Chapter 16

Cluster Analysis

16.1. BACKGROUND

16.1.1. Cluster Analysis vs. Discriminant Analysis

Cluster analysis deals with separating data into groups whose identities are not known in advance. This more limited state of knowledge is in contrast to the situation for discrimination methods, which require a training data set in which group memberships are known. In modern statistical parlance, cluster analysis is an example of unsupervised learning, whereas discriminant analysis is an instance of supervised learning. In general, in cluster analysis even the correct number of groups into which the data should be sorted is not known ahead of time. Rather, it is the degree of similarity and difference between individual observations $\boldsymbol{x}$ that is used to define the groups and to assign group membership. Examples of use of cluster analysis in the meteorological and climatological literature include grouping daily weather observations into synoptic types (Kalkstein et al., 1987), defining weather regimes from upper-air flow patterns (Mo and Ghil, 1988; Molteni et al., 1990), grouping members of forecast ensembles (Legg et al., 2002; Molteni et al., 1996; Tracton and Kalnay, 1993), and forecast evaluation (Kücken and Gerstengarbe, 2009; Marzban and Sandgathe, 2008). Gong and Richman (1995) have compared various clustering approaches in a climatological context and catalog the literature with applications of clustering to atmospheric data through 1993. Romesburg (1984) contains a general-purpose overview.

Given a sample of data vectors $\boldsymbol{x}$ defining the rows of a $(n \times K)$ data matrix $[X]$, a cluster analysis will define groups and assign group memberships at varying levels of aggregation. However, the method is primarily an exploratory data analysis tool, rather than an inferential tool. Indeed, the statistical underpinnings of most clustering approaches are vague. In most cases they are uncoupled from possible differences in underlying distributions, and useful associated inferential procedures are not generally available (e.g., Chacón, 2015). Unlike discriminant analysis, the procedure does not contain rules for assigning membership to future observations. However, a cluster analysis can bring out groupings in the data that might otherwise be overlooked, possibly leading to an empirically useful stratification of the data, or helping to suggest physical bases for observed structure in the data. For example, cluster analyses have been applied to geopotential height data in order to try to identify distinct atmospheric flow regimes (e.g., Cheng and Wallace, 1993; Mo and Ghil, 1988).

16.1.2. Distance Measures and the Distance Matrix

The idea of distance is central to the idea of the clustering of data points. Clusters should be composed of points separated by small distances, relative to the distances between clusters. However, there is a wide

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00016-X

variety of plausible definitions for distance in this context, and the results of a cluster analysis may depend quite strongly on the distance measure chosen.

The most intuitive and commonly used distance measure in cluster analysis is Euclidean distance (Equation 11.6) in the K-dimensional space of the data vectors. Euclidean distance is by no means the only available choice for measuring distance between points or clusters, and in some instances may be a poor choice. In particular, if the elements of the data vectors are unlike variables with inconsistent measurement units, the variable with the largest values will tend to dominate the Euclidean distance. A more general alternative is the weighted Euclidean distance between two vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$,

$$d_{i,j} = \left[\sum_{k=1}^{K} w_k \left(x_{i,k} - x_{j,k} \right)^2 \right]^{1/2}. \tag{16.1}$$

For $w_k = 1$ for each $k = 1, \ldots, K$, Equation 16.1 reduces to the ordinary Euclidean distance. If the weights are the reciprocals of the corresponding variances, that is, $w_k = 1/s_{k,k}$, the resulting function of the standardized variables is called the *Karl-Pearson distance*. Other choices for the weights are also possible. For example, if one or more of the K variables in $\boldsymbol{x}$ contains large outliers, it might be better to use weights that are reciprocals of the ranges of each of the variables.

Euclidean distance and Karl-Pearson distance are the most frequent choices in cluster analysis, but other alternatives are also possible. One alternative is to use the Mahalanobis distance (Equation 11.91), although deciding on an appropriate (pooled) dispersion matrix $[S]$ may be difficult, since group memberships are not known in advance. A yet more general form of Equation 16.1 is the *Minkowski metric*,

$$d_{i,j} = \left[\sum_{k=1}^{K} w_k \left| x_{i,k} - x_{j,k} \right|^{\lambda} \right]^{1/\lambda}, \quad \lambda \geq 1. \tag{16.2}$$

Again, the weights w_k can equalize the influence of variables with incommensurate units. For $\lambda = 2$, Equation 16.2 reduces to the weighted Euclidean distance in Equation 16.1. For $\lambda = 1$, Equation 16.2 is known as the *city-block distance*.

The angles between pairs of vectors (Equation 11.15), or its cosine, are other possible choices for distance measures, as are the many alternatives presented in Mardia et al. (1979) or Romesburg (1984). Tracton and Kalnay (1993) have used the anomaly correlation (Equation 9.95) to group members of forecast ensembles, and the ordinary Pearson correlation sometimes is used as a clustering criterion as well. These latter two criteria are inverse distance measures, which should be maximized within groups and minimized between groups.

Having chosen a distance measure to quantify dissimilarity or similarity between pairs of vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, the next step in cluster analysis is to calculate the distances between all $n(n-1)/2$ possible pairs of the n observations. It can be convenient, either organizationally or conceptually, to arrange these into a $(n \times n)$ matrix of distances, $[\Delta]$, called the *distance matrix*. This symmetric matrix has zeros along the main diagonal, indicating zero distance between each $\boldsymbol{x}$ and itself.

16.2. HIERARCHICAL CLUSTERING

16.2.1. Agglomerative Methods Using the Distance Matrix

The most commonly implemented cluster analysis procedures are *hierarchical* and *agglomerative*. That is, they construct a hierarchy of sets of groups, each level of which is formed by merging one pair from

the collection of previously defined groups. These procedures begin by considering that the n observations of $\boldsymbol{x}$ have no group structure or, equivalently, that the data set consists of n groups containing one observation each. The first step is to find the two groups (i.e., data vectors) that are closest in their K-dimensional space and to combine them into a new group. There are then $n-1$ groups, one of which has two members. On each subsequent step, the two groups that are closest are merged to form a larger group. Once a data vector $\boldsymbol{x}$ has been assigned to a group, it is not removed. Its group membership changes only when the group to which it has been assigned is merged with another group. This process continues until, at the final, $(n–1)^{st}$, step all n observations have been aggregated into a single group.

The n-group clustering at the beginning of this process and the one-group clustering at the end of this process are neither useful nor enlightening. Hopefully, however, a natural clustering of the data into a workable number of informative groups will emerge at some intermediate stage. That is, we hope that the n data vectors cluster or clump together in their K-dimensional space into some number G, $1 < G < n$, groups that reflect similar data-generating processes. The ideal result is a division of the data that both minimizes differences between members of a given cluster and maximizes differences between members of different clusters.

Distances between pairs of points can be unambiguously defined and stored in a distance matrix. However, even after choosing a distance measure and calculating a distance matrix there are alternative definitions for distances between groups of points if the groups contain more than a single member. The choice made for the distance measure, together with the criterion used to define cluster-to-cluster distances, essentially define the method of clustering. A few of the most common definitions and a less common but potentially interesting definition for intergroup distances based on the distance matrix are:

- *Single-linkage*, or *minimum-distance clustering*. Here the distance between clusters G_1 and G_2 is the smallest distance between one member of G_1 and one member of G_2. That is,

$$d_{G_1,G_2} = \min_{i \in G_1, j \in G_2} \left(d_{i,j}\right). \tag{16.3}$$

- *Complete-linkage*, or *maximum-distance clustering* groups data points on the basis of the largest distance between points in the two groups G_1 and G_2,

$$d_{G_1,G_2} = \max_{i \in G_1, j \in G_2} \left(d_{i,j}\right). \tag{16.4}$$

- *Average-linkage* clustering defines cluster-to-cluster distance as the average of distances between all possible pairs of points in the two groups being compared. If G_1 contains n_1 points and G_2 contains n_2 points, this measure for the distance between the two groups is

$$d_{G_1,G_2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d_{i,j}. \tag{16.5}$$

- *Centroid* clustering compares distances between the centroids, or vector averages, of pairs of clusters. According to this measure the distance between G_1 and G_2 is

$$d_{G_1,G_2} = \|\overline{\boldsymbol{x}}_{G_1} - \overline{\boldsymbol{x}}_{G_2}\|, \tag{16.6}$$

  where the vector means are taken over all members of each of the groups separately.

