Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Chapter 8 Offprint

Ward Clustering and *k*-means Clustering

First published: December 2013
ISBN: 978-84-92937-50-9

Supporting websites:
www.fbbva.es
www.multivariatestatistics.org

Fundación BBVA

CHAPTER 8

Ward Clustering and *k*-means Clustering

This chapter continues the theme of cluster analysis, first with a popular alternative to the hierarchical clustering methods presented in Chapter 7 – Ward clustering. This method is based on the same concepts as analysis of variance (ANOVA), so we shall give a brief introduction to ANOVA, specifically to the definitions of between-group and within-group variance, to motivate Ward clustering. Exactly the same concepts are used in a different clustering algorithm called *k-means clustering*. This form of clustering, which is an example of "nonhierarchical" clustering, is particularly useful when a very large number of objects need to be clustered, where the dendrogram would be so big that it becomes too burdensome to visualize and interpret. In this situation, all we really want is a partitioning of the objects into a set of groups. Nonhierarchical clustering algorithms such as *k*-means do not result in a dendrogram – the user specifies in advance how many groups are being sought (the *k* of *k*-means) and the final result is the allocation of each object to a group so that the groups are as internally homogeneous as possible. This measure of internal homogeneity is the same as in Ward clustering, hence our treatment of these two methods together in this chapter.

Contents

Analysis of variance (ANOVA)

To introduce the concepts inherent in Ward and nonhierarchical clustering, it is worthwhile to recall analysis of variance, abbreviated as ANOVA. ANOVA is concerned with testing the difference between means of a continuous variable observed in different groups. As an example, we can use

the continuous variable "pollution" and the categorical variable "sediment" from Exhibit 1.1, where sediment divides the sample into three groups: clay (11 sites), sand (11 sites) and gravel (8 sites). In the middle of Exhibit 1.5 the box-and-whisker plot for pollution shows the medians of the three groups, or subsamples, and their dispersions between first and third quartiles – here we shall be concerned with the means and variances in each subsample. Exhibit 8.1 shows an alternative graphical display of the same data, where each value is shown on the pollution scale and is coded according to its respective sedimentary group. The means of each group are indicated as well as the mean pollution of all 30 sites. In ANOVA, the separateness of the three groups is measured by how far the means are away from the overall mean, taking into account the size of the groups, the so-called *between-group sum of squares* BSS:

$$\mathrm{BSS} = \sum_{g=1}^{G} n_g (\bar{x}_g - \bar{x})^2 \tag{8.1}$$

where G = number of groups, n_g = the sample size in the g-th group, $\bar{x}_g$ is the g-th group mean and $\bar{x}$ is the overall mean. In this particular case the calculation gives a value of BSS = 37.6. In isolation this value tells nothing about how separate the groups are, because if the three groups of points were more tightly dispersed about their respective means, we would get the same value of BSS even though the groups appear more separate. The dispersion of the observations around their respective group means thus needs to be taken into account, and this is calculated by the *within-group sum of squares* WSS:

$$\mathrm{WSS} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{ig} - \bar{x}_g)^2 \tag{8.2}$$

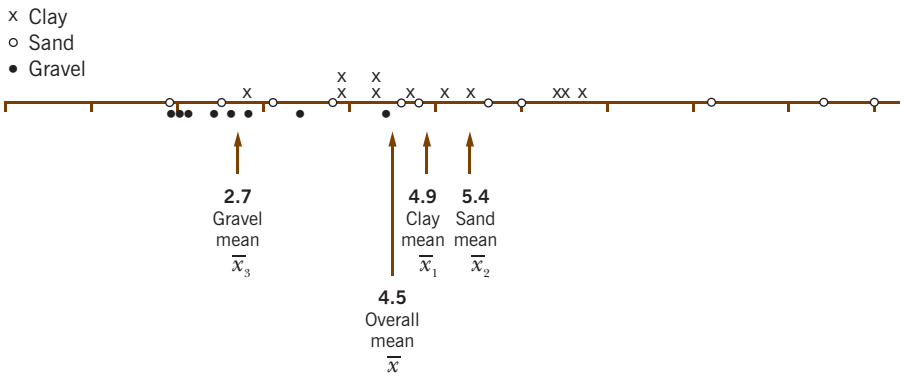


Exhibit 8.1: *Representation of the 30 values of pollution (see Exhibit 1.1), coded for the three sediment types. The means (to one decimal place) of the three subsets of data are indicated, as well as the overall mean (compare this graphical representation with that of the middle plot of Exhibit 1.5, where the medians and quartiles are displayed)*

