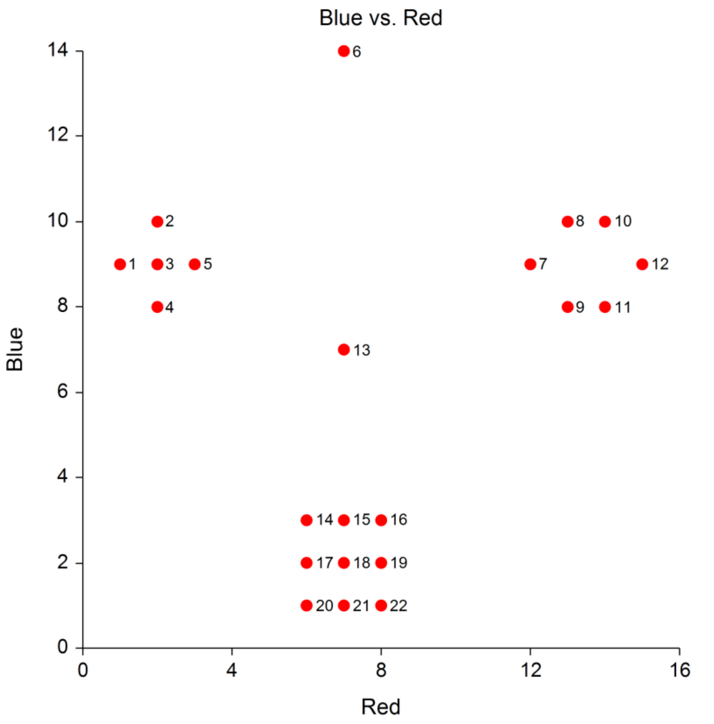
Chapter 445

Hierarchical Clustering / Dendrograms

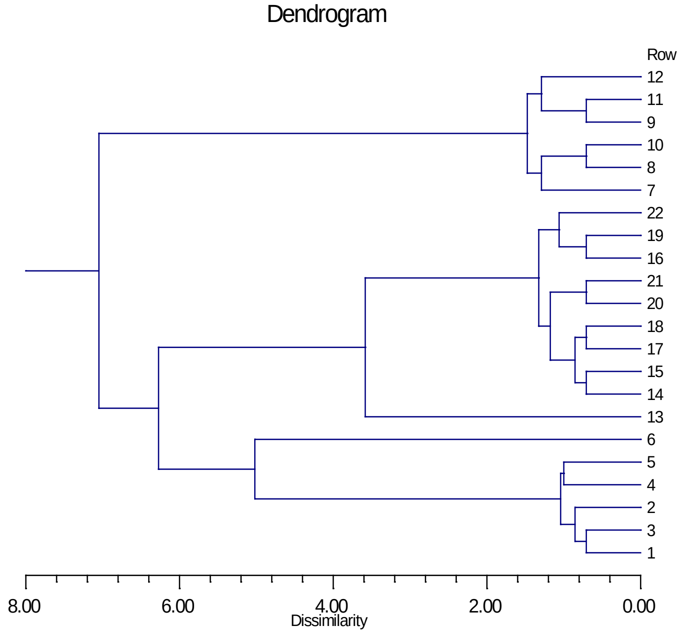
Introduction

The *agglomerative hierarchical clustering* algorithms available in this program module build a cluster hierarchy that is commonly displayed as a tree diagram called a *dendrogram*. They begin with each object in a separate cluster. At each step, the two clusters that are most similar are joined into a single new cluster. Once fused, objects are never separated. The eight methods that are available represent eight methods of defining the similarity between clusters.

Suppose we wish to cluster the bivariate data shown in the following scatter plot. In this case, the clustering may be done visually. The data have three clusters and two singletons, 6 and 13.



Following is a dendrogram of the results of running these data through the Group Average clustering algorithm.



The horizontal axis of the dendrogram represents the distance or dissimilarity between clusters. The vertical axis represents the objects and clusters. The dendrogram is fairly simple to interpret. Remember that our main interest is in similarity and clustering. Each joining (fusion) of two clusters is represented on the graph by the splitting of a horizontal line into two horizontal lines. The horizontal position of the split, shown by the short vertical bar, gives the distance (dissimilarity) between the two clusters.