- *Minimax linkage* (Bien and Tibshirani, 2011) is a relatively little known but potentially informative and useful alternative to the four more common distance measures,

$$d_{G_1,G_2} = \min_{\boldsymbol{x}_i \in G_1 \cup G_2} \left[\max_{\boldsymbol{x}_j \in G_1 \cup G_2} \left(d_{i,j} \right) \right]. \tag{16.7}$$

For each of the $n_1 + n_2$ vectors $\boldsymbol{x}_i$ in a pair of groups that are candidates for merger, the expression inside the square brackets defines the maximum among distances to other members of the potentially merged group. The minimax distance in Equation 16.7 is then the smallest of these $n_1 + n_2$ maximized pairwise distances. Use of minimax linkage allows the potentially informative definition of prototype vectors, which provides a natural single-vector characterization of the group. These prototype vectors are exactly the $\boldsymbol{x}_i$ satisfying Equation 16.7 for each group.

Figure 16.1a illustrates single-linkage, complete-linkage, and centroid clustering for two hypothetical groups G_1 and G_2 in a $K = 2$-dimensional space. The open circles denote data points, of which there are $n_1 = 2$ in G_1 and $n_2 = 3$ in G_2. The centroids of the two groups are indicated by the solid circles. The single-linkage distance between G_1 and G_2 is the distance $d_{2,3}$ between the closest pair of points in the two groups. The complete-linkage distance is that between the most distant pair, $d_{1,5}$. The centroid distance is the distance between the two vector means $||\overline{\boldsymbol{x}}_{G_1} - \overline{\boldsymbol{x}}_{G_2}||$. The average-linkage distance can also be visualized in Figure 16.1 as the average of the six possible distances between individual members of G_1 and G_2, $(d_{1,5} + d_{1,4} + d_{1,3} + d_{2,5} + d_{2,4} + d_{2,3})/6$.

Figure 16.1b indicates that the minimax distance for the merger of the two groups is $d_{1,3}$, which is the largest distance between $\boldsymbol{x}_3$ and any of the other points, but is the smallest of these maximum distances among the five points. Accordingly, the prototype vector for the merged group is $\boldsymbol{x}_3$. The prototype vector within G_2 alone is $\boldsymbol{x}_4$, because $d_{4,5}$ is smaller than the other two maximized distances within G_2, which are both $d_{3,5}$.

The results of a cluster analysis can depend strongly on which definition is chosen for the distances between clusters. Single-linkage clustering rarely is used, because it is susceptible to *chaining*, or the production of a few large clusters, which are formed by virtue of nearness to opposite edges of a cluster

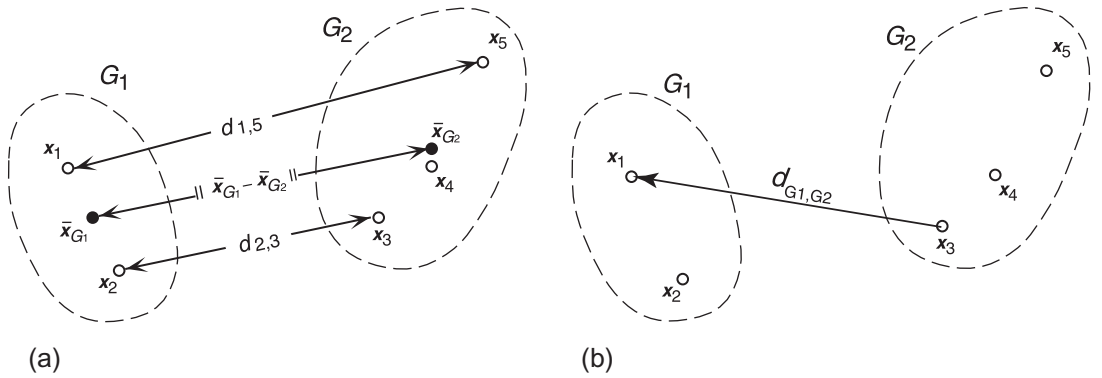

FIGURE 16.1 Illustration of four measures of the distance in $K = 2$-dimensional space, between a cluster G_1 containing the two elements $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and a cluster G_2 containing the elements $\boldsymbol{x}_3$, $\boldsymbol{x}_4$, and $\boldsymbol{x}_5$. The data points are indicated by open circles, and centroids of the two groups are indicated by the solid circles. (a) According to the maximum-distance, or complete-linkage criterion, the distance between the two groups is $d_{1,5}$, or the greatest distance between all of the six possible pairs of points in the two groups. The minimum-distance, or single-linkage criterion computes the distance between the groups as equal to the distance between the nearest pair of points, or $d_{2,3}$. According to the centroid method, the distance between the two clusters is the distance between the sample means of the points contained in each. (b) The minimax distance between the two groups is indicated by the arrow, which originates at the prototype vector for the merged groups, $\boldsymbol{x}_3$.

of points to be merged at different steps. At the other extreme, complete-linkage clusters tend to be more numerous, as the criterion for merging clusters is more stringent. Average-distance clustering is usually intermediate between these two extremes and appears to be the most commonly used approach to hierarchical clustering based on the distance matrix. Hastie et al. (2009) argue that average-distance clustering is the only statistically consistent method of the three, meaning that as the sample size becomes arbitrarily large, the group-average dissimilarities approach true population values.

16.2.2. Ward's Minimum Variance Method

Ward's minimum variance method, or simply Ward's method, is a popular hierarchical clustering method that does not operate on a distance matrix. As a hierarchical method, it begins with n single-member groups, and merges two groups at each step, until all the data are in a single group after n–1 steps. However, the criterion for choosing which pair of groups to merge at each step is that, among all possible ways of merging two groups, the pair to be merged is chosen that minimizes the sum of squared distances between the points and the centroids of their respective groups, summed over the resulting groups. That is, among all possible ways of merging two of $G+1$ groups to make G groups, that merger is made that minimizes

$$W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left\| \boldsymbol{x}_i - \overline{\boldsymbol{x}}_g \right\|^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{k=1}^{K} \left(x_{i,k} - \overline{x}_{g,k} \right)^2. \tag{16.8}$$

In order to implement Ward's method to choose the best pair from $G+1$ groups to merge, Equation 16.8 must be calculated for all of the $G(G+1)/2$ possible pairs of existing groups. For each trial pair, the centroid, or group mean, for the trial merged group is recomputed using the data for both of the previously separate groups, before the squared distances are calculated. In effect, Ward's method minimizes the sum, over the K dimensions of $\boldsymbol{x}$, of within-groups variances. At the first (n-group) stage this variance is zero, and at the last (1-group) stage this variance is tr[S_x], so that $W = n$ tr[S_x]. For data vectors whose elements have incommensurate units, operating on nondimensionalized values (dividing by standard deviations) will prevent artificial domination of the procedure by one or a few of the K variables.

16.2.3. The Dendrogram or Tree Diagram

The progress and intermediate results of a hierarchical cluster analysis are conventionally illustrated using a *dendrogram* or tree diagram. Beginning with the "twigs" at the beginning of the analysis, when each of the n observations $\boldsymbol{x}$ constitutes its own cluster, one pair of "branches" is joined at each step as the closest two clusters are merged. The distances between these clusters before they are merged are also indicated in the diagram by the distances of the points of merger from the initial n-cluster stage of the twigs.