which in this example is equal to WSS = 95.4. The beauty of using sums of squares is that BSS and WSS add up to the *total sum of squares*, TSS:

$$\text{TSS} = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{ig} - \bar{x})^2 \tag{8.3}$$

that is,

$$\text{BSS} + \text{WSS} = \text{TSS} \tag{8.4}$$

in this case: 37.6 + 95.4 = 133.0. TSS in (8.3) is the sum of squared deviations of all the observations from the overall mean, which measures the dispersion of all the data points around the overall mean. WSS in (8.2), on the other hand, is measuring the dispersion of the observations in the groups around their own respective group means, so WSS must be less than TSS. Notice that TSS divided by $n - 1$ (= 29) is the usual sample variance (of all the observations), while each of the G summations in (8.2), divided by its respective $n_g - 1$, is the variance of the g-th group. Furthermore, (8.1) divided by $G - 1$ is the variance of the $G = 3$ means weighted by their respective group sizes. The term *analysis of variance* derives from the fact that (8.4) implies this decomposition of variance into parts between and within the groups. In order to test whether there is significant separation of the groups, or whether the observed separation is compatible with random variation in the data, the values of BSS and WSS are combined into the classic F-statistic.[1] This *F-test* gives a p-value of 0.0112, indicating significant differences between the sediment groups in terms of pollution. We could also perform a permutation test, to be described in Chapter 17, which estimates the p-value as 0.0106, very close to that of the F-test.

Looking for the optimal solution

In ANOVA the grouping variable is prescribed (sediment type in the above example), but in cluster analysis we are looking for a grouping variable in the data. In the one-dimensional example of Exhibit 8.1, suppose we have no classification of the 30 values, what would be the optimal clustering of the data into three groups? Optimality could be defined as maximizing the ratio BSS/TSS, which is equivalent to optimizing any increasing function of that ratio, for example BSS itself (since TSS is fixed), or BSS/WSS, or the F-statistic defined in the footnote. Because there is only one variable and a fairly small sample size, we can investigate every pair of cutpoints that separates the data set into three groups (clearly,

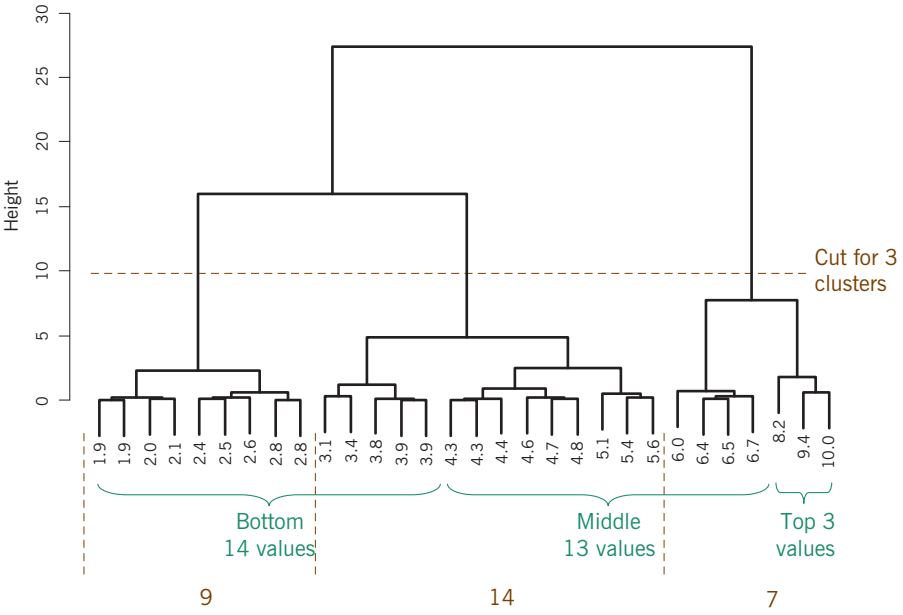
[1] The classical test in ANOVA for testing differences between means is the F-test, where $F = \frac{BSS/(G-1)}{WSS/(n-G)}$ has the F-distribution with $G - 1$ and $n - G$ "degrees of freedom". The observed value $F = (37.6/2)/(95.4/27) = 5.32$ has an associated p-value of 0.0112, which is very close to the p-value of 0.0106 of the permutation test.

