BasibregmaticHeight	BasiolveolarLength	NasalHeight
1	Early predynastic	131	138	89	49
2	Early predynastic	125	131	92	48
3	Early predynastic	131	132	99	50
4	Early predynastic	119	132	96	44
5	Early predynastic	136	143	100	54
6	Early predynastic	138	137	89	56
7	Early predynastic	139	130	108	48
8	Early predynastic	125	136	93	48
9	Early predynastic	131	134	102	51
10	Early predynastic	134	134	99	51
11	Early predynastic	129	138	95	50
12	Early predynastic	134	121	95	53
13	Early predynastic	126	129	109	51
14	Early predynastic	132	136	100	50
15	Early predynastic	141	140	100	51
16	Early predynastic	131	134	97	54

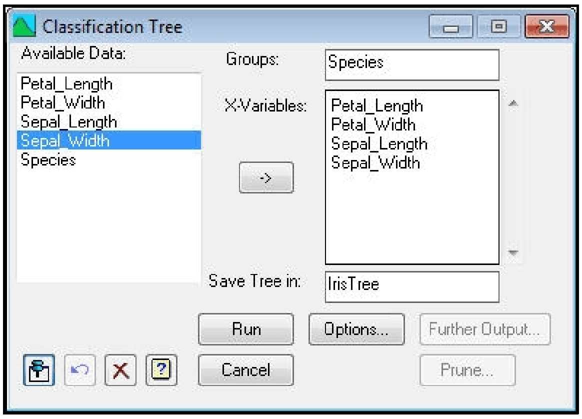
Figure 8.4

9 Classification trees

A classification tree is a device for predicting (or identifying) the class to which an unidentified object belongs. The starting point is a sample of objects from the various classes. Measurement recorded on the sample may be either continuous (supplied in variates) or discrete (supplied in factors). Below we shall illustrate the methods using the iris data from Chapter 3, where the data were all continuous (see Figure 3.1).

The Classification Tree menu (Figure 9.1) is in the Trees sub-option of the Multivariate Analysis option of the Stats menu on the menu bar.

In Figure 9.1, we have specified Species as the name of the factor defining the groups to be predicted, and entered the names of all the measurements into the X-variates box. The Save Tree in box allows you to specify a name for the *tree* structure that Genstat will generate to represent the classification tree. If you do not do this, Genstat will use its own private name, but you will not find it easy to use the tree outside the menus. Here we have specified the name IrisTree.



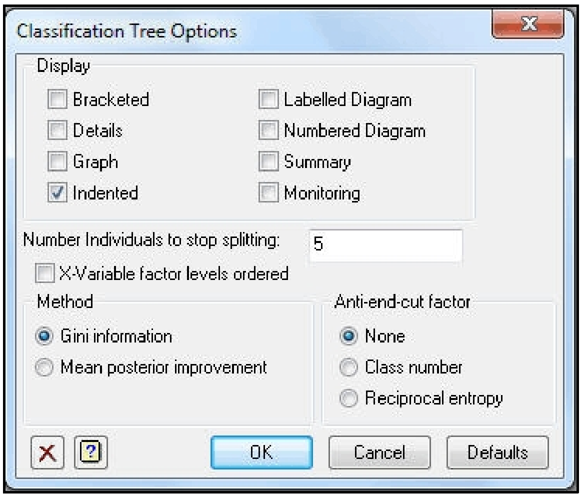
Groups: Species

X-Variables: Petal_Length, Petal_Width, Sepal_Length, Sepal_Width

Save Tree in: IrisTree

Figure 9.1

The tree progressively splits the objects into subsets based on their values for the measurements. Construction starts at a node known as the *root*, which contains all of the objects. A factor or variate is chosen to use there that "best" splits the individuals into two subsets. For example, in the tree for the irises, the first division is done by seeing whether the petal lengths are less than or greater then 2.450 (see the output below). The tree is then extended to contain two new nodes, one for each of the subsets, and factors or variates are selected for use at each of these nodes to subdivide the subsets further. The process stops when either no factor or variate provides any additional information, or the subset contains individuals all from the same group, or the subset contains fewer individuals than a limit specified by the Number of items to stop splitting field of the Classification Tree Options menu (Figure 9.2). The nodes where the construction ends are known as *terminal nodes*.



Display

□ Bracketed □ Labelled Diagram
□ Details □ Numbered Diagram
□ Graph □ Summary
☑ Indented □ Monitoring

Number Individuals to stop splitting: 5

□ X-Variable factor levels ordered

Method

● Gini information
○ Mean posterior improvement

Anti-end-cut factor

● None
○ Class number
○ Reciprocal entropy

Figure 9.2

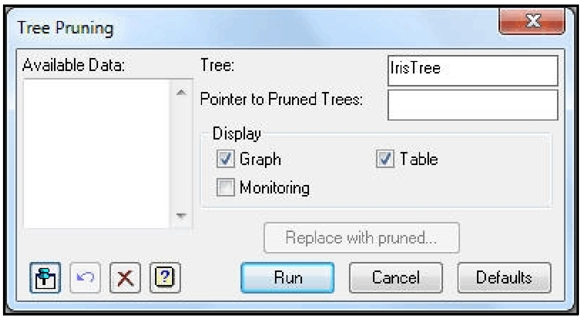
Factors may have either ordered or unordered levels, according to whether or not the X-Variable factor levels ordered box is checked. For example, a factor called Temperature with levels 5, 10 and 20 would usually be treated as having ordered levels, whereas levels labelled 'London', 'Moscow', 'New York', 'Ottawa' and 'Paris' of a factor called Town would be regarded as unordered. For unordered factors, all possible ways of dividing the levels into two sets are tried. With variates or ordered factors with more than 2 levels, a suitable value *p* is found to partition the individuals into those with values less than or greater than *p*. The radio buttons in the Method and Anti-end-cut-factor boxes in the Classification Tree Options menu allow you to choose how to assess the potential splits: whether to use Gini information or mean posterior improvement, and whether to use adaptive anti-end-cut factors. Details are given in the *Guide to the Genstat Command Language*, Part 2 Section 6.21.1.

The Display box of the Classification Tree Options menu (Figure 9.2) is set to print the tree only in "indented" format. This is a representation analogous to those used to display botanical trees. In the iris output, printed below, the first variable to examine is Petal_Length. If this is less than 2.450, the iris specimen is identified as *Setosa*. Otherwise you progress to index 2, and examine Petal_Width. So, a specimen of *Versicolor* might be identified by the sequence: (1) Petal_Length > 2.450; (2) Petal_Width < 1.750; (3) Petal_Length > 4.950; (5) Petal_Width > 1.550 *Versicolor*. Notice that the same variable can be used several times as the observed characteristics are refined on the way to an identification.

```
1 Petal_Length<2.450 Setosa
1 Petal_Length>2.450 2
 2 Petal_Width<1.750 3
 3 Petal_Length<4.950 4
 4 Petal_Width<1.650 Versicolor
 4 Petal_Width>1.650 Virginica
 3 Petal_Length>4.950 5
 5 Petal_Width<1.550 Virginica
 5 Petal_Width>1.550 Versicolor
 2 Petal_Width>1.750 6
 6 Petal_Length<4.850 Virginica
 6 Petal_Length>4.850 Virginica
```

Generally the construction of a classification tree will result in *over-fitting*. That is, it will form a tree that keeps selecting factors or variates to subdivide the individuals beyond the point that can be justified statistically. The solution is to prune the tree to remove the uninformative sub-branches. The pruning uses *accuracy* figures, which are stored for each node of the tree. The tree also stores a *prediction* for each node, which corresponds to the group with most individuals at the node. For each node of a classification tree, the accuracy is the number of misclassified individuals at the node, divided by the total number of individuals in the data set. It thus measures the "impurity" of the subset at that node (how far it is from it from being homogeneous i.e. having individuals all from a single group).

You can prune the tree using the Tree Pruning menu (Figure 9.3), which is


Tree Pruning

Available Data:

Tree: IrisTree

Pointer to Pruned Trees:

Display

☑ Graph ☑ Table

☐ Monitoring

Figure 9.3

accessible from Tree subsection in the Multivariate Analysis section of the menu bar or by clicking on the Prune button on the Classification Tree menu. As we have loaded the menu from the Classification Tree menu, Genstat has filled in the name of the tree automatically.

In the Display box, we have asked for the relationship between the impurity and the number of terminal nodes to be presented in a graph (Figure 9.4) and a table (below).

The table and graph show that the impurity of the pruned trees drops rapidly as the number of terminal nodes increases from one up to three, but then tails off more slowly. This suggests that we should prune down to three terminal nodes, but no further. This tree is the fifth in the sequence of pruned trees (count from the right of the graph, or notice the numbering in first column of the table).