Figure 16.2 illustrates a simple dendrogram, reflecting the clustering of the five points plotted as open circles in Figure 16.1a. The analysis begins at the left of Figure 16.2, when all five points constitute separate clusters. At the first stage, the closest two points, $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$, are merged into a new cluster. Their distance $d_{3,4}$ is proportional to the distance between the vertical bar joining these two points and the left edge of the figure. At the next stage, the points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are merged into a single cluster because the distance between them is smallest of the six distances among the four clusters that existed at the previous stage. The distance $d_{1,2}$ is necessarily larger than the distance $d_{3,4}$, since $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ were not chosen for merger on the first step, and the vertical line indicating the distance between them is plotted further to the

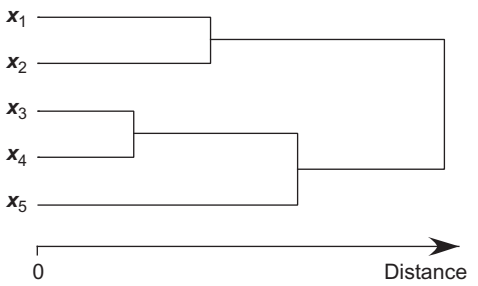

FIGURE 16.2 Illustration of a dendrogram or tree diagram, for a clustering of the five points plotted as open circles in Figure 16.1a. The results of the four clustering steps are indicated as the original five lines are progressively joined from left to right, with the distances between joined clusters indicated by the positions of the vertical lines.

right in Figure 16.2 than the distance between $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$. The third step merges $\boldsymbol{x}_5$ and the pair $(\boldsymbol{x}_3, \boldsymbol{x}_4)$, to yield the two-group stage indicated by the dashed ovals in Figure 16.1.

When minimax linkage is used to define distances, the interpretability of the dendrogram can be enhanced by indicating the prototype vector for each newly combined pair of groups at the point of the dendrogram indicating that merger. If the $\boldsymbol{x}$ vectors represent maps or other graphical objects, thumbnail plots of the prototypes can be shown. Otherwise, the name or case number of the prototype can be used. Showing prototypes only for the larger groups, or for the groups that are candidates for a final solution, might reduce clutter in the resulting plot.

16.2.4. How Many Clusters?

A hierarchical cluster analysis will produce a different grouping of n observations at each of the n–1 steps. At the first step each observation is in a separate group, and after the last step all the observations are in a single group. An important practical problem in cluster analysis is the choice of which intermediate stage will be chosen as the final solution. That is, we need to choose the level of aggregation in the tree diagram at which to stop further merging of clusters. The principle guiding this choice is to find that level of clustering that maximizes similarity within clusters and minimizes similarity between clusters, but in practice the best number of clusters for a given problem is usually not obvious. Generally the stopping point will require a subjective choice that will depend to some degree on the goals of the analysis.

One approach to the problem of choosing the best number of clusters is through summary statistics that relate to concepts in discrimination presented in Chapter 15. Several such criteria are based on the within-groups covariance matrix (Equation 15.16), either alone or in relation to the "between-groups" covariance matrix (Equation 15.18). Some of these objective stopping criteria are discussed in Jolliffe et al. (1986) and Fovell and Fovell (1993), who also provide references to the broader literature on such methods.

A traditional subjective approach to determination of the stopping level is to inspect a plot of the distances between merged clusters as a function of the stage of the analysis. When similar clusters are being merged early in the process, these distances are small and they increase relatively little from step to step. Late in the process there may be only a few clusters, separated by large distances. If a point can be discerned where the distances between merged clusters jump markedly, the process can be stopped just before these distances become large.

Wolter (1987) suggests a Monte Carlo approach, where sets of random numbers simulating the real data are subjected to cluster analysis. The distributions of clustering distances for the random numbers can be compared to the actual clustering distances for the data of interest. The idea here is that genuine clusters in the real data should be closer than clusters in the random data and that the clustering algorithm should be stopped at the point where clustering distances are greater than for the analysis of the random

data. Similarly, Tibshirani et al. (2001) propose defining the stopping point as that exhibiting the maximum difference between the logs of the distances W in Equation 16.8 to those obtained by averaging the results of many cluster analyses of K-dimensional uniform random numbers.

Example 16.1. A Cluster Analysis in Two Dimensions

The mechanics of cluster analysis are easiest to see when the data vectors have only $K = 2$ dimensions. Consider the data in Table 15.1, where these two dimensions are average July temperature and average July precipitation. These data were collected into three groups for use in the discriminant analysis worked out in Example 15.2. However, the point of a cluster analysis is to try to discern group structure within a data set, without prior knowledge or information about the nature of that structure. Therefore for purposes of a cluster analysis, the data in Table 15.1 should be regarded as consisting of $n = 28$ observations of two-dimensional vectors $\boldsymbol{x}$, whose natural groupings we would like to discern.

Because the temperature and precipitation values have different physical units, it is advisable to divide by the respective standard deviations before subjecting them to a clustering algorithm. That is, the temperature and precipitation values are divided by 4.42 °F and 1.36 in., respectively. The result is that the analysis is done using the Karl-Pearson distance, and the weights in Equation 16.1 are $w_1 = 4.42^{-2}$ and $w_2 = 1.36^{-2}$. The reason for this treatment of the data is to avoid the same kind of problem that can occur when conducting a principal component analysis using unlike data, where a variable with a much higher variance than the others will dominate the analysis even if that high variance is an artifact of the units of measurement. For example, if the precipitation had been expressed in millimeters there would be apparently more distance between points in the direction of the precipitation axis, and a clustering algorithm would focus on precipitation differences to define groups. If the precipitation were expressed in meters there would be essentially no distance between points in the direction of the precipitation axis, and a clustering algorithm would separate points almost entirely on the basis of the temperatures.

Figure 16.3 shows the results of clustering the data in Table 15.1, using the complete-linkage clustering criterion in Equation 16.4. On the left is a tree diagram for the process, with the individual stations listed at the bottom as the leaves. There are 27 horizontal lines in this tree diagram, each of which represents the merger of the two clusters it connects. At the first stage of the analysis the two closest points (Springfield and St. Louis) are merged into the same cluster, because their Karl-Pearson distance $d = [4.42^{-2}(78.8–78.9)^2 + 1.36^{-2}(3.58–3.63)^2]^{1/2} = 0.043$ is the smallest of the $(28)(28–1)/2 = 378$ distances between the possible pairs. This separation distance can be seen graphically in Figure 16.4: the distance $d = 0.043$ is the height of the leftmost dot in Figure 16.3b. At the second stage Huntsville and Athens are merged, because their Karl-Pearson distance $d = [4.42^{-2}(79.3–79.2)^2 + 1.36^{-2}(5.05–5.18)^2]^{1/2} = 0.098$ is the second-smallest separation of the points (cf. Figure 16.4), and this distance corresponds to the height of the second dot in Figure 16.3b. At the third stage, Worcester and Binghamton ($d = 0.130$) are merged, and at the fourth stage Macon and Augusta ($d = 0.186$) are merged. At the fifth stage, Concordia is merged with the cluster consisting of Springfield and St. Louis. Since the Karl-Pearson distance between Concordia and St. Louis is larger than the distance between Concordia and Springfield (but smaller than the distances between Concordia and the other 25 points), the complete-linkage criterion merges these three points at the larger distance $d = [4.42^{-2}(79.0–78.9)^2 + 1.36^{-2}(3.37–3.63)^2]^{1/2} = 0.193$ (height of the fifth dot in Figure 16.3b).

