Conduct and Interpret a Cluster Analysis

by James Lani

http://www.statisticssolutions.com/cluster-analysis-2/

Click here for to get help with your Thesis or Dissertation.

Click here for FREE Thesis and Dissertation resources (templates, samples, calculators).

What is the Cluster Analysis?

The Cluster Analysis is an explorative analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases, i.e., observations, participants, respondents. Cluster analysis is used to identify groups of cases if the grouping is not previously known. Because it is explorative it does make any distinction between dependent and independent variables. The different cluster analysis methods that SPSS offers can handle binary, nominal, ordinal, and scale (interval or ratio) data.

The Cluster Analysis is often part of the sequence of analyses of factor analysis, cluster analysis, and finally, discriminant analysis. First, a factor analysis that reduces the dimensions and therefore the number of variables makes it easier to run the cluster analysis. Also, the factor analysis minimizes multicollinearity effects. The next analysis is the cluster analysis, which identifies the grouping. Lastly, a discriminant analysis checks the goodness of fit of the model that the cluster analysis found and profiles the clusters. In almost all analyses a discriminant analysis follows a cluster analysis because the cluster analysis does not have any goodness of fit measures or tests of significance. The cluster analysis relies on the discriminant analysis to check if the groups are statistically significant and if the variables significantly discriminate between the groups. However, this does not ensure that the groups are actually meaningful; interpretation and choosing the right clustering is somewhat of an art. It is up to the understanding of the researcher and how well he/she understands and makes sense of his/her data! Furthermore, the discriminant analysis builds a predictive model that allows us to plug in the numbers of new cases and to predict the cluster membership.

Typical research questions the Cluster Analysis answers are as follows:

- **Medicine** – What are the diagnostic clusters? To answer this question the researcher would devise a diagnostic questionnaire that entails the symptoms (for example in psychology standardized scales for anxiety, depression etc.). The cluster analysis can then identify groups of patients that present with similar symptoms and simultaneously maximize the difference between the groups.
- **Marketing** – What are the customer segments? To answer this question a market researcher conducts a survey most commonly covering needs, attitudes, demographics, and behavior of customers. The researcher then uses the cluster analysis to identify homogenous groups of customers that have similar needs and attitudes but are distinctively different from other customer

segments.

- **Education** – What are student groups that need special attention? The researcher measures a couple of psychological, aptitude, and achievement characteristics. A cluster analysis then identifies what homogeneous groups exist among students (for example, high achievers in all subjects, or students that excel in certain subjects but fail in others, etc.). A discriminant analysis then profiles these performance clusters and tells us what psychological, environmental, aptitudinal, affective, and attitudinal factors characterize these student groups.
- **Biology** – What is the taxonomy of species? The researcher has collected a data set of different plants and noted different attributes of their phenotypes. A hierarchical cluster analysis groups those observations into a series of clusters and builds a taxonomy tree of groups and subgroups of similar plants.

Other techniques you might want to try in order to identify similar groups of observations are *Q-analysis*, *multi-dimensional scaling (MDS)*, and *latent class analysis*.

Q-analysis, also referred to as *Q factor analysis*, is still quite common in biology but now rarely used outside of that field. Q-analysis uses factor analytic methods (which rely on *R*—the correlation between variables to identify homogenous dimensions of variables) and switches the variables in the analysis for observations (thus changing the R into a Q).

Multi-dimensional scaling for scale data (interval or ratio) and correspondence analysis (for nominal data) can be used to map the observations in space. Thus, it is a graphical way of finding groupings in the data. In some cases MDS is preferable because it is more relaxed regarding assumptions (normality, scale data, equal variances and covariances, and sample size).

Lastly, *latent class analysis* is a more recent development that is quite common in customer segmentations. Latent class analysis introduces a dependent variable into the cluster model, thus the cluster analysis ensures that the clusters explain an outcome variable, (e.g., consumer behavior, spending, or product choice).

The Cluster Analysis in SPSS

Our research question for the cluster analysis is as follows:

When we examine our standardized test scores in mathematics, reading, and writing, what do we consider to be homogenous clusters of students?