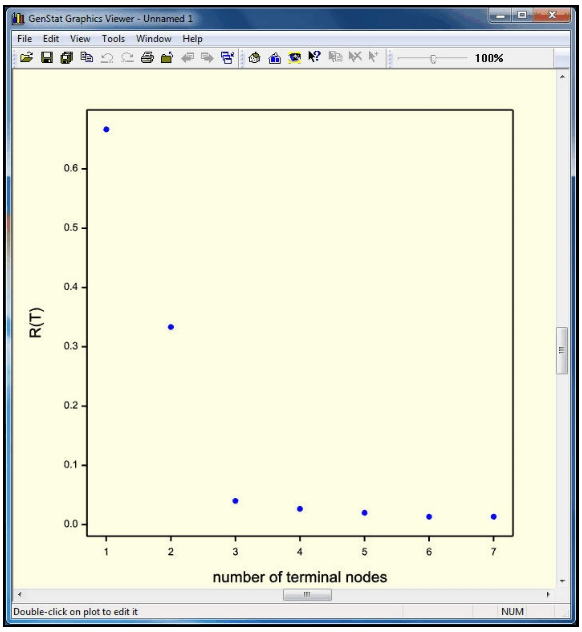
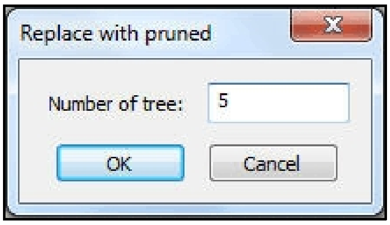

GenStat Graphics Viewer - Unnamed 1
File Edit View Tools Window Help
100%
R(T)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
6
7
number of terminal nodes
Double-click on plot to edit it
NUM

Figure 9.4

Characteristics of the pruned trees

Tree no.	RT	Number of terminal nodes
1	0.0133	7
2	0.0133	6
3	0.0200	5
4	0.0267	4
5	0.0400	3
6	0.3333	2
7	0.6667	1

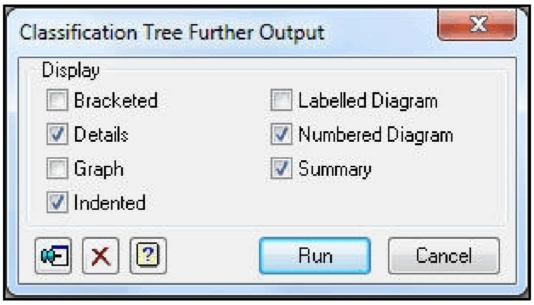
By clicking the button Replace with pruned we can replace contents of the tree IrisTree with this smaller tree. We simply need to fill in the number of the tree (5) in the resulting menu (Figure 9.5), click on OK, and then cancel the Tree Pruning menu.



Number of tree: 5

Figure 9.5

The pruned tree can be displayed using the Classification Tree Further Output menu (Figure 9.6), obtained by clicking on the Further Output button on the Classification Tree menu.


Classification Tree Further Output

Display

☐ Bracketed
☑ Details
☐ Graph
☑ Indented
☐ Labelled Diagram
☑ Numbered Diagram
☑ Summary

Figure 9.6

Summary of classification tree: IrisTree

```
Number of nodes: 5
Number of terminal nodes: 3
Misclassification rate: 0.040
Variables in the tree: Petal_Length, Petal_Width.
```

Details of classification tree: IrisTree

```
1 Current prediction: 1.000
  Number of observations: 150
               Species   SetosaVersicolor Virginica
         Proportions     0.333     0.333     0.333
  Test: Petal_Length<2.450
        Next nodes:23

 2 Current prediction: 1.000
   Number of observations: 50
                Species   SetosaVersicolor Virginica
          Proportions     1.000     0.000     0.000
   Conclusion: Setosa

 3 Current prediction: 2.000
   Number of observations: 100
                Species   SetosaVersicolor Virginica
          Proportions     0.000     0.500     0.500
   Test: Petal_Width<1.750
         Next nodes:45

  4 Current prediction: 2.000
    Number of observations: 54
                 Species   SetosaVersicolor Virginica
           Proportions     0.000     0.907     0.093
    Conclusion: Versicolor

  5 Current prediction: 3.000
    Number of observations: 46
                 Species   SetosaVersicolor Virginica
           Proportions     0.000     0.022     0.978
    Conclusion: Virginica
```

```
1 Petal_Length<2.450 Setosa
1 Petal_Length>2.450 2
2 Petal_Width<1.750 Versicolor
2 Petal_Width>1.750 Virginica


Tree diagram

1   2
->  3   4
    ->  5
```

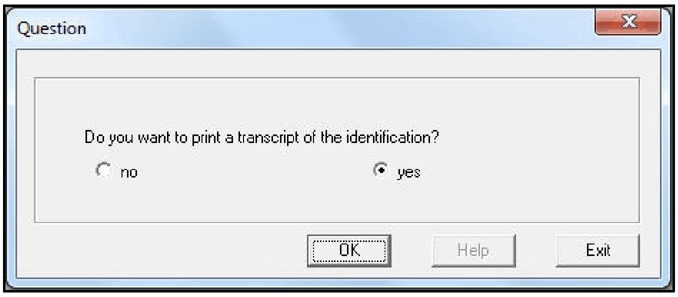
The initial summary, generated by the Summary check box, lists the number of nodes (5) and terminal nodes (3) in the tree, its misclassification rate and which variables it uses. The details section (from the Details check box) gives information about each node, referring to the numbering displayed in the tree diagram at the end of the output (which is generated by the Numbered Diagram check box).

Note, if possible, it is best to use “accuracy” figures that are derived from a different set or sets of data from that which was used to construct the tree. This cannot be done through the menus, but you can use the BCVALUES procedure, which is described in the *Guide to the Genstat Command Language*, Part 2 Section 6.21.3.

Another useful procedure, which also cannot be accessed currently through the menus is BCIDENTIFY. This has a convenient interactive interface, that asks you to enter the information required by the tree as and when it is needed. (For details of the options and parameters that allow you to use it in batch mode, see the *Guide to the Genstat Command Language*, Part 2 Section 6.21.3). To run the procedure in this way, you merely need to set the TREE option to the name of the tree, here IrisTree. If we type the command

```
BCIDENTIFY [TREE=IrisTree]
```

and execute it, for example by clicking on the Submit Line line in the Run menu on the menu bar, Genstat asks the question in Figure 9.7. (Our answer is yes.)

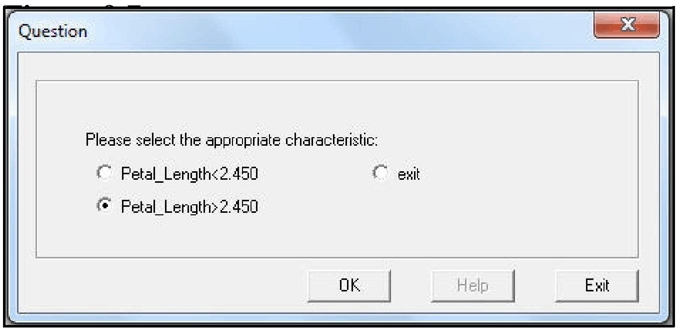


Question

Do you want to print a transcript of the identification?

○ no ● yes

The next question is in Figure 9.8, to which we shall answer that the petal length is greater than 2.450. (Check the box and click on OK.)



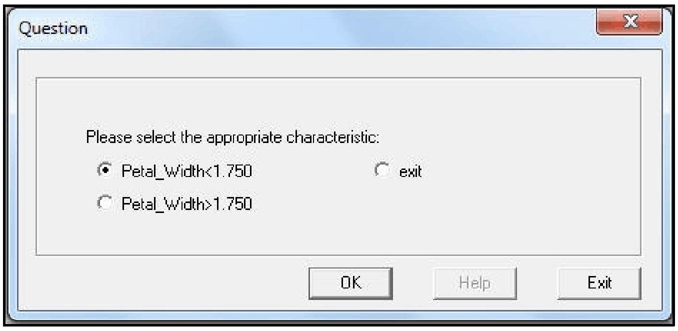
Question

Please select the appropriate characteristic:

○ Petal_Length<2.450 ○ exit
● Petal_Length>2.450

Figure 9.8

This generates the question in Figure 9.9, to which we shall answer that the petal width is less than 1.750.



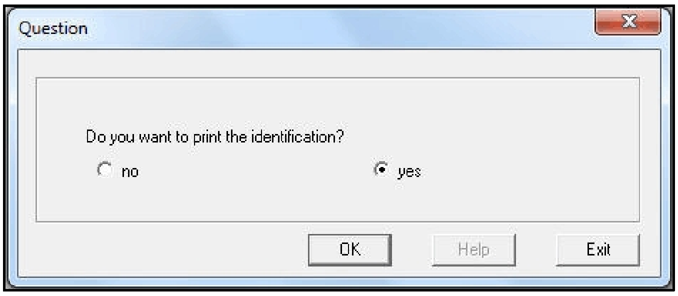
Please select the appropriate characteristic:

● Petal_Width<1.750 ○ exit

○ Petal_Width>1.750

Figure 9.9

We have now reached the terminal node, and Genstat asks if we want to print the identification (Figure 9.10). It would be best to take the default suggestion, of yes, here as we have not set the option of `BCIDENTIFY` that would save the information!



Do you want to print the identification?

○ no ● yes

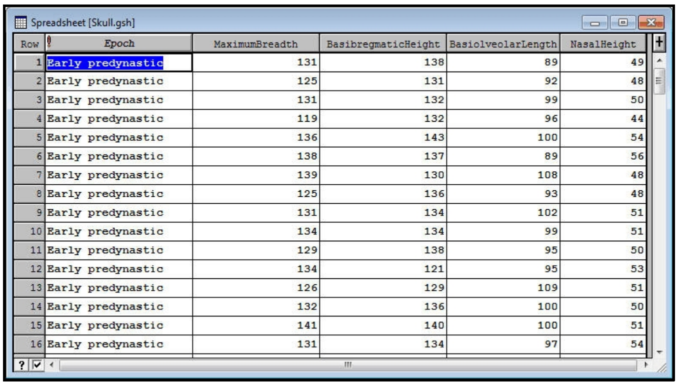
Figure 9.10

The output shows first a transcript of the questions and answers (as requested in Figure 9.7), and then the identification of *Versicolor*.

```
Identification using a classification tree

Observations:
Petal_Length>2.450
Petal_Width<1.750

Identification:
Versicolor
```

9.1 Practical

Genstat spreadsheet file `Skull.gsh` (Figure 9.11) contains data on 150 male Egyptian skulls from five different epochs (see Practical 3.1 and pages 4 and 5 of Manly, 1986, *Multivariate Statistical Methods a Primer*, Chapman & Hall, London).