The heights of the horizontal lines in Figure 16.3a, indicating group mergers, also correspond to the distances between the merged clusters. Since the merger at each stage is between the two closest clusters, these distances become greater at later stages. Figure 16.3b shows these same distances between merged clusters as a function of the stage in the analysis. Subjectively, these distances climb gradually until perhaps stage 22 or stage 23, where the distances between combined clusters begin to become noticeably

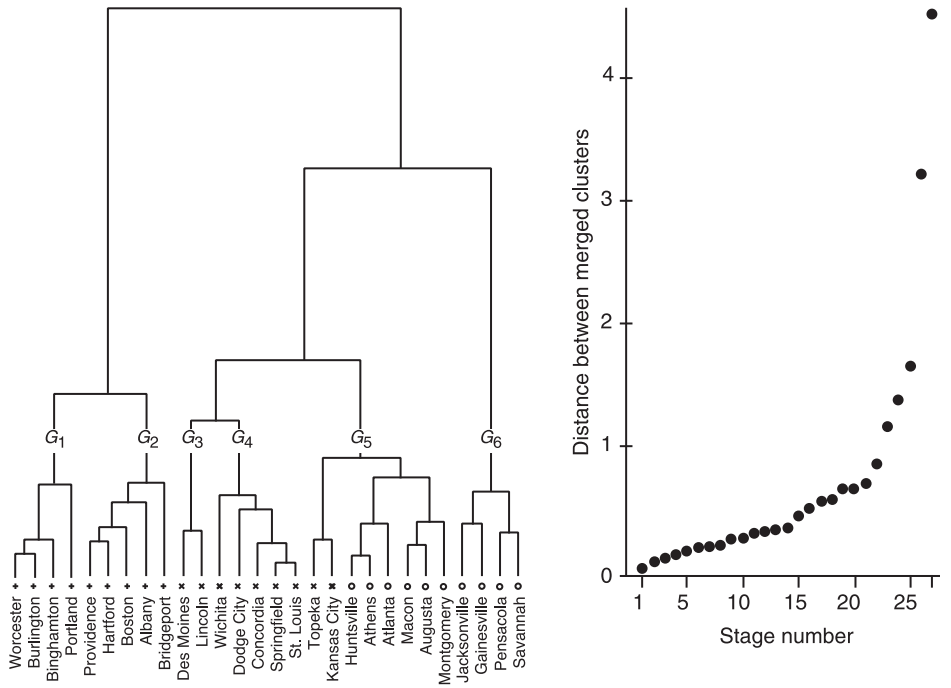

FIGURE 16.3 Dendrogram (a) and the corresponding plot of the distances between merged clusters as a function of the stage of the cluster analysis (b) for the data in Table 15.1. Standardized data (i.e., Karl-Pearson distances) have been clustered according to the complete-linkage criterion. The distances between merged groups appear to increase markedly at stage 22 or 23, indicating that the analysis should stop after 21 or 22 stages, which for these data would yield seven or six clusters, respectively. The six numbered clusters correspond to the grouping of the data shown in Figure 16.4. The seven-cluster solution would split Topeka and Kansas City from the Alabama and Georgia stations in G_5. The five-cluster solution would merge G_3 and G_4.

larger. A reasonable interpretation of this change in slope is that natural clusters have been defined at this point in the analysis, and that the larger distances at later stages indicate mergers of unlike clusters that should be distinct groups. Note, however, that a single change in slope does not occur in every cluster analysis, so that the choice of where to stop group mergers may not always be so clear-cut. It is possible, for example, for there to be two or more relatively flat regions in the plot of distance versus stage, separated by segments of larger slope. Different clustering criteria may also produce different breakpoints. In such cases the choice of where to stop the analysis is more ambiguous.

If Figure 16.3b is interpreted as exhibiting its first major slope increase between stages 22 and 23, a plausible point at which to stop the analysis would be after stage 22. This stopping point would result in the definition of the six clusters labeled $G_1 - G_6$ on the tree diagram in Figure 16.3a. This level of clustering assigns the nine northeastern stations (+ symbols) into two groups, assigns seven of the nine central stations (x symbols) into two groups, allocates the central stations Topeka and Kansas City to Group 5 with six of the southeastern stations (o symbols), and assigns the remaining four southeastern stations to a separate cluster.

Figure 16.4 indicates these six groups in the $K = 2$-dimensional space of the standardized data, by separating points in each cluster with dashed lines. If this solution seemed too highly aggregated on the basis of the prior knowledge and information available to the analyst, the seven-cluster solution produced after stage 21 could be chosen, which separates the central U.S. cities Topeka and Kansas City

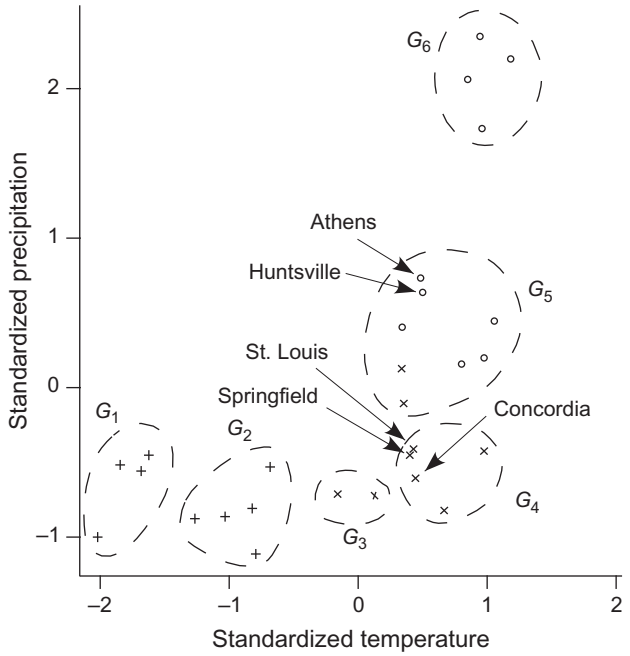

FIGURE 16.4 Scatterplot of the data in Table 15.1 expressed as standardized anomalies, with dashed lines showing the six groups defined in the cluster analysis tree diagram in Figure 16.3a. The five-group clustering would merge the central U.S. stations in Groups 3 and 4. The seven-group clustering would split the two central U.S. stations in Group 5 from six southeastern U.S. stations.

(x's) from the six southeastern cities in Group 5. If the six-cluster solution seemed too finely split, the five-cluster solution produced after stage 23 would merge the central U.S. stations in Groups 3 and 4. None of the groupings indicated in Figure 16.3a corresponds exactly to the group labels in Table 15.1, and we should not necessarily expect them to. It could be that limitations of the complete-linkage clustering algorithm operating on Karl-Pearson distances have produced some misclassifications, or that the group labels in Table 15.1 have been imperfectly defined, or both.

Finally, Figs. 16.5 and 16.6 illustrate the fact that different clustering algorithms will usually yield somewhat different results. Figure 16.5a shows distances at which groups are merged for the data in Table 15.1, according to single linkage operating on Karl-Pearson distances. There is a large jump after stage 21, suggesting a possible natural stopping point with seven groups. These seven groups are indicated in Figure 16.5b, which can be compared with the complete-linkage result in Figure 16.4. The clusters denoted G_2 and G_6 in Figure 16.4 occur also in Figure 16.5b. However, one long and thin group has developed in Figure 16.5b, composed of stations from G_3, G_4, and G_5. This result illustrates the chaining phenomenon to which single-linkage clusters are prone, as additional stations or groups are accumulated that are close to a point at one edge or another of a group, even though the added points may be quite far from other points in the same group.

Figure 16.6 shows the results when the same data are clustered using Ward's method. Here the results are very similar to those in Figure 16.3, which were derived using complete linkage. Ward's method groups the four southernmost stations on the far right of the dendrogram in Figure 16.6a somewhat differently than does complete linkage, and the first relatively large gap after stage 22 in Figure 16.6b suggests that a 5-group solution might be appropriate. ◇