Looking at this dendrogram, you can see the three clusters as three branches that occur at about the same horizontal distance. The two outliers, 6 and 13, are fused in rather arbitrarily at much higher distances. This is the interpretation.

In this example we can compare our interpretation with an actual plot of the data. Unfortunately, this usually will not be possible because our data will consist of more than two variables.

Dissimilarities

The first task is to form the distances (dissimilarities) between individual objects. This is described in the Medoid Clustering chapter and will not be repeated here.

Hierarchical Clustering Algorithms

The algorithm used by all eight of the clustering methods is outlined as follows. Let the distance between clusters *i* and *j* be represented as d_{ij} and let cluster *i* contain n_i objects. Let ***D*** represent the set of all remaining d_{ij}. Suppose there are *N* objects to cluster.

1. Find the smallest element d_{ij} remaining in ***D***.
2. Merge clusters *i* and *j* into a single new cluster, *k*.
3. Calculate a new set of distances d_{km} using the following distance formula.

   $$d_{km} = \alpha_i d_{im} + \alpha_j d_{jm} + \beta d_{ij} + \gamma |d_{im} - d_{jm}|$$

   Here *m* represents any cluster other than *k*. These new distances replace d_{im} and d_{jm} in ***D***. Also let $n_k = n_i + n_j$.

   Note that the eight algorithms available represent eight choices for $\alpha_i, \alpha_j, \beta$, and γ.
4. Repeat steps 1 - 3 until ***D*** contains a single group made up of all objects. This will require *N-1* iterations.

We will now give brief comments about each of the eight techniques.

Single Linkage

Also known as *nearest neighbor* clustering, this is one of the oldest and most famous of the hierarchical techniques. The distance between two groups is defined as the distance between their two closest members. It often yields clusters in which individuals are added sequentially to a single group.

The coefficients of the distance equation are

$$\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = -0.5.$$

Complete Linkage

Also known as furthest neighbor or maximum method, this method defines the distance between two groups as the distance between their two farthest-apart members. This method usually yields clusters that are well separated and compact.

The coefficients of the distance equation are

$$\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = 0.5.$$

Simple Average

Also called the weighted pair-group method, this algorithm defines the distance between groups as the average distance between each of the members, weighted so that the two groups have an equal influence on the final result.

The coefficients of the distance equation are

$$\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = 0.$$

Centroid

Also referred to as the unweighted pair-group centroid method, this method defines the distance between two groups as the distance between their centroids (center of gravity or vector average). The method should only be used with Euclidean distances.

The coefficients of the distance equation are

$$\alpha_i = \frac{n_i}{n_k}, \alpha_j = \frac{n_j}{n_k}, \beta = -\alpha_i\alpha_j, \gamma = 0.$$

Backward links may occur with this method. These are recognizable when the dendrogram no longer exhibits its simple tree-like structure in which each fusion results in a new cluster that is at a higher distance level (moves from right to left). With backward links, fusions can take place that result in clusters at a lower distance level (move from left to right). The dendrogram is difficult to interpret in this case.

Median

Also called the weighted pair-group centroid method, this defines the distance between two groups as the weighted distance between their centroids, the weight being proportional to the number of individuals in each group. Backward links (see discussion under Centroid) may occur with this method. The method should only be used with Euclidean distances.

The coefficients of the distance equation are

$$\alpha_i = \alpha_j = 0.5, \beta = -0.25, \gamma = 0.$$

Group Average

Also called the unweighted pair-group method, this is perhaps the most widely used of all the hierarchical cluster techniques. The distance between two groups is defined as the average distance between each of their members.

The coefficients of the distance equation are

$$\alpha_i = \frac{n_i}{n_k}, \alpha_j = \frac{n_j}{n_k}, \beta = 0, \gamma = 0.$$

Ward’s Minimum Variance

With this method, groups are formed so that the pooled within-group sum of squares is minimized. That is, at each step, the two clusters are fused which result in the least increase in the pooled within-group sum of squares.

