

Jason & Glenwick, eds.

Methodological Approaches to Community-Based Research

Chapter 5/Web Appendix guide

Web Appendix Guide for Chapter 5/Cluster Analysis

The Web Appendix for Chapter 5 (on cluster analysis) includes a data set and corresponding SPSS syntax. These materials provide the information necessary to run the analyses presented in the chapter. More specific directions are embedded within the files.

Datasets

The complete data set is ready for analysis in SPSS format and is also readable by SAS and Excel. As described in the text of the chapter, the first step to conducting cluster analysis is data preparation. We have done this step for you in this practice data by using theoretically grounded and nonredundant measures (e.g., using four indicators to create a neighborhood poverty scale); examining 3-D scatterplots, histograms, and dendrograms to determine if there is clustered data (which there is); and using three scales with the same range. We remind readers of the importance of this step when conducting cluster analysis.

SPSS Syntax

All SPSS syntax is given in a SPSS syntax file.

The second step is to select a clustering method that is appropriate for the research question and the data. In the chapter, we conducted hierarchical clustering using Ward's Linkage as the way of conceptualizing the locations of clusters and Squared Euclidean Distance as a way of measuring distances between cases and clusters. There are a number of other linkage methods (Nearest Neighbor Linkage, Furthest Neighbor Linkage, Between-Groups Linkage) and other distance measures (Pearson Correlation, Cosine, Block) that could also be selected in SPSS. The difference between Single Neighbor and Ward's method is the point in the cluster used to represent the cluster in measuring distances. As its name suggests, Nearest Neighbor Linkage measures the distances between clusters by the distance between the closest points. Ward's method uses an analysis of variance approach to evaluate the distances between clusters and tries to minimize the sum of squares of any two (hypothetical) clusters that can be formed at each step. It is regarded as an efficient approach and is similar to the approach used in K-means clustering. Euclidean and Squared Euclidean Distances are the most commonly used distance measures. The former is the geometric distance in the multidimensional space. The latter places progressively greater weight on objects that are further apart.

Syntax to Conduct Hierarchical Clustering using Ward's Linkage and Squared Euclidean Distance:

```
CLUSTER  ConPov EtHet BusDis
/METHOD WARD
/MEASURE=SEUCLID
/ID=CommId
/PRINT SCHEDULE
/PLOT DENDROGRAM
/SAVE CLUSTER(3).
```

variable labels clu3_1 'Hierarchical 3 cluster solution' .
execute.

- Use this syntax to do hierarchical clustering using Ward's Linkage and Squared Euclidean Distance. You can compare different linkage methods by replacing WARD (Ward's method) with COMPLETE (Furthest Neighbor Linkage) or BAVERAGE (between groups linkage). Replacing SEUCLID (Squared Euclidean Distance) with EUCLID (Euclidean distance), CORRELATION (Pearson correlation) or COSINE (Cosine) will compare different distance metrics. The syntax provided automatically saves identifiers from a three-cluster solution but it could be changed to select a different solution: For example, the syntax /SAVE CLUSTER (3) could be changed to /SAVE CLUSTER (2) to select a two-cluster solution. We chose a three-cluster solution by looking at the dendrogram (see Figure 5.1 in Chapter 5) and 3-D scatterplots (see Online Figure 5.1 in this Web Appendix).

Syntax to Conduct Non-Hierarchical K-means Clustering with Three Clusters:

```
QUICK CLUSTER ConPov EtHet BusDis
/MISSING=LISTWISE
/CRITERIA=CLUSTER(3) MXITER(10) CONVERGE(0)
/METHOD=KMEANS(NOUPDATE)
/SAVE CLUSTER
/PRINT ID(CommId) INITIAL.
variable labels qcl_1 'Non-hierarchical K-means solution' .
execute.
```

- Use this syntax to do non-hierarchical K-means clustering with three clusters. Cluster identifiers will be saved to the data file (e.g., /SAVE CLUSTER will create a new variable in your SPSS file that will tell you which of the three clusters each community would be placed into). K-means clustering should be used when you want to form exactly three clusters that are to be as distinct as possible and should be run several times after sorting cases in different random orders. In most cases, TWOSTEP (which first pre-clusters the cases into small sub-clusters and then clusters the sub-clusters from the first step into the desired number of clusters) is preferable to QUICK CLUSTER. Also, do not select "Use running means" under the "Iterate" function when using K-means clustering.

The third step to cluster analysis is to find the appropriate solution. As we have provided syntax for the reader to conduct cluster analysis using different techniques, it is important to compare the hierarchical and non-hierarchical solutions by cross-tabulating results.

Syntax to See Differences between Solutions Produced by Hierarchical and Non-Hierarchical Clustering:

```
CROSSTABS
/TABLES=CLU3_1 BY QCL_1
/FORMAT=AVALUE TABLES
```

```
/STATISTICS=CHISQ KAPPA
/CELLS=COUNT TOTAL
/COUNT ROUND CELL.
```

- Use the syntax to see the differences between the solutions produced by hierarchical and non-hierarchical clustering. The crosstabs table produced in the SPSS output indicates there are seven times that both approaches placed communities into Cluster 1, 12 times that both approaches placed communities in Cluster 2, and 10 times that both approaches placed communities in Cluster 3. Thus, 29 of the 38 communities would be put into the same cluster (76% accuracy), while the other nine communities would be placed into different clusters based on whether a hierarchical or nonhierarchical clustering approach was used.

Syntax to Create a Variable that Indicates Whether the Results of the Two Clustering Approaches Match

```
(clu3_1 ne qcl_1) match=0.
if (clu3_1 eq qcl_1) match=1.
execute.
```

- Use this syntax to create a variable that indicates whether the results of the two solutions match for each of the 38 communities. There will be 0s for new variable in the SPSS data set for the nine communities that were placed into different clusters based on the approach used (Community ID 4, 6, 8, 18, 25, 27, 28, 29, and 30).

Syntax to Create a Scatterplot Matrix of Variables Used in Clustering

GRAPH

```
/SCATTERPLOT(MATRIX)=ConPov EtHet BusDis BY match
/MISSING=LISTWISE.
```

- Use this syntax to create a scatterplot matrix or SPLOM of the variables used in the clustering, coloring the markers by whether or not the results of the hierarchical and nonhierarchical solutions match. This scatterplot matrix presents different visual representations of the three cluster solution. Although it is somewhat challenging to identify three distinct clusters in all of the plots, the graphs for Concentrated Poverty by Ethnic Heterogeneity present a clear picture of a three-cluster solution.