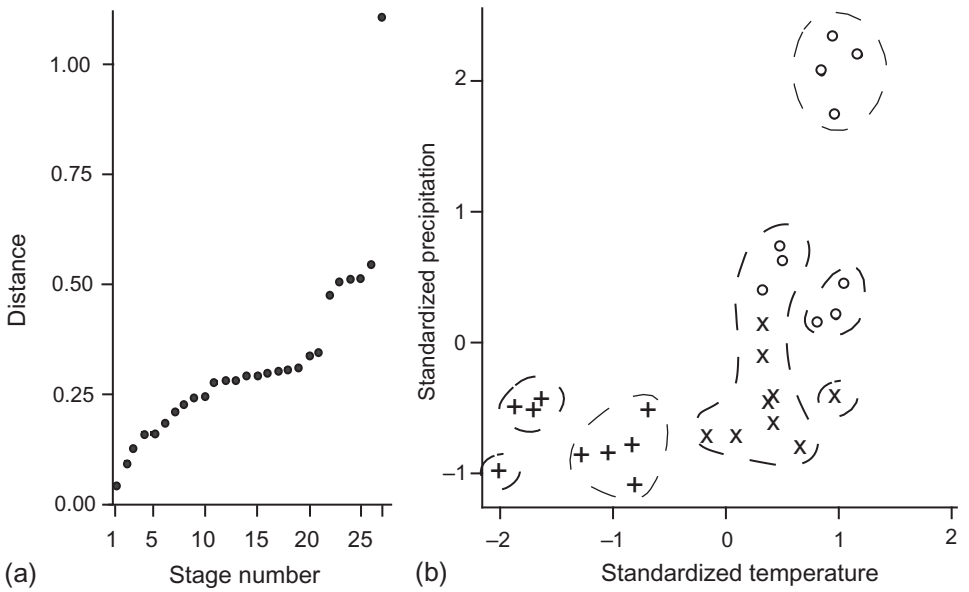

FIGURE 16.5 Clustering of the data in Table 15.1, using single linkage. (a) Merger distances as a function of stage, showing a large jump after 22 stages. (b) The seven clusters existing after stage 22, illustrating the chaining phenomenon.

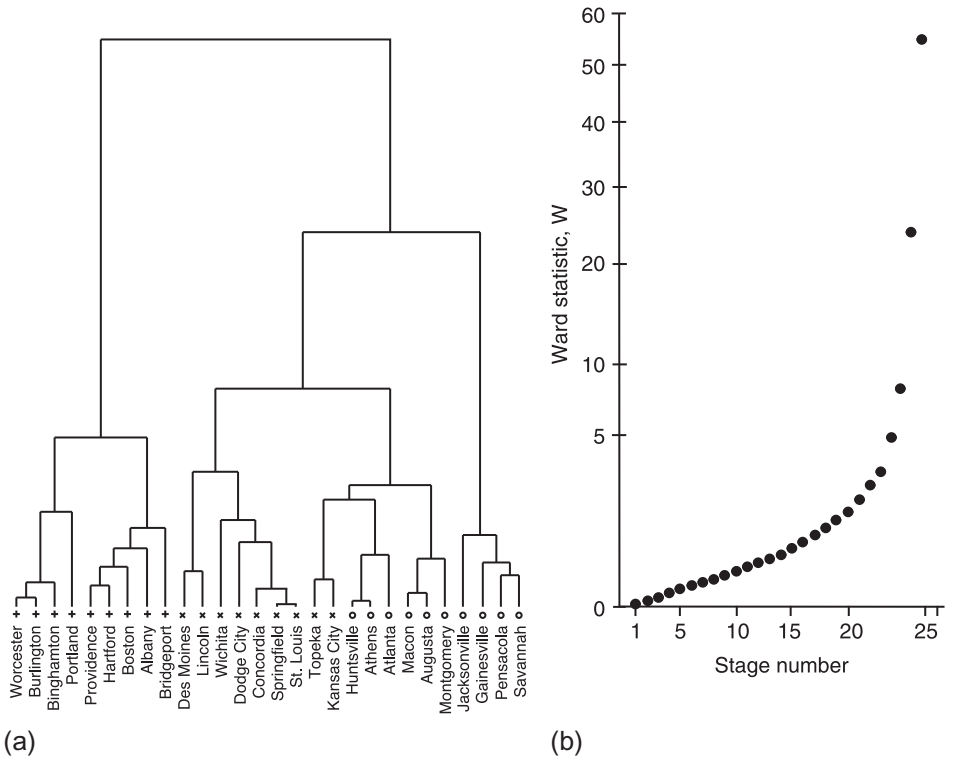

FIGURE 16.6 Clustering of the data in Table 15.1, using Ward's method, in the same format as Figure 16.4. The Ward sum-of-squares statistic W (Equation 16.8) is plotted in the vertical on a square-root scale.

Chapter 8 describes ensemble forecasting, in which the effects of uncertainty about the initial state of the atmosphere on the evolution of a forecast are addressed by calculating multiple dynamical forecasts beginning at an ensemble of similar initial conditions. The method has proved to be an extremely useful advance in forecasting technology, but requires extra effort to absorb the large amount of additional information produced. One way to summarize the information in a large collection of maps from a forecast ensemble is to group them according to a cluster analysis. If the smooth contours on each map have been interpolated from K gridpoint values, then each (Kx1) vector $\boldsymbol{x}$ included in the cluster analysis corresponds to one of the forecast maps.

Figure 16.7 shows the results of the use of Ward's method to group $n = 33$ ensemble members forecasting 500 mb heights over Europe, at a lead time of six days. An innovation in the approach is that it is

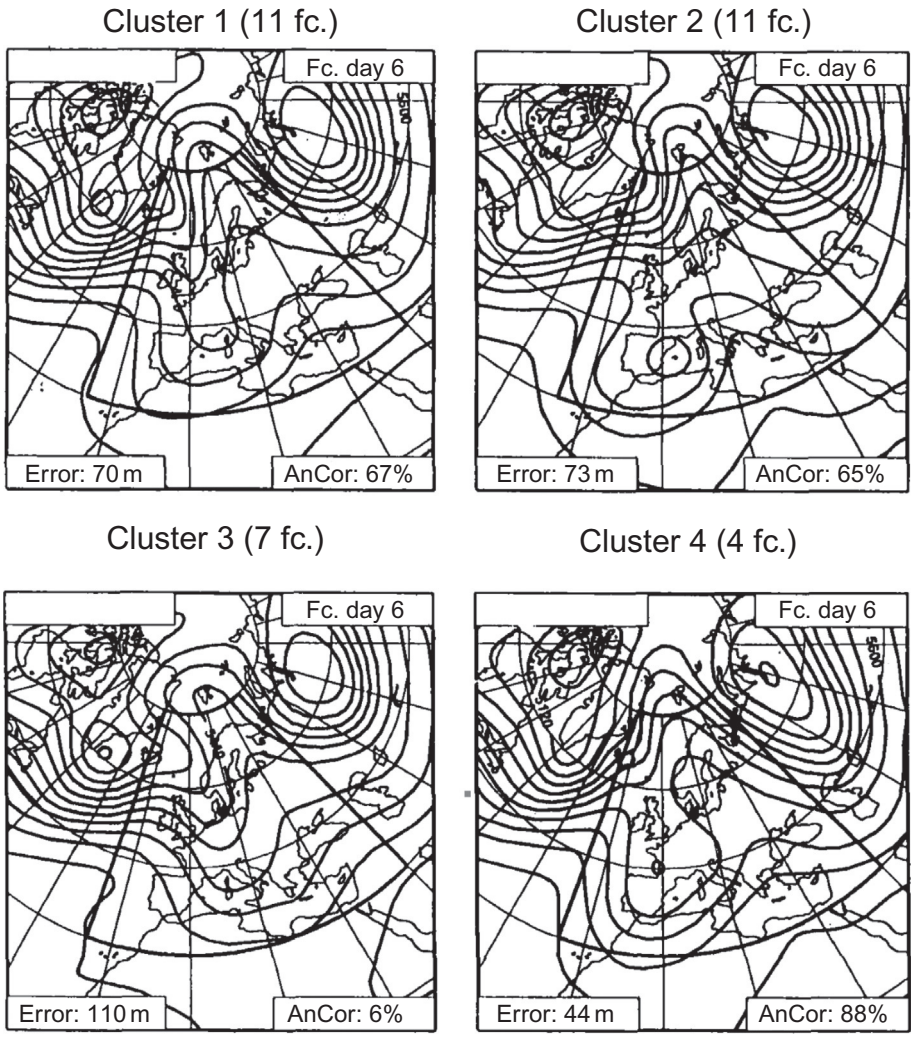

FIGURE 16.7 Centroids (ensemble means) of four clusters for an ensemble forecast for European 500 mb heights at a lead time of six days. Numbers of ensemble members in each cluster are indicated in the captions above each panel. *From Molteni et al. (1996).*

computed in a way that brings out the time trajectories of the forecasts, by simultaneously clustering maps for the five-, six-, and seven-day forecasts, although only the day 6 clusters are shown in Figure 16.7. That is, if each forecast map consists of K gridpoints, the $\boldsymbol{x}$ vectors being clustered would be dimensioned ($3\,K \times 1$), with the first K elements pertaining to day 5, the second K elements pertaining to day 6, and the last K elements pertaining to day 7. Each of the ensemble members remains in the same cluster throughout the forecast period, and clustering their joint behavior over the three-day period allows representation of the coherence in time of the flows that are represented. Because there are a large number of gridpoints underlying each map, the analysis actually was conducted using the first $K = 10$ principal components of the height fields, which was sufficient to capture 80% of the variance, so the clustered vectors had dimension (30×1). The currently operational counterpart of this method is described by Ferranti and Corti (2011).