Row	Epoch	MaximumBreadth	BasibregmaticHeight	BasiolveolarLength	NasalHeight
1	Early predynastic	131	138	89	49
2	Early predynastic	125	131	92	48
3	Early predynastic	131	132	99	50
4	Early predynastic	119	132	96	44
5	Early predynastic	136	143	100	54
6	Early predynastic	138	137	89	56
7	Early predynastic	139	130	108	48
8	Early predynastic	125	136	93	48
9	Early predynastic	131	134	102	51
10	Early predynastic	134	134	99	51
11	Early predynastic	129	138	95	50
12	Early predynastic	134	121	95	53
13	Early predynastic	126	129	109	51
14	Early predynastic	132	136	100	50
15	Early predynastic	141	140	100	51
16	Early predynastic	131	134	97	54

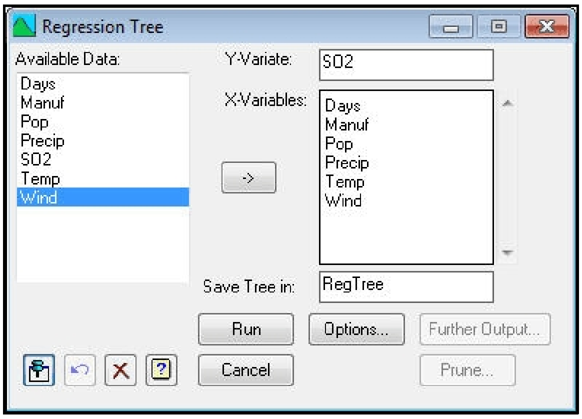
Figure 9.11

Form a classification tree. Prune down to 20 terminal nodes. What is the misclassification probability?

10 Regression trees

Regression trees are very similar to classification trees, except that the attribute to predict is the value of a response variate rather than the level of a group factor. So the starting point is now a sample of observations with various values of the response. As in a classification tree, the measurements recorded on the sample may be either continuous (supplied in variates) or discrete (supplied in factors). Below we shall illustrate the methods using the pollution data from Chapters 1 and 2, where the data were all continuous (see Figure 1.1).

The Regression Tree menu (Figure 10.1) is a sub-option of the Regression Analysis option of the Stats menu on the menu bar. In Figure 10.1, we have specified SO2 as the response variate, and entered the names of all the measurements into the X-variates box. The Save Tree in box allows you to specify a name for the *tree* structure that Genstat will generate to represent the classification tree. If you do not do this, Genstat will use its own private name, but you will not find it easy to use the tree outside the menus. Here we have specified the name RegTree.



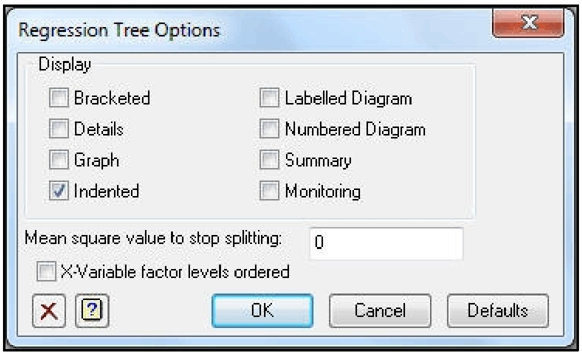
Y-Variate: SO2

X-Variables: Days Manuf Pop Precip Temp Wind

Save Tree in: RegTree

Figure 10.1

The tree progressively splits the observations into subsets based on their values for the measurements. Construction starts at a node known as the *root*, which contains all of the observations. A factor or variate is chosen to use there that “best” splits the observations into two subsets. The aim is to form subsets that have similar values for the response variate. The predicted value of the response variable at each node of the tree is the mean of its value for the subset of observations at that node. The *accuracy* of the node is the squared distance of the values of the response variate from their mean for the observations at the node, divided by the total number of observations. The potential splits at the node are assessed by their effect on the accuracy, that is the difference between the accuracy of the node and the sum of the accuracies of the two potential successor nodes. The node will become a terminal node if none of the splits provides any improvement in accuracy, or if the mean square of the observations at the node is less than a limit that can be specified in the Regression Tree Options menu (Figure 10.2). As in a classification tree, factors may have either ordered or unordered levels, according to whether or not the X-Variable factor levels ordered box is checked (see Chapter 9 for more details).



☐ Bracketed

☐ Details

☐ Graph

☑ Indented

☐ Labelled Diagram

☐ Numbered Diagram

☐ Summary

☐ Monitoring

Mean square value to stop splitting: 0

☐ X-Variable factor levels ordered

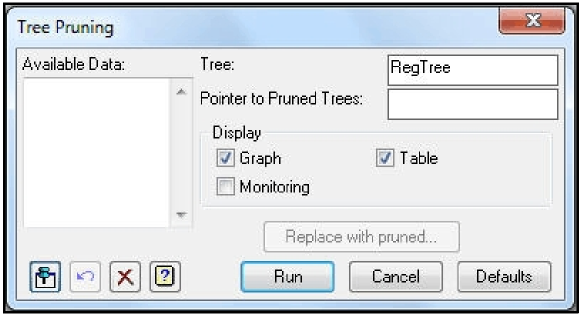
Figure 10.2

The menu also allows you to select the output to display. Here we have asked for the tree in “indented” format (as we did earlier for the classification tree in Chapter 9). So, in the output below, the first variable to examine is Manuf. If this is less than 748.0, you progress to index 2, and examine Pop. Otherwise, you find the other line for index 1, further down the tree, which tells you to go to index 33 and form another split involving Manuf. The terminal nodes (at which predictions of SO2 are made) are identified by the fact that they are followed by real numbers (with decimal points) rather than integers. So, for example, at index 5, SO2 is predicted to be 13.

```
1 Manuf<748.0 2
 2 Pop<190.0 3
 3 Wind<9.800 4
 4 Temp<-50.05 5
 5 Temp<-58.10 13.
 5 Temp>-58.10 6
 6 Wind<8.850 7
 7 Days<118.5 28.
 7 Days>118.5 31.
 6 Wind>8.850 36.
 4 Temp>-50.05 8
 8 Days<131.0 56.
 8 Days>131.0 46.
 3 Wind>9.800 94.
 2 Pop>190.0 9
 9 Days<108.0 10
 10 Temp<-55.55 11
 11 Wind<10.85 12
 12 Temp<-59.20 10.
 12 Temp>-59.20 13
 13 Days<62.50 11.
 13 Days>62.50 12.
 11 Wind>10.85 14
 14 Days<80.00 9.
 14 Days>80.00 8.
 10 Temp>-55.55 15
 15 Days<101.0 16
 16 Days<92.00 17.
 16 Days>92.00 14.
 15 Days>101.0 17.
 9 Days>108.0 17
 17 Temp<-59.35 18
 18 Manuf<241.0 19
 19 Manuf<170.0 14.
 19 Manuf>170.0 20
 20 Days<120.5 9.
 20 Days>120.5 10.
 18 Manuf>241.0 21
 21 Days<117.0 24.
 21 Days>117.0 18.
 17 Temp>-59.35 22
 22 Wind<11.20 23
 23 Days<142.0 24
 24 Pop<831.0 25
 25 Wind<7.950 26
 26 Days<123.5 26.
 26 Days>123.5 23.
 25 Wind>7.950 27
 27 Precip<41.00 28
 28 Manuf<313.5 26.
 28 Manuf>313.5 29
 29 Manuf<397.5 28.
 29 Manuf>397.5 29.
 27 Precip>41.00 30
 30 Days<119.5 31.
 30 Days>119.5 30.
 24 Pop>831.0 47.
 23 Days>142.0 31
 31 Days<155.5 61.
 31 Days>155.5 29.
 22 Wind>11.20 32
 32 Days<144.5 16.
 32 Days>144.5 11.
1 Manuf>748.0 33
33 Manuf<2518 34
 34 Precip<32.98 35.
 34 Precip>32.98 35
 35 Days<110.0 56.
 35 Days>110.0 36
 36 Days<135.0 69.
 36 Days>135.0 65.
33 Manuf>2518 110.
```

As with classification trees (Chapter 9), the construction of a regression tree will generally result in *over-fitting*. That is, it will form a tree that keeps selecting factors or variates to subdivide the individuals beyond the point that can be justified statistically. The solution is again to prune the tree to remove the uninformative sub-branches. The pruning uses *accuracy* of the nodes of the tree, as defined above.

You can prune the tree using the Tree Pruning menu (Figure 10.3), which can be opened from Tree subsection in the Multivariate Analysis section of the menu bar or by clicking on the Prune button on the Regression Tree menu. As we have loaded the menu from the Regression Tree menu, Genstat has filled in the name of the tree automatically.