The coefficients of the distance equation are

$$\alpha_i = \frac{n_i + n_m}{n_k + n_m}, \alpha_j = \frac{n_j + n_m}{n_k + n_m}, \beta = \frac{-n_m}{n_k + n_m}, \gamma = 0.$$

Flexible Strategy

Lance and Williams (1967) suggested that a continuum could be made between single and complete linkage. The program lets you try various settings of these parameters which do not conform to the constraints suggested by Lance and Williams.

The coefficients of the distance equation should conform to the following constraints

$$\alpha_i = 1 - \beta - \alpha_j, \alpha_j = 1 - \beta - \alpha_i, -1 \leq \beta \leq 1, \gamma = 0.$$

One interesting exercise is to vary these values, trying to find the set that maximizes the cophenetic correlation coefficient.

Goodness-of-Fit

Given the large number of techniques, it is often difficult to decide which is best. One criterion that has become popular is to use the result that has largest *cophenetic correlation coefficient*. This is the correlation between the original distances and those that result from the cluster configuration. Values above 0.75 are felt to be good. The Group Average method appears to produce high values of this statistic. This may be one reason that it is so popular.

A second measure of goodness of fit called *delta* is described in Mather (1976). These statistics measure degree of distortion rather than degree of resemblance (as with the cophenetic correlation). The two delta coefficients are given by

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} \left| d_{jk} - d_{jk}^* \right|^{1/A}}{\sum_{j<k} \left(d_{jk}^* \right)^{1/A}} \right]^A$$

where *A* is either 0.5 or 1 and d_{ij}^* is the distance obtained from the cluster configuration. Values close to zero are desirable.

Mather (1976) suggests that the Group Average method is the safest to use as an exploratory method, although he goes on to suggest that several methods should be tried and the one with the largest cophenetic correlation be selected for further investigation.

Number of Clusters

These techniques do not let you explicitly set the number of clusters. Instead, you pick a distance value that will yield an appropriate number of clusters. This will be discussed further when we discuss the Dendrogram and the Linkage report.

Limitations and Criticisms

We have attempted problems with up to 1,000 objects. Running times will vary with computer speed, with larger problems running several hours. Problems with 100 objects or less should run in a few seconds.

Hierarchical clustering methods are popular because they are relatively simple to understand and implement. However, this simplicity yields one of their strongest criticisms. Once two objects are joined, they can never be separated. As Kaufman (1990) complains, “once the damage is done, it can never be repaired.”

Data Structure

The data are entered in the standard columnar format in which each column represents a single variable.

The data given in the following table contain information on twelve superstars in basketball. The stats are on a per game basis for games played through the 1989 season.

BBall Dataset (Subset)

Player	Height	FgPct	Points	Rebounds
Jabbar K.A.	86.0	55.9	24.6	11.2
Barry R	79.0	44.9	23.2	6.7
Baylor E	77.0	43.1	27.4	13.5
Bird L	81.0	50.3	25	10.2
Chamberlain W	85.0	54.0	30.1	22.9
Cousy B	72.5	37.5	18.4	5.2
Erving J	78.5	50.6	24.2	8.5
Johnson M	81.0	53.0	19.5	7.4
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Data Input Formats

A number of input formats are available.

Raw Data

The variables are in the standard format in which each row represents an object, and each column represents a variable.

Distances

The variables containing a distance matrix are specified in the Interval Variables option. Note that this matrix contains the distances between each pair of objects. Each object is represented by a row and the corresponding column. Also, the matrix must be complete. You cannot use only the lower triangular portion, for example.

Correlations - 1

The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

$$d_{ij} = \frac{1 - r_{ij}}{2}$$

Correlations - 2

The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

$$d_{ij} = 1 - \left| r_{ij} \right|$$

Correlations - 3

The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

$$d_{ij} = 1 - r_{ij}^2$$

Note that all three types of correlation matrices must be completely specified. You cannot specify only the lower or upper triangular portions. Also, the rows correspond to variables. That is, the values along the first row represent the correlations of the first variable with each of the other variables. Hence, you cannot rearrange the order of the matrix.