Another interesting aspect of the example in Figure 16.7 is that the use of Ward's method provided an apparently natural stopping criterion for the clustering, which is related to forecast accuracy. Ward's method (Equation 16.8) is based on the sum of squared differences between the $\boldsymbol{x}$'s being clustered and their respective group means. Regarding the group means as forecasts, these squared differences would be contributions to the overall expected mean squared error if the ensemble members $\boldsymbol{x}$ were different realizations of plausible observed maps. The clustering in Figure 16.7 was stopped at the point where Equation 16.8 yields squared errors comparable to (the typically modest) 500 mb forecast errors obtained at the three-day lead time, so that their medium-range ensemble forecasts were grouped together if their differences were comparable to or smaller than typical short-range forecast errors.

16.2.5. Divisive Methods

In principle, hierarchical clustering can be achieved by reversing the agglomerative clustering process. That is, beginning with a single cluster containing all n observation vectors, this cluster can be split into the two most similar possible groups. At the third stage one of these groups could be split to yield the three most similar groups possible, and so on. The procedure would proceed, in principle, to the point of n clusters each populated by a single data vector, with an appropriate intermediate solution determined by a stopping criterion. This approach to clustering, which is opposite to agglomeration, is called *divisive clustering*.

Divisive clustering is almost never used, because it is computationally impractical for all except the smallest sample sizes. Agglomerative hierarchical clustering requires examination of all $G(G-1)/2$ possible pairs of G groups, in order to choose the most similar two for merger. In contrast, divisive clustering requires examination, for each group of size n_g members, all $2^{n_g-1}-1$ possible ways to make a split. This number of potential splits is 511 for $n_g = 10$, and rises to 524,287 for $n_g = 20$, and 5.4×10^8 for $n_g = 30$. Macnaughton-Smith et al. (1964) propose an alternative approach to divisive clustering that is much faster computationally but explores the possible splits less comprehensively.

16.3. NONHIERARCHICAL CLUSTERING

16.3.1. The *K*-Means Method

A potential drawback of hierarchical clustering methods is that once a data vector $\boldsymbol{x}$ has been assigned to a group it will remain in that group, and in groups with which it is merged. That is, hierarchical methods have no provision for reallocating points that may have been misclassified at an early stage. Clustering

methods that allow reassignment of observations as the analysis proceeds are called *nonhierarchical*. Like hierarchical methods, nonhierarchical clustering algorithms also group observations according to some distance measure in the K-dimensional space of $\boldsymbol{x}$.

The most widely used nonhierarchical clustering approach is called the *K-means* method. The "K" in K-means refers to the number of groups, called G in this text, and not to the dimension of the data vector. The K-means method is named for the number of clusters into which the data will be grouped, because this number must be specified in advance of the analysis, together with an initial guess for the group membership of each of the $\boldsymbol{x}_i$, $i = 1, \ldots, n$.

The K-means algorithm can begin either from a random partition of the n data vectors into the prespecified number G of groups, or from an initial selection of G seed points. The seed points might be defined by a random selection of G of the n data vectors or by some other approach that is unlikely to bias the results. Initial group memberships are then decided according to minimum distances to the seed points. Another possibility is to define the initial groups as the result of a hierarchical clustering that has been stopped at G groups, allowing reclassification of $\boldsymbol{x}$'s from their initial placement by the hierarchical clustering.

Having defined the initial membership of the G groups in some way, the K-means algorithm proceeds as follows:

(1) Compute the centroids (i.e., vector means) $\overline{\boldsymbol{x}}_g$, $g = 1, \ldots, G$; for each cluster.
(2) Calculate the distances between the current data vector $\boldsymbol{x}_i$ and each of the G $\overline{\boldsymbol{x}}_g$'s. Usually Euclidean or Karl-Pearson distances are used, but distance can be defined by any measure that might be appropriate to the particular problem.
(3) If $\boldsymbol{x}_i$ is already a member of the group whose mean is closest, repeat step 2 for $\boldsymbol{x}_{i+1}$ (or for $\boldsymbol{x}_1$, if $i = n$). Otherwise, reassign $\boldsymbol{x}_i$ to the group whose mean is closest, and return to step 1.

The algorithm is iterated until each $\boldsymbol{x}_i$ is closest to its group mean, that is, until a full cycle through all n data vectors produces no reassignments.

The need to prespecify the number of groups and their initial memberships can be a disadvantage of the K-means method, which may or may not compensate its ability to reassign potentially misclassified observations. Unless there is prior knowledge of the correct number of groups, and/or the clustering is a precursor to subsequent analyses requiring a particular number of groups, it is probably wise to repeat K-means clustering for a range of initial group numbers G. Because a particular set of initial guesses for group membership may yield a local rather than global minimum for the sum of distances to group centroids, K-means analyses should be repeated for different initial assignments of observations for each of the trial values of G. Hastie et al. (2009) suggest choosing that G minimizing an overall dissimilarity measure, such as the sum of squared distances between each $\boldsymbol{x}$ and its group mean in Equation 16.8.

16.3.2. Nucleated Agglomerative Clustering

Elements of agglomerative clustering and K-means clustering can be combined in an iterative procedure called *nucleated agglomerative clustering*. This method reduces somewhat the effects of arbitrary initial choices for group seeds in the K-means method and automatically produces a sequence of K-means clusters through a range of group sizes G.

The nucleated agglomerative method begins by specifying a number of groups G_{init} that is larger than the number of groups G_{final} that will exist at the end of the procedure. A K-means clustering into G_{init} groups is calculated, as described in Section 16.3.1. The following steps are then iterated:

(1) The two closest groups are merged according to Ward's method. That is, the two groups are merged that minimize the increase in Equation 16.8.
(2) K-means clustering is performed for the reduced number of groups, using the result of step 1 as the initial point. If the result is G_{final} groups, the algorithm stops. Otherwise, step 1 is repeated.

This algorithm produces a hierarchy of clustering solutions for the range of group sizes $G_{\text{init}} \geq G \geq G_{\text{final}}$, while allowing reassignment of observations to different groups at each stage in the hierarchy.

16.3.3. Clustering Using Mixture Distributions

The fitting of mixture distributions (see Section 4.4.9) (e.g., Everitt and Hand, 1981; McLachlan and Basford, 1988; Titterington et al., 1985) is another approach to nonhierarchical clustering. In the statistical literature, this approach to clustering is called "model-based," referring to the statistical model embodied in the mixture distribution (Banfield and Raftery, 1993; Fraley and Raftery, 2002). For multivariate data the most usual approach is to fit mixtures of multivariate normal distributions, for which maximum likelihood estimation using the EM algorithm (see Section 4.6.3) is straightforward (the algorithm is outlined in Hannachi and O'Neill, 2001, and Smyth et al., 1999). This approach to clustering has been applied to atmospheric data to identify large-scale flow regimes by Haines and Hannachi (1995), Hannachi (1997), and Smyth et al. (1999).