for maximal separation of the three groups, each group is defined as a contiguous interval on the pollution scale). There are $29 \times 28/2 = 406$ pairs of cutpoints, and the maximum BSS/TSS turns out to be 0.867 obtained when the three groups are defined as (i) the first 14 values; (ii) the next 13 values; (iii) the last 3 values, as shown in green in the lower part of Exhibit 8.2. This exhaustive way of looking for a given number of clusters is actually the nonhierarchical clustering treated in a later section, but we do it here to contrast with Ward clustering, which is the hierarchical version of this search.

Ward clustering in one dimension

Ward clustering also tries to maximize BSS (or, equivalently, minimize WSS) but does it at each step of a hierarchical clustering like the ones described in Chapter 7. So we start with single objects and look for the pair that are the "closest", in terms of keeping the WSS as small as possible (and thus the BSS as large as possible), and proceed stepwise in this way until a dendrogram is constructed. Exhibit 8.2 shows the dendrogram constructed by Ward clustering, and the associated three-cluster solution using a cutting of the tree at about level 10, giving clusters of 9, 14 and 7 sites. Notice that the Ward procedure does not necessarily find the optimal solution – this is because the hierarchical clustering is stepwise and every merging of clusters depends on what has happened previously. For example, the values 6.0; 6.4; 6.5 and 6.7 join the smaller cluster on the right formed by the three top values 8.2; 9.4 and 10.0, whereas in the optimal solution these three top values form a

Exhibit 8.2:
Ward clustering of the 30 sites in Exhibit 1.1 according to the single variable "pollution", showing the cutpoint for a 3-cluster solution (partitioning of 9; 14 and 7 values, shown by vertical dashed lines), with between-to-total sum of squares ratio, BSS/TSS = 0.825. The sites are labelled by their pollution values. The curly brackets show the globally optimal 3-cluster solution (partitioning of 14; 13 and 3 values) for which BSS/TSS = 0.867



cluster alone. The Ward clustering solution, with BSS/TSS = 0.825, is actually quite far from the optimal partitioning of the 30 sites, where BSS/TSS = 0.867, computed above.

Ward clustering in several dimensions

The type of exhaustive search that we could do above in one dimension, looking at all possible cutpoints, becomes much more difficult when the data are multidimensional: for example, for the three-dimensional (continuous) environmental data of Exhibit 1.1, we would have to consider all pairs of planes dividing the sample into three subsamples. Hierarchical Ward clustering, however, is still very simple to execute, even though it is unlikely to find the optimal solution. The algorithm proceeds in the same way as for the unidimensional case, with the BSS and TSS measures using squared distances in multidimensional space, which are the natural generalizations of the squared differences in one dimension. For example, BSS in (8.1) becomes, in the multidimensional version:

$$\mathrm{BSS} = \sum_{g=1}^{G} n_g d(\bar{\mathbf{x}}_g, \bar{\mathbf{x}})^2 \qquad (8.5)$$

where $\bar{\mathbf{x}}_g$ and $\bar{\mathbf{x}}$ are now the g-th mean vector and overall mean vector, respectively. When there are more than one variable, then the issue of standardization becomes important when defining the distance, as explained in Chapter 4. Exhibit 8.3 shows the Ward clustering of the 30 samples based on Euclidean distance using the three standardized variables (depth, pollution and temperature) – part of the distance matrix has been given in Exhibit 4.5.