In SPSS Cluster Analyses can be found in *Analyze/Classify…*. SPSS offers three methods for the cluster analysis: *K-Means Cluster*, *Hierarchical Cluster*, and *Two-Step Cluster*.

K-means cluster is a method to quickly cluster large data sets, which typically take a while to compute with the preferred hierarchical cluster analysis. The researcher must to define the number of clusters in advance. This is useful to test different models with a different assumed number of clusters (for example, in customer segmentation).

Hierarchical cluster is the most common method. We will discuss this method shortly. It takes time to

calculate, but it generates a series of models with cluster solutions from *1* (all cases in one cluster) to *n* (all cases are an individual cluster). *Hierarchical cluster* also works with variables as opposed to cases; it can cluster variables together in a manner somewhat similar to factor analysis. In addition, *hierarchical cluster* analysis can handle nominal, ordinal, and scale data, however it is not recommended to mix different levels of measurement.

Two-step cluster analysis is more of a tool than a single analysis. It identifies the groupings by running pre-clustering first and then by hierarchical methods. Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods. In this respect, it combines the best of both approaches. Also *two-step clustering* can handle scale and ordinal data in the same model. *Two-step cluster* analysis also automatically selects the number of clusters, a task normally assigned to the researcher in the two other methods.

The hierarchical cluster analysis follows three basic steps: 1) calculate the distances, 2) link the clusters, and 3) choose a solution by selecting the right number of clusters.

Before we start we have to select the variables upon which we base our clusters. In the dialog we add math, reading, and writing test to the list of variables. Since we want to cluster cases we leave the rest of the tick marks on the default.

In the dialog box *Statistics…* we can specify whether we want to output the proximity matrix (these are the distances calculated in the first step of the analysis) and the predicted cluster membership of the cases in our observations. Again, we leave all settings on default.

In the dialog box *Plots…* we should add the *Dendrogram*. The *Dendrogram* will graphically show how the clusters are merged and allows us to identify what the appropriate number of clusters is.

The dialog box *Method…* is very important! Here we can specify the distance measure and the clustering method. First, we need to define the correct distance measure. SPSS offers three large blocks of distance measures for interval (scale), counts (ordinal), and binary (nominal) data.

For scale data, the most common is *Square Euclidian Distance*. It is based on the Euclidian Distance between two observations, which uses Pythagoras' formula for the right triangle: the distance is the square root of squared distance on dimension x and y. The Squared Euclidian Distance is this distance squared, thus it increases the importance of large distances, while weakening the importance of small distances.

If we have ordinal data (counts) we can select between Chi-Square (think cross-tab) or a standardized Chi-Square called *Phi-Square*. For binary data SPSS has a plethora of distance measures. However, the Square Euclidean distance is a good choice to start with and quite commonly used. It is based on the number of discordant cases.

In our example we choose *Interval* and *Square Euclidean Distance*.

Next we have to choose the *Cluster Method*. Typically choices are Between-groups linkage (distance between clusters is the average distance of all data points within these clusters), nearest neighbor (single linkage: distance between clusters is the smallest distance between two data points), furthest neighbor (complete linkage: distance is the largest distance between two data points), and Ward's method (distance is the distance of all clusters to the grand average of the sample). Single linkage works best with long chains of clusters, while complete linkage works best with dense blobs of clusters and between-groups linkage works with both cluster types. The usual recommendation is to use single linkage first. Although single linkage tends to create chains of clusters, it helps in identifying outliers. After excluding these outliers, we can move onto Ward's method. Ward's method uses the F value (like an ANOVA) to maximize the significance of differences between cluster, which gives it the highest statistical power of all methods. The downside is that it is prone to outliers and creates small clusters.

A last consideration is *standardization*. If the variables have different scales and means we might want to standardize either to *Z scores* or just by centering the scale. We can also transform the values to absolute measures if we have a data set where this might be appropriate.

Output, syntax, and interpretation can be found in our downloadable manual: Statistical Analysis: A Manual on Dissertation Statistics in SPSS (included in our member resources). Click here to download
.