Tree: RegTree

Pointer to Pruned Trees:

Display

☑ Graph ☑ Table

☐ Monitoring

Figure 10.3

In the Display box, we have asked for the relationship between the accuracy and the number of terminal nodes to be presented in a graph (Figure 10.4) and a table (below). The table and graph show that the impurity of the pruned trees drops rapidly as the number of terminal nodes increases from one up to about ten, but then tails off more slowly. This suggests that we should prune down to ten terminal nodes, but no further. This tree is the 26th in the sequence of pruned trees.

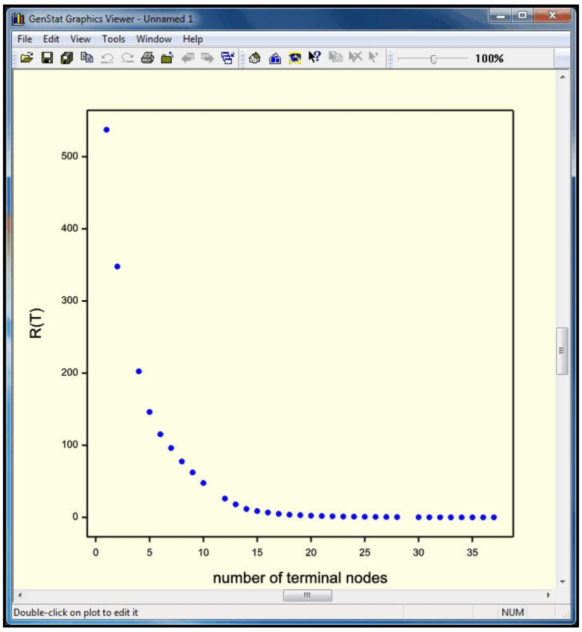

GenStat Graphics Viewer - Unnamed 1
File Edit View Tools Window Help
100%
R(T)
500
400
300
200
100
0
0
5
10
15
20
25
30
35
number of terminal nodes
Double-click on plot to edit it
NUM

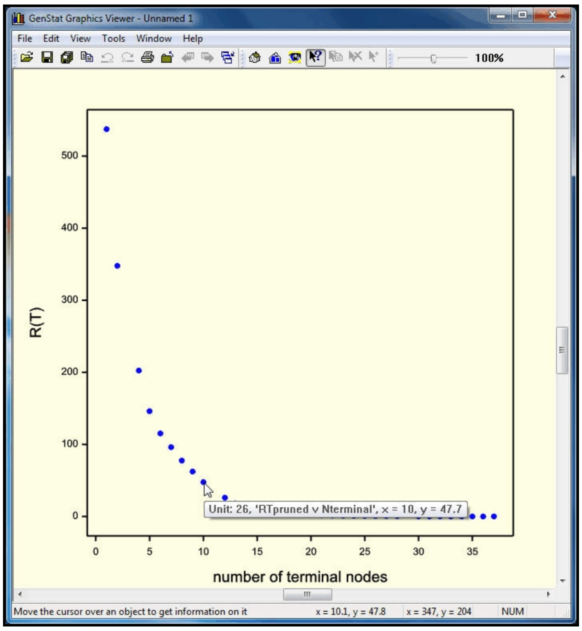
Figure 10.4

Characteristics of the pruned trees

Tree no.	RT	Number of terminal nodes
1	0.00	37
2	0.01	36
3	0.02	35
4	0.04	34
5	0.05	33
6	0.07	32
7	0.13	31
8	0.24	30
9	0.46	28
10	0.57	27
11	0.70	26
12	0.85	25
13	1.05	24
14	1.25	23
15	1.56	22
16	1.89	21
17	2.33	20
18	3.01	19
19	3.70	18
20	4.92	17
21	6.80	16
22	8.76	15
23	11.69	14
24	18.07	13
25	26.10	12
26	47.70	10
27	62.39	9
28	77.50	8

29	96.23	7
30	115.24	6
31	146.10	5
32	202.47	4
33	347.87	2
34	537.51	1

You can get the number of the tree from the index in the first column of the table, or by counting from the right of the graph, or by using the Data-info tool on the graph, as shown in Figure 10.5. To select the tool, you click on the icon with the arrow and question mark at the left-hand end of the Graphics Toolbar. The viewer will then display details of a point when you rest the pointer nearby.

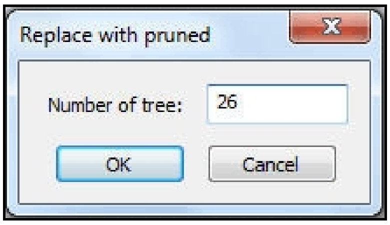

GenStat Graphics Viewer - Unnamed 1
File Edit View Tools Window Help
100%
R(T)
500
400
300
200
100
0
0
5
10
15
20
25
30
35
number of terminal nodes

Unit: 26, 'RTpruned v Nterminal', x = 10, y = 47.7

Move the cursor over an object to get information on it x = 10.1, y = 47.8 x = 347, y = 204 NUM

Figure 10.5

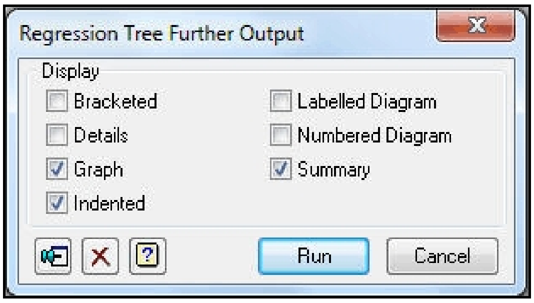
By clicking the button Replace with pruned we can replace contents of the tree RegTree with this smaller tree. We simply need to fill in the number of the tree (26) in the resulting menu (Figure 10.6), click on OK, and then cancel the Tree Pruning menu.



Number of tree: 26

Figure 10.6

The pruned tree can be displayed using the Regression Tree Further Output menu (Figure 10.7), obtained by clicking on the Further Output button on the Regression Tree menu. Here we have asked to print the tree in indented form again, to print a summary of its properties and to display it in a graph (see Figure 10.8).



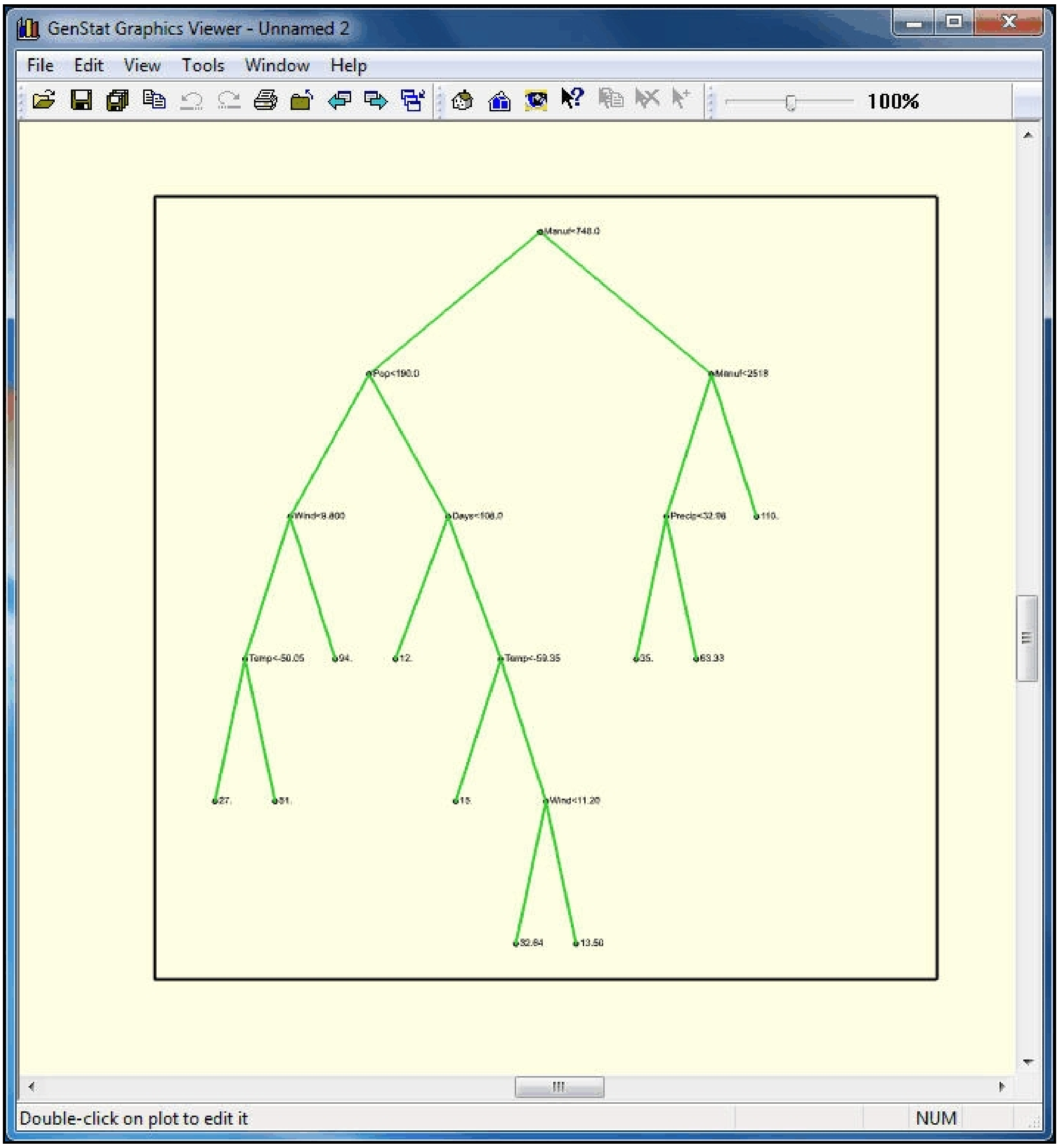
Display

□ Bracketed ☐ Labelled Diagram
□ Details □ Numbered Diagram
☑ Graph ☑ Summary
☑ Indented