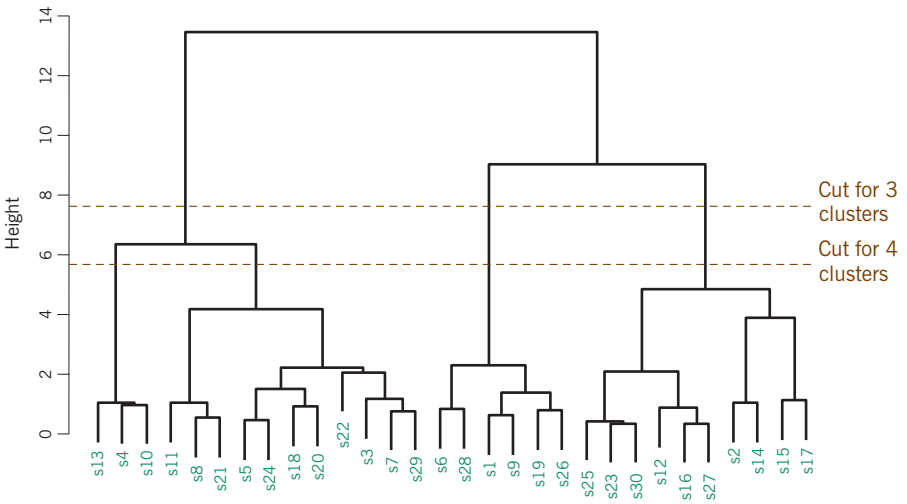


Exhibit 8.3: *Ward clustering of the 30 sites in Exhibit 1.1 according to the three variables depth, pollution and temperature, using standardized Euclidean distances (Exhibit 4.5). Cuts are shown which give three and four clusters*

Notice how different this result is in appearance from the complete linkage clustering of Exhibit 7.10, although on the left in both dendrograms we can see that the cluster of three sites (s13,s4,s10) corresponding to the three highest pollution values remains a separate cluster in both analyses. In the next section we shall show that the cluster solutions are, in fact, very similar. We can again perform a permutation test for clusteredness, to be described more fully in Chapter 17. For example, for the four-cluster solution, the permutation test estimates a p-value of 0.465, so there is no evidence of clustering in these data, and the analysis simply serves to partition the sites into four groups in their three-dimensional spatial continuum.

Comparing cluster solutions

One cluster analysis, which yields p clusters, can be compared to another cluster analysis on the same data, giving q clusters, by cross-tabulating the categories from the two solutions. For example, let us compare the four-category complete linkage solution from Exhibit 7.10 ($p = 4$) with the four-cluster Ward solution from Exhibit 8.3 ($q = 4$), leading to the following cross-tabulation of the 30 sites:

		Ward clustering			
		Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*
Complete linkage clustering	*Cluster 1*	2	0	11	0
	Cluster 2	0	10	0	0
	Cluster 3	0	0	0	3
	Cluster 4	4	0	0	0

Apart from the two sites in the first cell of the table, all the sites fall into the same clusters in both solutions – the two solutions would agree perfectly if these two (identified as sites s1 and s9) were either in cluster 4 of the complete linkage solution, or in cluster 3 of the Ward solution. There are several ways to measure the agreement between the two solutions, as summarized in such a cross-tabulation; for example, Cramer's V statistic given in (6.3) – which is equal to 1 for perfect agreement, is equal to 0.925 in this example.

Nonhierarchical clustering by k-means

Instead of constructing a dendrogram, nonhierarchical clustering searches for a prescribed number of clusters in the data. We shall describe the most popular of the nonhierarchical algorithms, called *k-means clustering*. The k refers to the specified number of groups we are looking for in the data set, and *means* refers to the fact that in each iteration of the algorithm objects are allocated to the closest group mean. The k-means algorithm proceeds as follows, where n objects need to be clustered into k groups, and we have a distance function between any pair of

objects (which should be of the Euclidean type, weighted or unweighted, for the decomposition (8.4) to be valid):

1. Choose k objects at random as starting *seeds*; or use k prespecified objects as seeds.

2. Calculate the distances of all n objects to the k seeds, and allocate each object to its closest seed – this gives a first clustering of the objects into k groups.

3. Calculate the means of the k clusters, used as seeds for the next iteration.

4. Repeat steps 2. and 3. until convergence, that is when there is no change in the group allocation from one iteration to the next.

It can be proved that when the distance function is of the Euclidean type, optionally weighted, then the value of the between-groups sum of squares (BSS) must increase from one iteration to the next. Since BSS cannot be higher than TSS, the algorithm must converge, but there is no guarantee that the convergence is at the global optimum – we say that the algorithm converges at a local optimum, which could be the global one, we simply do not know.