Missing Values

When an observation has missing values, appropriate adjustments are made so that the average dissimilarity across all variables with non-missing data is computed. Hence, rows with missing values are not omitted unless all variables have missing values. Note that the distances require that at least one variable have non-missing values for each pair of rows.

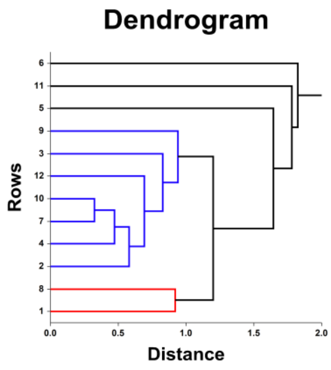
Dendrogram Window Options

This section describes the specific options available on the Dendrogram window, which is displayed when the Dendrogram Format button is clicked. Common options, such as axes, labels, legends, and titles are documented in the Graphics Components chapter.

Dendrogram Plot Tab

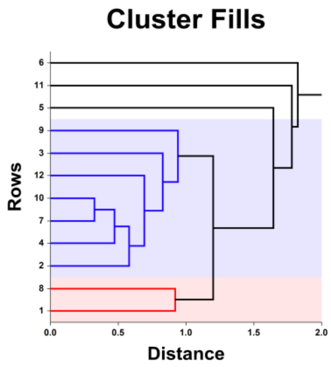
Lines Section

You can modify the color, width, and pattern of dendrogram lines. Lines that join at a distance less than the cutoff value are said to be "clustered." Other lines are "non-clustered."



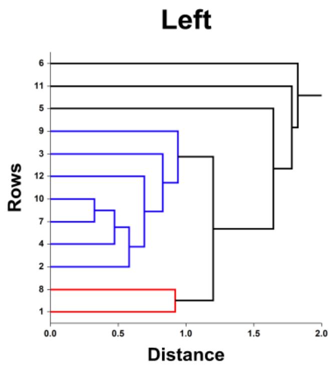
Fills Section

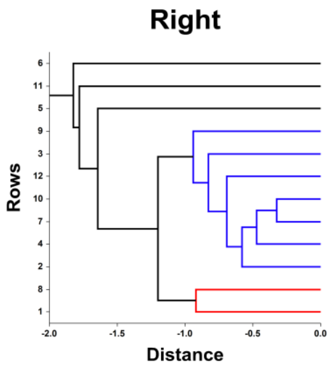
You can use a different fill color for each cluster and each set of contiguous non-clustered.

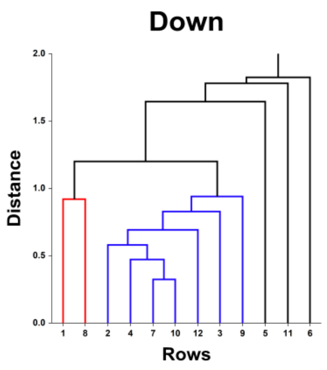


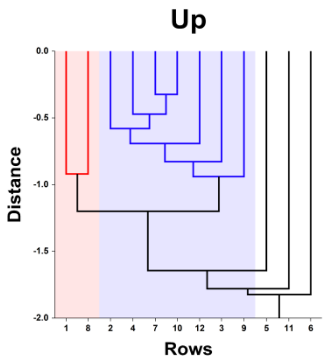
Orientation Section

You can specify where the cluster lines end.


Left


Right


Down


Up

Titles, Legend, Numeric Axis, Cluster (Group) Axis, Grid Lines, and Background Tabs

Details on setting the options in these tabs are given in the Graphics Components chapter.

Example 1 – Hierarchical Clustering

This section presents an example of how to run a cluster analysis of the basketball superstars data. The data are found in the BBall dataset.