Figure 10.7

```
Summary of regression tree: RegTree

Number of nodes: 19
Number of terminal nodes: 10
Residual sum of squares: 1956
Residual degrees of freedom:  31
Residual mean square: 63.09
Percentage variance accounted for: 88.55
Variables in the tree: Manuf, Pop, Wind, Days, Precip, Temp.


1 Manuf<748.0 2
 2 Pop<190.0 3
 3 Wind<9.800 4
 4 Temp<-50.05 27.
 4 Temp>-50.05 51.
 3 Wind>9.800 94.
 2 Pop>190.0 5
 5 Days<108.0 12.
 5 Days>108.0 6
 6 Temp<-59.35 15.
 6 Temp>-59.35 7
 7 Wind<11.20 32.64
 7 Wind>11.20 13.50
1 Manuf>748.0 8
 8 Manuf<2518 9
 9 Precip<32.98 35.
 9 Precip>32.98 63.33
 8 Manuf>2518 110.
```

The initial summary, generated by the Summary check box, lists the number of nodes (19) and terminal nodes (10) in the tree, its residual sum of squares, degrees of freedom and mean square, and the variables that it uses. Note, if possible, it is best to use “accuracy” figures that are derived from a different set or sets of data from that which was used to construct the tree. This cannot be done through the menus, but you can use the BCVALUES procedure, which is described in the *Guide to the Genstat Command Language*, Part 2 Section 6.21.3.



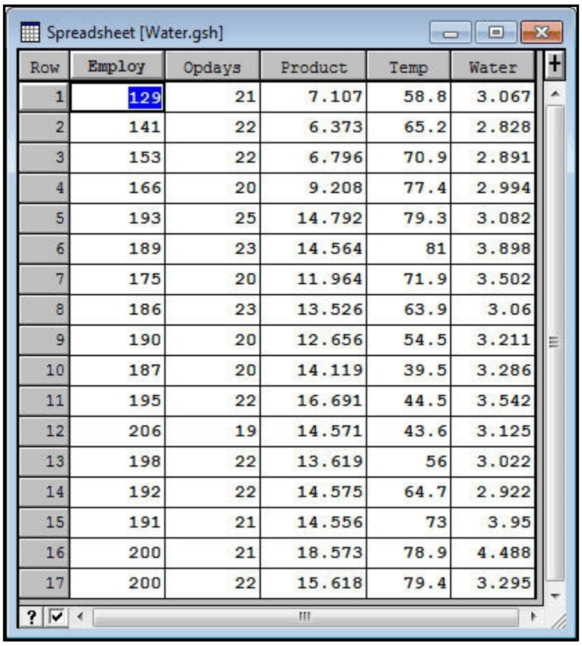
Double-click on plot to edit it

Figure 10.8

10.1 Practical

Genstat spreadsheet file Water.gsh (Figure 10.9) contains data about the water usage of a production plant (last column of the sheet). There are also four variates that may be associated with the amount of water that has been used: the average temperature, the amount of production, the number of operating days and the number of employees. (See page 352 of *Applied Regression Analysis* by Draper & Smith, 1981, Wiley, New York.)

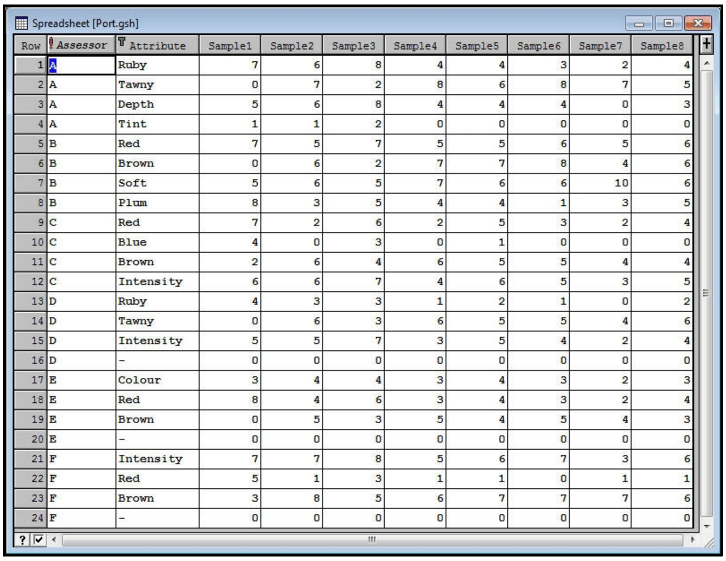
Form a regression tree to predict water usage from the other variates.



Row	Employ	Opdays	Product	Temp	Water
1	129	21	7.107	58.8	3.067
2	141	22	6.373	65.2	2.828
3	153	22	6.796	70.9	2.891
4	166	20	9.208	77.4	2.994
5	193	25	14.792	79.3	3.082
6	189	23	14.564	81	3.898
7	175	20	11.964	71.9	3.502
8	186	23	13.526	63.9	3.06
9	190	20	12.656	54.5	3.211
10	187	20	14.119	39.5	3.286
11	195	22	16.691	44.5	3.542
12	206	19	14.571	43.6	3.125
13	198	22	13.619	56	3.022
14	192	22	14.575	64.7	2.922
15	191	21	14.556	73	3.95
16	200	21	18.573	78.9	4.488
17	200	22	15.618	79.4	3.295

Figure 10.9

11 Generalized Procrustes analysis



Spreadsheet [Port.gsh]

Row	Assessor	Attribute	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
1	A	Ruby	7	6	8	4	4	3	2	4
2	A	Tawny	0	7	2	8	6	8	7	5
3	A	Depth	5	6	8	4	4	4	0	3
4	A	Tint	1	1	2	0	0	0	0	0
5	B	Red	7	5	7	5	5	6	5	6
6	B	Brown	0	6	2	7	7	8	4	6
7	B	Soft	5	6	5	7	6	6	10	6
8	B	Plum	8	3	5	4	4	1	3	5
9	C	Red	7	2	6	2	5	3	2	4
10	C	Blue	4	0	3	0	1	0	0	0
11	C	Brown	2	6	4	6	5	5	4	4
12	C	Intensity	6	6	7	4	6	5	3	5
13	D	Ruby	4	3	3	1	2	1	0	2
14	D	Tawny	0	6	3	6	5	5	4	6
15	D	Intensity	5	5	7	3	5	4	2	4
16	D	–	0	0	0	0	0	0	0	0
17	E	Colour	3	4	4	3	4	3	2	3
18	E	Red	8	4	6	3	4	3	2	4
19	E	Brown	0	5	3	5	4	5	4	3
20	E	–	0	0	0	0	0	0	0	0
21	F	Intensity	7	7	8	5	6	7	3	6
22	F	Red	5	1	3	1	1	0	1	1
23	F	Brown	3	8	5	6	7	7	7	6
24	F	–	0	0	0	0	0	0	0	0

Figure 11.1

Generalized Procrustes analysis allows you to produce a consensus configuration of points from several input configurations. It is often used in sensory analysis, for example of food or wine, where the input configurations will be assessments made of various attributes of the food or wine samples.

Figure 11.1 shows an example data set from an evaluation of the appearance of port-wines by Williams & Langron (1984, *Journal of the Science of Food and Agriculture*, **35**, 558-568), stored in spreadsheet file Port.gsh. There were six assessors (labelled A-F in column 1), eight samples of port (columns Sample1 - Sample8), and the attributes measured by each assessor are described in the Attribute column. Notice that this is an example of *free choice profiling*; the assessors were not required to observe the same attributes, but they could each define their own. The only constraint is that each assessor must be consistent in their definition of an attribute over the samples. Also, for the analysis to work, each assessor must observe the same number of attributes. However, it is valid to include a "null" attribute of zero observations for assessors that have observed too few (assessors D, E and F here).

The analysis treats the observations from each assessor as a configuration of n points (one for each sample) in p dimensions (one for each attribute), and forms a *centroid* configuration that gives a consensus view of how the assessors perceive the ports. The basic data for the analysis is a set of attribute × sample matrices, containing the measurements made by each assessor.

These can be formed by first deleting the Attribute column (put the cursor into any cell in the column, then select the Current Column sub-option of the Delete option of the Spread menu, as shown in Figure 11.2.