The *k*-means algorithm is very fast and is generally used for very large data sets, typically found in the social sciences, for example looking for clusters of attitudes in a political survey of thousands of people. If a large ecological data set is encountered, this is a way to find some simple structure in the sample units. In the clustering of the 30 sites described previously, in terms of the single variable pollution, we did know the global optimum (BSS/TSS = 0.867, which is the highest possible value for this example – see Exhibit 8.2) because we could do an exhaustive search of all three-cluster solutions. We already saw that Ward clustering gave a nonoptimal solution, with BSS/TSS = 0.825. Even though *k*-means is usually used for much bigger data sets and many variables, we applied it to the same example, specifying three clusters as before. The result was BSS/TSS = 0.845, which is an improvement over the hierarchical clustering solution, but still not the global optimum. In *k*-means clustering the starting set of seeds is quite crucial to the result – unless we have some prior information on what constitutes good seeds to start growing clusters, the initial seeds are chosen randomly. So it is recommended to use several random sets of starting seeds and then take the best result. When we repeated *k*-means clustering using 10 different random starts, we did indeed find the optimal solution with BSS/TSS = 0.867.

Similarly, we can compare the *k*-means result, after several random starts, with the Ward clustering on the three-dimensional data. The four-cluster solution

obtained in the latter result, shown in Exhibit 8.3, gives a BSS/TSS ratio of 0.637. The best *k*-means solution, after 10 random starts, has an improved BSS/TSS equal to 0.648. So it seems, just after these few examples, that if one is interested in obtaining only a partition of the objects, then *k*-means clustering with several random starts does perform better than hierarchical Ward clustering. It does no harm, however, to do both and check which gives the better solution.

Weighting the objects in Ward and *k*-means clustering

The central concepts of Ward and *k*-means clustering are the measures BSS, WSS and TSS, where the objective is to maximize BSS, or equivalently minimize WSS because they add up to TSS, which is a constant. In the definitions (8.1), (8.2) and (8.3) of these measures, each object is counted, or weighted, equally. But in some situations (we shall see one of these when we treat correspondence analysis) we would like to count some objects differently from others, that is weight the objects differentially. If $w_1, w_2, \ldots, w_n$ denote positive weights assigned to the n objects, then (8.1)–(8.3) can be generalized as:

$$\mathrm{BSS} = \sum_{g=1}^{G} w_g (\bar{x}_g - \bar{x})^2 \tag{8.6}$$

where w_g is the total weight of the objects in the g-th group: $w_g = \sum_{i=1}^{n_g} w_i$.

$$\mathrm{WSS} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} w_i (x_{ig} - \bar{x}_g)^2 \tag{8.7}$$

$$\mathrm{TSS} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} w_i (x_{ig} - \bar{x})^2 \tag{8.8}$$

The equally weighted versions used before are thus a simple case when $w_i = 1$. The multidimensional equivalents – for example, BSS in (8.5) – are generalized in a similar fashion.

SUMMARY: Ward clustering and *k*-means clustering

1. Ward clustering is a hierarchical cluster analysis where the criterion for merging two clusters at each node of the tree is to maximize the separation of the new cluster's mean from the means of the other clusters. The separation between clusters is measured by the between-group sum of squares (BSS).
2. Equivalently, the criterion is based on minimizing the dispersion within the newly combined cluster. The dispersion within clusters is measured by the within-group sum of squares (WSS).
3. BSS and WSS sum to a constant, the total sum of squares (TSS). Thus, maximization of BSS is equivalent to minimization of WSS.

4. *k*-means clustering is a nonhierarchical cluster analysis based on exactly the same criteria as Ward clustering, with the difference that a solution is sought by an iterative procedure which successively allocates the set of observations to a set of *k* seeds, where *k* is the number of clusters specified by the user.

5. The initial seeds are *k* observations randomly chosen, or specified by the user, from which the algorithm can start to allocate observations to their nearest seed, providing a first clustering of the observations. Mean points in each cluster are calculated, which provide the seeds for the next iteration, and this process is repeated until there is no change in the clustering from one iteration to the next.

6. If interest is just in finding a set of clusters rather than visualizing the complete clustering process, then *k*-means clustering seems to find better solutions, but the analysis should be repeated several times with different random sets of initial seeds.

7. Both Ward clustering and *k*-means can be generalized to include observation weights, which give observations varying importance in the cluster analysis.

List of Exhibits