Setup

To run this example, complete the following steps:

1 Open the BBall example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **BBall** and click **OK**.

2 Specify the Hierarchical Clustering / Dendrograms procedure options

- Find and open the **Hierarchical Clustering / Dendrograms** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Interval Variables..**Height,FgPct,Points,Rebounds**
Label Variable...**Player**

Reports Tab

Distance Between Rows.................................**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Cluster Detail

```
Cluster Detail
───────────────────────────────────────────────────────────────────────
Row      Cluster    Player
1        1          Jabbar K.A.
8        1          Johnson M
2        2          Barry R
3        2          Baylor E
4        2          Bird L
7        2          Erving J
9        2          Jordan M
10       2          Robertson O
12       2          West J
5                   Chamberlain W
6                   Cousy B
11                  Russell B
───────────────────────────────────────────────────────────────────────
```

This report displays the cluster number associated with each row. The report is sorted by row number within cluster number. The cluster number of rows that cannot be classified are left blank. The cluster configuration depends on the Cluster Cutoff value that was used.

Linkage

```
Linkage
───────────────────────────────────────────────────────────────────────
                               Distance
          Number of    ─────────────────────────────
Link      Clusters          Value    Bar                               Rows Linked
11        1              1.822851    |IIIIIIIIIIIIIIIIIIIIIIIIIIIIII   1,8,2,4,7,10,12,3,9,5,11,6
10        2              1.780810    |IIIIIIIIIIIIIIIIIIIIIIIIIIIII    1,8,2,4,7,10,12,3,9,5,11
9         3              1.642553    |IIIIIIIIIIIIIIIIIIIIIIIIIII      1,8,2,4,7,10,12,3,9,5
8         4              1.199225    |IIIIIIIIIIIIIIIIIIII             1,8,2,4,7,10,12,3,9
7         5              0.941566    |IIIIIIIIIIIIIII                  2,4,7,10,12,3,9
6         6              0.919016    |IIIIIIIIIIIIIII                  1,8
5         7              0.826883    |IIIIIIIIIIIIII                   2,4,7,10,12,3
4         8              0.693822    |IIIIIIIIIII                      2,4,7,10,12
3         9              0.579517    |IIIIIIIIII                       2,4,7,10
2         10             0.470534    |IIIIIIII                         4,7,10
1         11             0.325592    |IIIII                            7,10

Clustering Fit Metrics
Cophenetic Correlation    0.830472
Delta(0.5)                0.171620
Delta(1.0)                0.223057
───────────────────────────────────────────────────────────────────────
```

This report displays the subgroup that is formed at each fusion that took place during the cluster analysis. The links are displayed in reverse order so that you can quickly determine an appropriate number of clusters to use. It displays the distance level at which the fusion took place. It will let you precisely determine the best value of the Cluster Cutoff value.

For example, looking down the Distance Value column of the report, you can see that the cutoff value that we used (the default value is 1.0) occurs between Links 7 and 8. Hence, the cutoff value of 1.0 results in five clusters. Looking at the Cluster Detail report (above), you will see that we obtained two real clusters and three outliers. These outliers are called clusters even though they consist of only one individual.

The cophenetic correlation and the two delta goodness of fit statistics are reported at the bottom of this report. As discussed earlier, these values let you compare the fit of various cluster configurations.

Link

This is the sequence number of the fusion.

Number of Clusters

This is the number of clusters that would result if the Cluster Cutoff value were set to the corresponding Distance Value or higher. Note that this number includes outliers.

Distance: Value

This is the distance value between the two joining clusters that is used by the algorithm. Normally, this value is monotonically increasing. When backward linking occurs, this value will no longer exhibit a strictly increasing behavior.

As discussed above, these values are used to determine an appropriate number of clusters.

Distance: Bar

This is a bar graph of the Distance Values. Choose the number of clusters by finding a jump in the decreasing pattern shown in this bar chart.

Rows Linked

These are the rows that were joined at this step. Remember that the links are presented in reverse order, so, in our example, rows 7 and 10 were joined first, row 4 was added, and so on.

Cophenetic Correlation

This is the Pearson correlation between the actual distances and the predicted distances based on this particular hierarchical configuration. A value of 0.75 or above needs to be achieved in order for the clustering to be considered useful.

Delta (0.5, 1)

These are the values of the goodness of fit deltas. When comparing to clustering configurations, the configuration with the smallest delta value fits the data better.