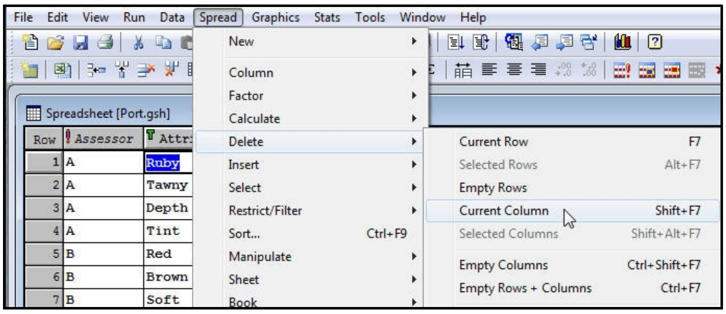
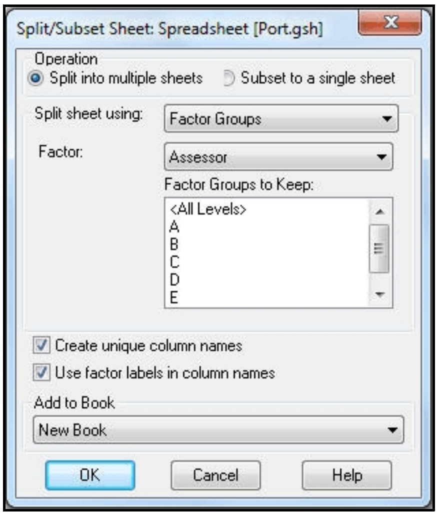


Figure 11.2

Then open the Split/Subset Sheet menu (Figure 11.3), by selecting the Split/Subset sub-option of the Manipulate option of the Spread menu. Figure 11.3 is set to split the spreadsheet into multiple sheets using all levels of the Assessor factor.



Split/Subset Sheet: Spreadsheet [Port.gsh]

Operation

● Split into multiple sheets ○ Subset to a single sheet

Split sheet using: Factor Groups

Factor: Assessor

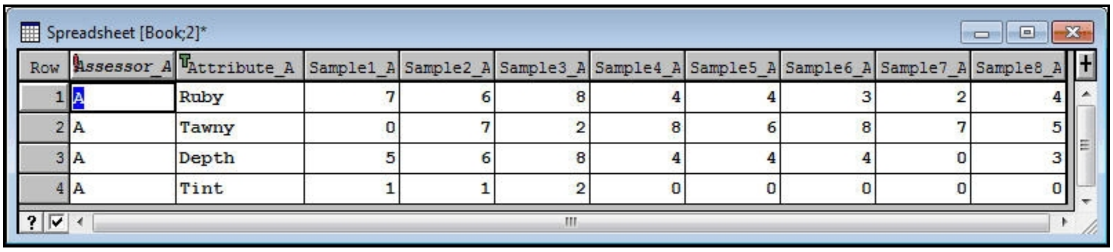
Factor Groups to Keep:
<All Levels>
A
B
C
D
E

☑ Create unique column names

☑ Use factor labels in column names

Add to Book: New Book

Figure 11.3

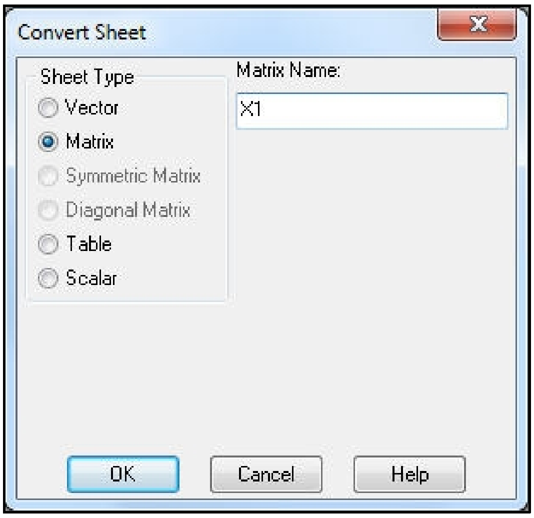


Spreadsheet [Book;2]*

Row	Assessor_A	Attribute_A	Sample1_A	Sample2_A	Sample3_A	Sample4_A	Sample5_A	Sample6_A	Sample7_A	Sample8_A
1	A	Ruby	7	6	8	4	4	3	2	4
2	A	Tawny	0	7	2	8	6	8	7	5
3	A	Depth	5	6	8	4	4	4	0	3
4	A	Tint	1	1	2	0	0	0	0	0

Figure 11.4

The first of the six new spreadsheets is shown in Figure 11.4. We need to delete the assessor column. We then convert the spreadsheet into a matrix by selecting the Convert sub-option of the Manipulate option of the Spread menu to open the Convert Sheet menu. In that menu (Figure 11.5), we select Matrix as the Sheet Type, give it a name (here M1), and click on OK. We now need to transpose the spreadsheet, by selecting the Transpose sub-option of the Manipulate option of the Spread menu, and we can rename the transposed matrix (e.g. to X1) by using the Sheet Properties menu (opened by selecting the Properties sub-option of the Sheet option of the Spread menu. We then need to repeat the process for the other new spreadsheets, giving each transposed matrix a different name (X2 - X6).



Convert Sheet

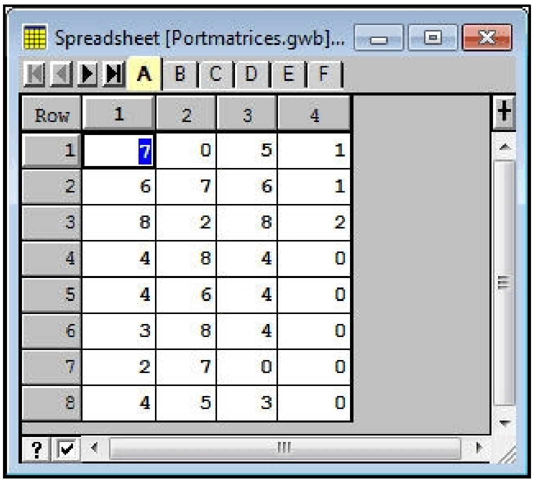
Sheet Type

○ Vector
● Matrix
○ Symmetric Matrix
○ Diagonal Matrix
○ Table
○ Scalar

Matrix Name: X1

Figure 11.5

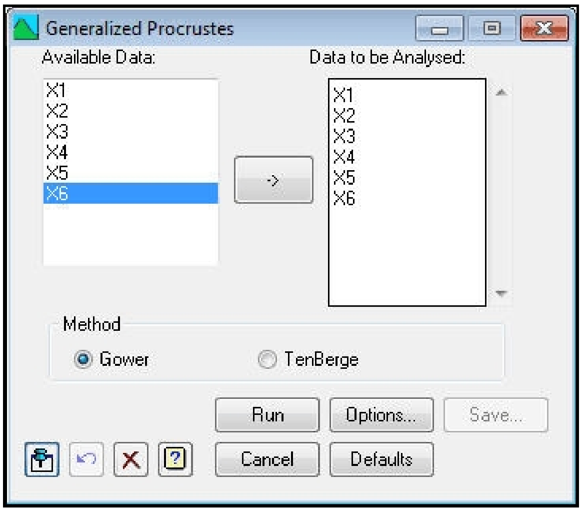
However, as spreadsheet manipulation is not the main point of this Chapter, the transposed matrices can be found in spreadsheet book Portmatrices.gwb, as shown in Figure 11.6. (The original data were presented in Figure 11.1 in order to display their structure more clearly.)



Spreadsheet [Portmatrices.gwb]...

Row	1	2	3	4
1	7	0	5	1
2	6	7	6	1
3	8	2	8	2
4	4	8	4	0
5	4	6	4	0
6	3	8	4	0
7	2	7	0	0
8	4	5	3	0

Figure 11.6



Generalized Procrustes

Available Data: X1 X2 X3 X4 X5 X6

Data to be Analysed: X1 X2 X3 X4 X5 X6

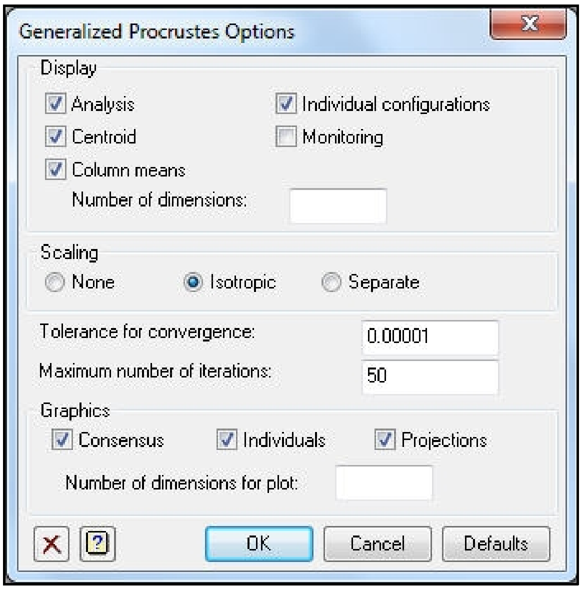
Method

● Gower ○ TenBerge

Figure 11.7

The Generalized Procrustes menu (Figure 11.7) is opened by selecting the Generalized Procrustes sub-option of the Multivariate Analysis option of the Stats menu. The main task is to set the Data to be Analysed to the matrices containing the configurations (here X1 - X6).

The common centroid configuration is formed by the operations of translation to a common origin, rotation and reflection of axes, and possibly also scale changes. It is found iteratively, using either Gower's or Tenberge's method, by minimizing the sum of the squared distances between the centroid and each individual configuration. To give a unique representation, the final centroid is defined using its principal axes.



Generalized Procrustes Options

Display

☑ Analysis ☑ Individual configurations

☑ Centroid □ Monitoring

☑ Column means

Number of dimensions:

Scaling

○ None ● Isotropic ○ Separate

Tolerance for convergence: 0.00001

Maximum number of iterations: 50

Graphics

☑ Consensus ☑ Individuals ☑ Projections

Number of dimensions for plot:

Figure 11.8

The Generalized Procrustes Options menu (Figure 11.8) allows you to select the type of scaling to be done. *Isotropic* scaling, which scales the all the dimensions of each configuration by an equal amount, takes place during the Procrustes analysis. The alternative is to scale each configuration prior to the analysis so that the trace of each matrix is one. If this *separate* scaling is used, the subsequent residuals represent pure lack-of-fit and the scaling factors given in the results represent differences in relative size/spread of the original (centred) configurations, whereas for overall isotropic scaling the scaling factor contains components of both size and lack-of-fit.

