Note on Cluster Analysis Results

***Multivariate Data Analysis*, Sixth edition**

"The world's leading authority on applied multivariate data analysis based on number of citations, as reported by Google Scholar"

As noted in the text, cluster analysis is a non-statistical procedure, using specified algorithms to calculate similarity and form clusters through either hierarchical or nonhierarchical procedures. As a result, differing solutions can result from such actions as random versus specified seed points in nonhierarchical procedures or the type of partitioning procedure used in the nonhierarchical analysis.

One other characteristic of the data that can potentially impact certain results is the order of cases in the dataset. As we will discuss below, SPSS and other programs can be impacted by the order of cases in both hierarchical and nonhierarchical procedures.

The user should note that this characteristic exemplifies the differing results that can be obtained from cluster analysis due to any number of factors. As stressed in the text's discussion of validating the results, the researcher is encouraged to vary as many of these characteristics as possible in differing analyses to assess the changes that occur between cluster solutions.

Impacts on Results in Chapter 8

In the HBAT analysis in Chapter 8, the dataset was resorted from the original order in some exploratory analysis before both the hierarchical and nonhierarchical procedures. While the exploratory analyses (and the data resorting) themselves do not impact the data in any way, the characteristics of cluster analysis discussed above require that this resorting must be done to exactly replicate the analysis shown in the text. The implications of not resorting the dataset before running the analyses are discussed below, ranging from impacting just the labeling of individual cases to an actual slight change in cluster membership.

Hierarchical Procedure:

> In SPSS and some other procedures, labeling for the output is done based on case order (i.e., position in the dataset) rather than a specific variable label (e.g., questionnaire ID). As such, if the dataset is in a different order, the results will be the same (i.e., agglomeration coefficient and cluster membership), but the labeling on the output and diagrams (e.g., dendrogram) may differ. This is because the same case will have a different case order. <u>Thus, while the results are equivalent, certain aspects of the output (i.e., labeling of cases) will differ.</u>

Nonhierarchical Procedure:

Nonhierarchical procedures follow a partitioning procedure to assign individual cases to clusters as discussed in the text. While the text discusses that differing seed points can result in slightly differing solutions, it should also be noted that the actual order of the dataset can also impact the results, even with specified seed points. This results from the initial partitioning and updating procedures. Thus, to exactly replicate the nonhierarchical results illustrated in the text, the dataset must be resorted for the nonhierarchical analysis as well.

Resolution

The most direct solution is to resort the data to the order used when performing the analysis in the text. Thus, the command syntax available for download from the textbook website (www.mvstats.com) has this already included. If you are not using this syntax, the order of cases used for each procedure is listed below.

Dataset order for hierarchical analysis (listed by questionnaire number (ID) from first to last in the dataset):

1 6 8 30 33 34 39 40 41 42 49 50 55 63 68 69 70 75 86 95 97 100 2 3 16 23 27 29 45 47 52 53 58 60 61 79 85 88 94 4 5 11 12 14 17 21 25 26 28 31 35 37 38 51 54 56 62 65 67 74 77 78 80 82 89 91 93 96 7 87 92 9 10 15 18 19 46 48 57 66 73 13 20 22 43 71 99 24 32 36 59 64 72 76 81 83 98 44 90 84

Dataset order for nonhierarchical analysis (listed by questionnaire number (ID) from first to last in the dataset). All 100 observations are included in reordering the dataset even though seven cases are omitted from the actual nonhierarchical analysis.

37 86 75 95 55 83 97 8 99 93 54 89 11 5 10 70 82 51 65 56 25 28 68 6 18 32 50 9 27 14 17 39 88 15 42 34 69 85 19 23 46 20 91 12 80 58 26 31 66 64 61 53 38 29 79 1 73 67 30 21 78 71 16 33 40 63 77 94 3 98 96 41 4 72 45 60 43 57 100 36 49 48 52 81 76 74 2 90 7 87 62 35 92 47 13 44 22 59 84 24 .