The basic idea in this approach to clustering is that each of the component PDFs $f_g(\boldsymbol{x})$, $g = 1, \ldots, G$, represents one of the G groups from which the data have been drawn. As illustrated in Example 4.14, using the EM algorithm to estimate a mixture distribution produces (in addition to the distribution parameters) posterior probabilities (Equation 4.88) for membership in each of the component PDFs, given each of the observed data values $\boldsymbol{x}_i$. Using these posterior probabilities, a "hard" (i.e., nonprobabilistic) classification can be achieved by assigning each data vector $\boldsymbol{x}_i$ to that PDF $f_g(\boldsymbol{x})$ having the largest probability. However, in many applications retention of these probability estimates regarding the group memberships may be informative.

As is the case for other nonhierarchical clustering approaches, the number of groups G (in this case, the number of component PDFs $f_g(\boldsymbol{x})$) typically is specified in advance. However, the number of mixture densities to include can be cast as a model-selection exercise, using the BIC statistic (Equation 7.39) in connection with appropriate (e.g., multivariate normal) likelihood functions (Fraley and Raftery, 2002). Alternatively, Banfield and Raftery (1993) and Smyth et al. (1999) describe algorithms for choosing the number of groups, using a cross-validation approach.

16.4. SELF-ORGANIZING MAPS (SOM)

The method of Self-Organizing Maps (SOM) is a "machine learning" approach that is commonly used for clustering data sets in which the membership of the training data vectors in some prespecified number of groups G is not known. Accordingly, the method bears similarities to the K-means method described in Section 16.3.1.

The goal in a SOM analysis is to find a 1- or (more usually) 2-dimensional manifold within the K-dimensional data space that best conforms to the data distributions. Initially this manifold is represented by a regular planar grid containing G points, or nodes. As the analysis proceeds this grid is bent out of its initial planar arrangement and deformed away from its initial regular spacing. At the end of the

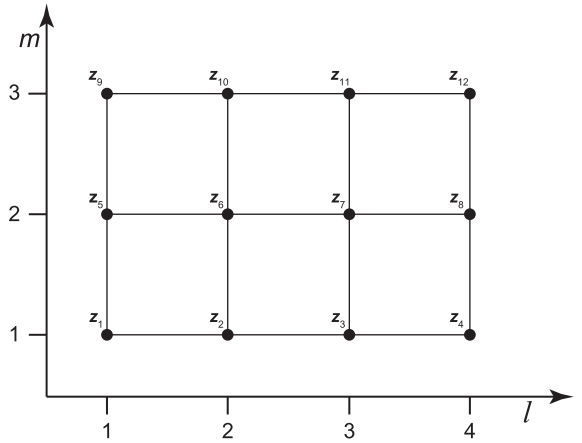

FIGURE 16.8 An example (4×3) initial rectangular grid containing $G = 12$ nodes, z_g.

procedure the data vectors $\boldsymbol{x}$ closest to each of the G nodes are allocated into a group associated with that node.

Figure 16.8 shows an example $(L \times M) = (4 \times 3)$ point grid, of nodes z_g, in which the grid index g might be defined as $g = (m-1)L + \ell$, for $m = 1, \ldots, M$, and $\ell = 1, \ldots, L$, so that $G = LM$. To initialize the calculations, this grid is often placed on the plane defined by the leading two principal components of the data set overall, in the K-dimensional space of the data vectors $\boldsymbol{x}$. The initial grid collapses to a line, and the analysis will utilize a 1-dimensional manifold, for $M = 1$, in which case the initial placement might be along the axis of the leading principal component. In either case, the nodes z_g are $(K \text{x} 1)$ vectors in the K-dimensional data space.

The SOM calculations proceed iteratively through the $\boldsymbol{x}_i$, $i = 1, \ldots, n$, and generally more than one pass is made through the data. At each step the nodes are moved closer to the current data point $\boldsymbol{x}_i$ by first locating the node z_g that is closest (i.e., the node minimizing $||\boldsymbol{x}_i - z_g||$), and then adjusting z_g and all other nodes $z_{g'}$ sufficiently close to $\boldsymbol{x}_i$ toward $\boldsymbol{x}_i$ according to

$$z_{g'} \leftarrow z_{g'} + \alpha\, h\left(\left\| \begin{bmatrix} \ell \\ m \end{bmatrix} - \begin{bmatrix} \ell' \\ m' \end{bmatrix} \right\|\right) (\boldsymbol{x}_i - z_g), \tag{16.9}$$

where α is a learning rate parameter, $h(\bullet)$ is a kernel function, the indices ℓ and m pertain to the closest node z_g, and the indices ℓ' and m' pertain to the node $z_{g'}$ being moved toward $\boldsymbol{x}_i$. Notice that the proximity of a node $z_{g'}$ to the closest node z_g is defined in the argument of the kernel function with respect to distances between them in the original grid system, so that for example the distance between z_2 and z_7 in Figure 16.8 is always $\sqrt{2}$, regardless of how extremely the initial rectangular arrangement of nodes has been distorted by the iterative adjustments. Typical choices for the kernel function are circular, in which case only nodes $z_{g'}$ nearest to the closest node z_g are adjusted, and those adjustments are of equal magnitudes; or Gaussian, in which case all $z_{g'}$ are adjusted toward the current $\boldsymbol{x}_i$ but closer nodes are displaced more. In either case $z_{g'} = z_g$ in Equation 16.9 for the closest node, so that z_g is always adjusted toward the current data point $\boldsymbol{x}_i$. As the iterations proceed, the learning rate parameter α is reduced from 1 to 0, and the width or standard deviation of the kernel function decreases, so that the adjustments toward each $\boldsymbol{x}_i$ tend to be smaller as the calculations proceed. These adjustments

can be made either gradually and continuously as the iterations progress, or as step functions that change after each set of n iterations.

Figure 16.9 shows an example $G = 6$ result, obtained using an SOM based on a (2×3) grid. Here 104 locations in the northeastern United States have been clustered based on data vectors $\boldsymbol{x}_i$, $i = 1, \ldots, 104$, containing the pairwise station covariances of daily precipitation data for the years 1900–1999. That is, each $\boldsymbol{x}_i$ is a row or column of the full (104×104) covariance matrix. The analysis has succeeded in regionalizing the data into six overlapping groups, which turn out to be spatially contiguous. The upper pair of diagrams in Figure 16.10 shows the arrangement of the 6-node grid, and the correspondence between nodes and the regions labeled A–F in Figure 16.9. The arrangement of the regions in the space of the grid mirrors their relative geographical positions, as a consequence of covariances between stations tending to decrease as the station separations increase. The lower pair of diagrams in Figure 16.10 shows the corresponding results from a $G = 12$-group clustering based on a (3×4) grid. The hatching patterns correspond to