The Display boxes control the output:

Monitoring — gives monitoring information during the fitting process;

Column means — prints the column (i.e. attribute) means of the configurations;

Centroid — prints the latent roots and coordinates of the centroid configuration;

Individual configurations — prints rotations of the individual configurations to the principal axes;

Analysis — prints an analysis of variation for the configurations and entities (i.e. samples);

All of these, except monitoring, are shown in the output below.

Generalized Procrustes analysis

Isotropic scaling

Column means of the configurations

Configuration 1

```
        1         2         3         4
    4.750     5.375     4.250     0.500
```

Configuration 2

```
         1           2           3           4
     5.750       5.000       6.375       4.125
```

Configuration 3

```
         1           2           3           4
     3.875       1.000       4.500       5.250
```

Configuration 4

```
         1           2           3           4
     2.000       4.375       4.375       0.000
```

Configuration 5

```
         1           2           3           4
     3.250       4.250       3.625       0.000
```

Configuration 6

```
         1           2           3           4
     6.125       1.625       6.125       0.000
```

Rotation of centroid to principal axes

Latent roots

```
         1           2           3           4
     0.714       0.174       0.011       0.006
```

Percentage variance

```
         1           2           3           4
     78.88       19.25        1.22        0.65
```

Coordinates of the consensus configuration

```
                   1           2           3           4
         1    0.6173     -0.1389     -0.0072      0.0143
         2   -0.1303      0.1879      0.0438      0.0374
         3    0.3595      0.1338      0.0361     -0.0260
```

```
     4   -0.2390   -0.0142   -0.0166    0.0308
     5   -0.0600    0.0759   -0.0397    0.0094
     6   -0.2297    0.0764    0.0040   -0.0491
     7   -0.2663   -0.2994    0.0432   -0.0055
     8   -0.0516   -0.0213   -0.0636   -0.0113
```

Final coordinates for configuration 1

Variance

```
      1         2         3         4
  0.747     0.243     0.011     0.006
```

Percentage variance

```
      1         2         3         4
  74.19     24.12      1.12      0.56
```

Coordinates

```
               1         2         3         4
     1    0.5081   -0.0939   -0.0208    0.0072
     2   -0.0224    0.2261    0.0610    0.0496
     3    0.4854    0.2358    0.0555   -0.0105
     4   -0.2349    0.0622   -0.0049    0.0261
     5   -0.0872   -0.0149   -0.0407   -0.0158
     6   -0.2734    0.0376   -0.0274   -0.0453
     7   -0.3373   -0.3246    0.0151   -0.0073
     8   -0.0382   -0.1282   -0.0378   -0.0041
```

Rotation matrix

```
               1         2         3         4
     1     0.439     0.280     0.257     0.814
     2    -0.842     0.440     0.204     0.239
     3     0.283     0.852    -0.236    -0.372
     4     0.137     0.044     0.915    -0.378
```

Final coordinates for configuration 2

Variance

```
      1         2         3         4
  0.693     0.184     0.061     0.028
```

Percentage variance

```
      1         2         3         4
  71.83     19.03      6.28      2.86
```

Coordinates

	1	2	3	4
1	0.6147	-0.1257	-0.0145	0.0315
2	-0.1278	0.0857	0.0445	0.0361
3	0.3291	0.0049	0.0878	-0.0635
4	-0.1837	0.0129	-0.0916	0.0531
5	-0.1463	0.0992	-0.0848	0.0666
6	-0.3401	0.2011	0.0508	-0.1167
7	-0.1411	-0.3280	0.1315	-0.0165
8	-0.0048	0.0498	-0.1237	0.0094

Rotation matrix

	1	2	3	4
1	0.248	0.058	-0.250	-0.934
2	-0.716	0.359	-0.599	-0.007
3	-0.393	-0.906	-0.071	-0.141
4	0.522	-0.218	-0.757	0.327

Final coordinates for configuration 3

Variance

1	2	3	4
0.817	0.144	0.025	0.021

Percentage variance

1	2	3	4
81.17	14.27	2.46	2.10

Coordinates

	1	2	3	4
1	0.6009	-0.1247	0.0045	-0.0076
2	-0.2587	0.1799	0.0926	-0.0067
3	0.3982	0.1224	0.0138	0.0217
4	-0.3332	-0.0110	0.0251	0.1044
5	0.0943	0.1167	-0.0892	0.0108
6	-0.1722	0.0182	-0.0088	-0.0231
7	-0.2808	-0.2537	0.0386	-0.0043
8	-0.0485	-0.0480	-0.0765	-0.0951

Rotation matrix

	1	2	3	4
1	0.650	0.061	-0.752	0.084
2	0.589	-0.097	0.564	0.571
3	-0.369	0.607	-0.195	0.677
4	0.307	0.787	0.278	-0.458

Final coordinates for configuration 4

Variance

```
     1        2        3        4
 0.678    0.266    0.011    0.031
```

Percentage variance

```
     1        2        3        4
 68.78    26.98     1.12     3.13
```

Coordinates

```
                  1         2         3         4
        1    0.5996   -0.1990   -0.0120    0.0479
        2   -0.0664    0.2300   -0.0580    0.0902
        3    0.3452    0.1843    0.0596   -0.1009
        4   -0.3017   -0.0458   -0.0210    0.0291
        5   -0.0285    0.1086    0.0111   -0.0246
        6   -0.1461   -0.0293    0.0296   -0.0552
        7   -0.2182   -0.3409    0.0304   -0.0463
        8   -0.1840    0.0921   -0.0395    0.0597
```

Rotation matrix

```
                  1         2         3         4
        1     0.480     0.327    -0.403     0.707
        2    -0.770     0.602    -0.126     0.172
        3     0.421     0.728     0.261    -0.474
        4    -0.016     0.020     0.868     0.496
```

Final coordinates for configuration 5

Variance

```
     1        2        3        4
 0.885    0.133    0.010    0.000
```

Percentage variance

```
     1        2        3        4
 86.10    12.91     0.98     0.02
```

Coordinates

```
                  1         2         3         4
        1    0.7446   -0.1389    0.0040   -0.0006
        2   -0.1387    0.1825    0.0190    0.0044
        3    0.2660    0.1170    0.0028    0.0054
        4   -0.2672    0.0248    0.0444   -0.0030
        5   -0.0515    0.0985   -0.0609    0.0062
        6   -0.2672    0.0248    0.0444   -0.0030
```

```
       7     -0.3085     -0.2169     -0.0101     -0.0086
       8      0.0224     -0.0919     -0.0436     -0.0007
```

Rotation matrix

```
                   1           2           3           4
       1       0.092       0.734      -0.670       0.061
       2       0.794       0.353       0.495      -0.009
       3      -0.601       0.579       0.551      -0.013
       4      -0.006      -0.034       0.052       0.998
```

Final coordinates for configuration 6

Variance

```
           1           2           3           4
       0.709       0.263       0.025       0.010
```

Percentage variance

```
           1           2           3           4
       70.41       26.14        2.46        0.99
```

Coordinates

```
                   1           2           3           4
       1      0.6363     -0.1511     -0.0045      0.0073
       2     -0.1678      0.2228      0.1036      0.0510
       3      0.3333      0.1380     -0.0030     -0.0081
       4     -0.1135     -0.1284     -0.0513     -0.0249
       5     -0.1407      0.0472      0.0262      0.0131
       6     -0.1795      0.2060     -0.0644     -0.0509
       7     -0.3116     -0.3326      0.0538      0.0497
       8     -0.0565     -0.0018     -0.0605     -0.0371
```

Rotation matrix

```
                   1           2           3           4
       1       0.408       0.906      -0.066      -0.087
       2       0.686      -0.231       0.582       0.371
       3      -0.602       0.351       0.620       0.359
       4      -0.003       0.045      -0.522       0.852
```

Analysis of variation for the configurations

```
             Scaling    Residual       Total
   1           0.771       0.094       1.007
   2           0.838       0.126       0.965
   3           1.067       0.088       1.007
   4           1.148       0.104       0.985
   5           1.275       0.068       1.028
   6           1.228       0.085       1.007
```

Analysis of variation for the entities

```
        Consensus    Residual       Total
 1          2.405       0.037       2.442
 2          0.334       0.072       0.406
 3          0.894       0.076       0.970
 4          0.351       0.077       0.428
 5          0.066       0.069       0.136
 6          0.366       0.097       0.463
 7          0.975       0.057       1.032
 8          0.044       0.080       0.123


Initial within-configuration sum of squares                        463.750
Initial between-configuration sum of squares                       509.875
Final residual sum of squares                                        0.566
Number of steps to convergence 7
```

The first graph (Figure 11.9) plots the positions of the eight port-wines in the consensus configuration; the default is to display the first three principal axes. This allows you to study the similarities of the port-wines, as observed overall by the assessors.