Distance Between Rows

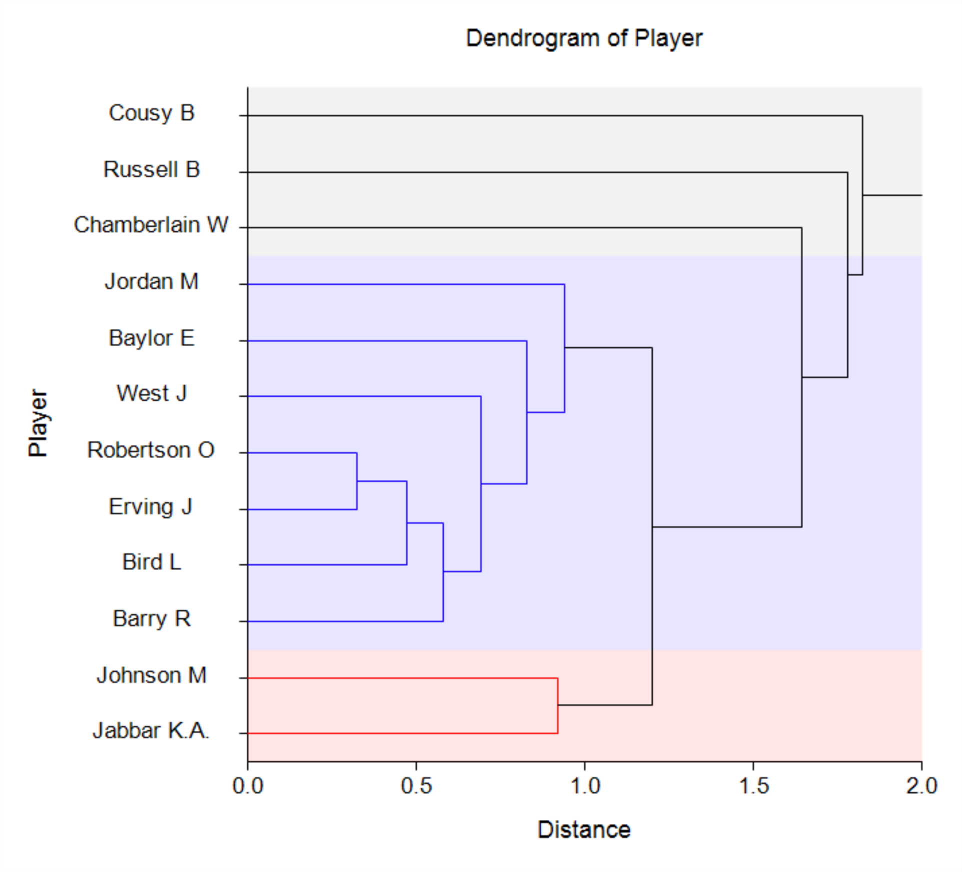
Distance Between Rows

Row		Distance		Difference	
First	Second	Actual	Dendrogram	Actual	Percent
1	2	1.427013	1.199225	0.227788	15.96%
1	3	1.703276	1.199225	0.504050	29.59%
1	4	0.833498	1.199225	-0.365727	-43.88%
1	5	1.126296	1.642553	-0.516257	-45.84%
1	6	2.575167	1.822851	0.752316	29.21%
1	7	1.100763	1.199225	-0.098462	-8.94%
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

This report displays the actual and predicted distance for each pair of rows. It also includes their difference and percent difference. Since the report grows very long for even a modest number of rows, it is usually omitted.

Dendrogram

Dendrogram



This section displays the dendrogram which visually displays a particular cluster configuration. Rows that are close together (have small dissimilarity) will be linked near the right side of the plot. For example, we notice the Oscar Robertson and Julius Erving are very similar.

Rows that link up near the left side are very different. For example, Bob Cousy appears to be quite different from any of the other players.

The number of clusters that will be formed at a particular Cluster Cutoff value may be quickly determined from this plot by drawing a vertical line at that value and counting the number of lines that the vertical line intersects. For example, you can see that if we draw a vertical line at the value 1.0, five clusters will result. One cluster will contain two objects, one will contain seven objects, and three clusters each will contain only one object.

We strongly recommend that you compare the dendrograms from several different methods and on several different datasets with known cluster patterns so that you can get the feel of the technique.