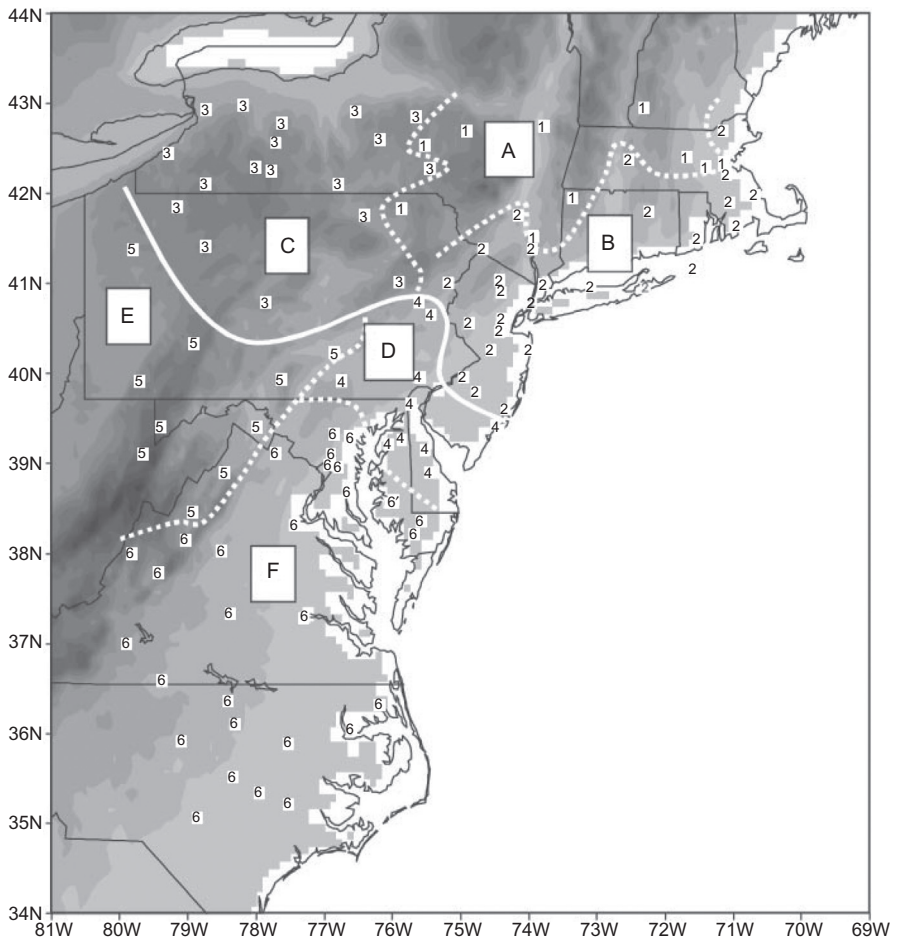

FIGURE 16.9 Clustering of 104 locations in the northeastern United States into $G = 6$ groups, according to daily precipitation data, 1900–1999, using a SOM based on a (2×3) grid. Shading indicates the topography. *From Crane and Hewitson (2003).*

the groups indicated in the upper diagram and in Figure 16.9, showing that the groups maintain their relative positions in the two grid systems, but their overlap in the lower-right diagram illustrates that the SOM clustering is nonhierarchical.

Although Figure 16.9 shows a geographic regionalization based on an SOM, the "map" in the name SOM usually refers to a display of representative data vectors which are those closest to each node z_g, on the original and undistorted grid. Figure 16.11 shows one such map, depicting fields of scatterometer-

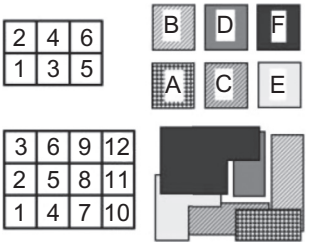

FIGURE 16.10 The relationship between the $G = 6$-group regionalization in Figure 16.9 and the (2×3) SOM grid is indicated in the upper pair of panels. The lower pair of panels shows corresponding results for the $G = 12$-group SOM, where the overlapping shading illustrates that the SOM clustering is nonhierarchical. *From Crane and Hewitson (2003).*

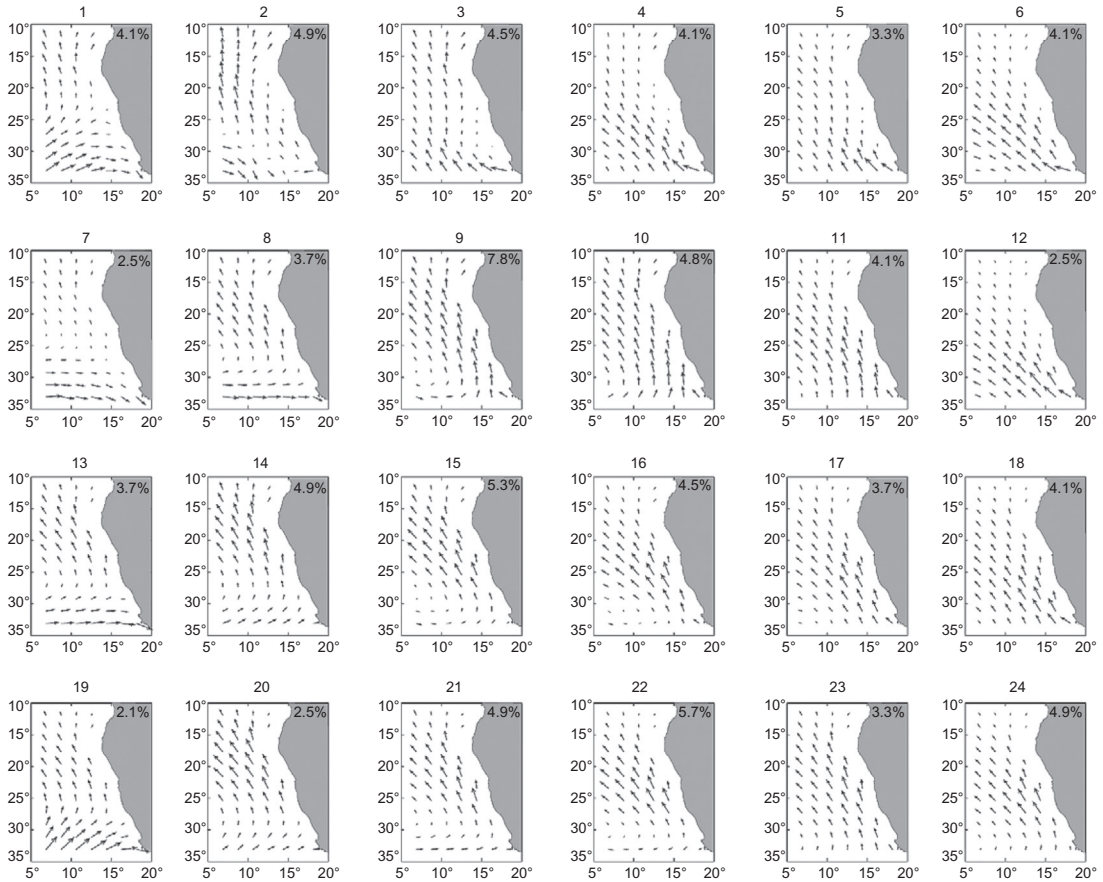

FIGURE 16.11 A SOM depicting a $G = 24$-group clustering of satellite-derived sea-surface winds in a portion of the southeastern Atlantic Ocean. The edge of the gray shading locates the west coast of southern Africa. *From Richardson et al. (2003).*

derived surface wind vectors for a region of the southeast Atlantic Ocean. Here the SOM has been computed using a (6×4) grid, and the proportion of the underlying (u, v) wind pair fields classified into each node are indicated in the upper-right corners. A close study of the individual panels reveals that nearby nodes are very similar to each other, reflecting the fact that the number density of SOM nodes in the K-dimensional data space will tend to follow the underlying density of data vectors. On the other hand, the strong similarities among nearby panels in Figure 16.11 suggest that a smaller initial grid might yield a more parsimonious clustering of the underlying data.

16.5. EXERCISES

16.1. Compute the distance matrix $[\Delta]$ for the Guayaquil temperature and pressure data in Table A.3 for the six years 1965–1970, using Karl-Pearson distance.

16.2. From the distance matrix computed in Exercise 16.1, cluster the six years using
   a. Single linkage.
   b. Complete linkage.
   c. Average linkage.

16.3. Cluster the Guayaquil pressure data (Table A.3) for the six years 1965–1970, using
   a. The centroid method and Euclidean distance.
   b. Ward's method operating on the raw data.

16.4. Cluster the Guayaquil temperature data (Table A.3) for the six years 1965–1970 into two groups using the K-means method, beginning with $G_1 = \{1965, 1966, 1967\}$ and $G_2 = \{1968, 1969, 1970\}$.

16.5. Consider the mixture distribution $f(x) = w\, f_1(x) + (1\text{–}w)\, f_2(x)$, where x is a nonnegative scalar, and each of the two constituent PDFs is a univariate exponential distribution. Applying the E-M algorithm to a collection of rainfall data, you have estimated the parameters as $\mu_1 = 2\,\text{mm}$, $\mu_2 = 20\,\text{mm}$, and $w = 0.7$, for the purpose of "model-based" clustering.
   a. Make a "hard" classification of a new datum, $x_0 = 5\,\text{mm}$, into one of the two clusters.
   b. Calculate probabilities of membership for $x_0 = 5\,\text{mm}$ in each of the two clusters.