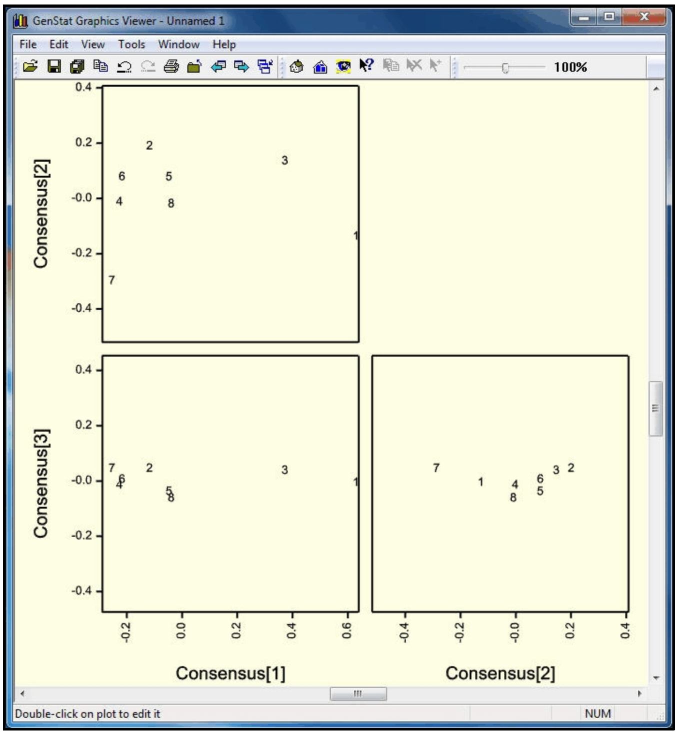


Figure 11.9

The individuals plot (Figure 11.10) is similar, but also shows how the points in each configuration are mapped to the equivalent points of the consensus. This allows you to assess the consistency of the assessors.

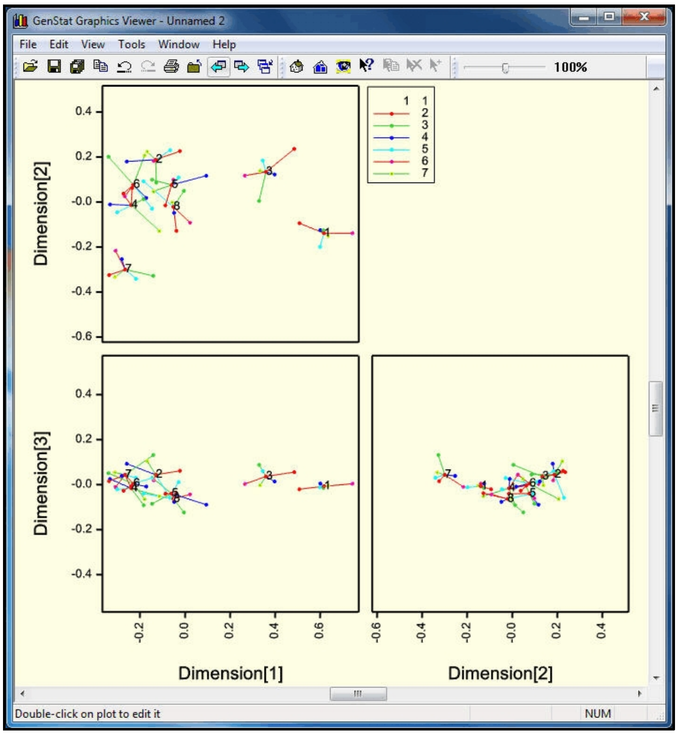


Figure 11.10

The projection plot (Figure 11.11) shows the fitted projections of the attributes, as observed by each assessor, onto the first two principal axes. A different colour is used for each attribute, taking the standard order of the Genstat pens (by default red, green, blue, cyan, mauve, yellow, brown etc.). Each line is numbered by its assessor. This shows how each of the attributes contributes to the consensus picture.

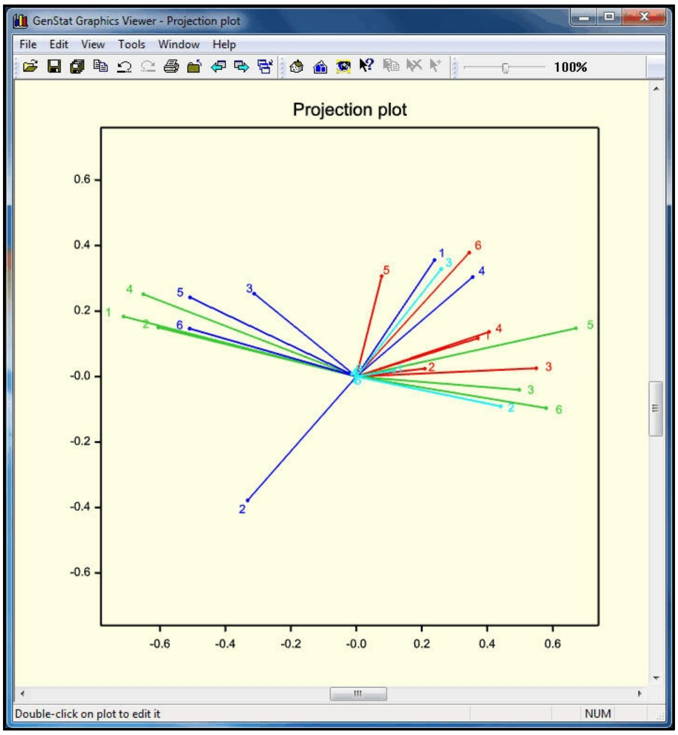
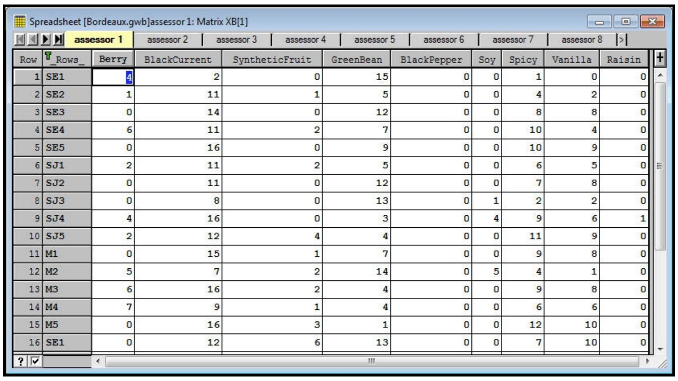

GenStat Graphics Viewer - Projection plot
File Edit View Tools Window Help
100%
Projection plot
0.6
0.4
0.2
-0.0
-0.2
-0.4
-0.6
-0.6
-0.4
-0.2
-0.0
0.2
0.4
0.6
Double-click on plot to edit it
NUM

Figure 11.11

Notice that the fitted attributes (plotted in cyan) for the null attributes of assessors 4-6 are negligible. The scores fitted to these null attributes can be regarded as representing random variation, and used as a yardstick for assessing the other scores. (However, if one of the null fitted scores were found to be noticeable, this would indicate a large Procrustes deviation between this assessor and the consensus −which might suggest that this assessor should treated separately.)

11.1 Practical

Genstat spreadsheet file `Bordeaux.gsh` contains results of a sensory assessment of aroma attributes of 24 Bordeaux wines. Do a generalized Procrustes analysis and examine a consensus plot of the first two dimensions. (Hint: use the Ten Berge method to speed convergence.) Rows 1-5 are St Estèphe wines, 6-10 are St Julien, 11-15 are Margaux, 16-20 are St Emilion, and 21-24 are regional Bordeaux wines (see the row labels of the matrices). Can you see this structure reflected in the plot?



Row	_Rows_	Berry	BlackCurrent	SyntheticFruit	GreenBean	BlackPepper	Soy	Spicy	Vanilla	Raisin
1	SE1	4	2	0	15	0	0	1	0	0
2	SE2	1	11	1	5	0	0	4	2	0
3	SE3	0	14	0	12	0	0	8	8	0
4	SE4	6	11	2	7	0	0	10	4	0
5	SE5	0	16	0	9	0	0	10	9	0
6	SJ1	2	11	2	5	0	0	6	5	0
7	SJ2	0	11	0	12	0	0	7	8	0
8	SJ3	0	8	0	13	0	1	2	2	0
9	SJ4	4	16	0	3	0	4	9	6	1
10	SJ5	2	12	4	4	0	0	11	9	0
11	M1	0	15	1	7	0	0	9	8	0
12	M2	5	7	2	14	0	5	4	1	0
13	M3	6	16	2	4	0	0	9	8	0
14	M4	7	9	1	4	0	0	6	6	0
15	M5	0	16	3	1	0	0	12	10	0
16	SE1	0	12	6	13	0	0	7	10	0

Figure 11.12

12 Other facilities

This chapter illustrates menus from most of the main areas of multivariate analysis provided by Genstat. Other menus are listed below with references to sections in the *Guide to the Genstat Command Language* describing the associated commands and methodology:

Discriminant analysis	Part 2 Section 6.5,
Factor analysis	Part 2 Section 6.11,
Correspondence analysis	Part 2 Section 6.13,
Canonical correlation analysis	Part 2 Section 6.9,
Redundancy analysis	Part 2 Section 6.14,
Canonical correspondence analysis	Part 2 Section 6.15,
Partial least squares regression	Part 2 Section 6.8, and
Multivariate analysis of distance	Part 2 Section 6.6.3.

Other multivariate facilities, not available through the menus, include ridge and principal-component regression (procedure `RIDGE`; Part 2 Section 6.7), analysis of skew symmetry (procedure `SKEWSYMMETRY`; Part 2 Section 6.17), the construction of identification keys (procedure `BKEY`; Part 2 Section 6.22) and random classification forests (procedure `BCFOREST`).

Index